KSCE Journal of Civil and Environmental Engineering Research
Vol. 46 No. 6: 1-17/ December 2026
DOI: https://doi.org/10.12652/Ksce.2026.46.6.01

ISSN 1015-6348 (Print)
ISSN 2799-9629 (Online)
www.kscejournal.or.kr

Information Technology **정보기술**

WIM 운영데이터 조회를 위한 하이브리드 LLM 질의응답 체계의 클라우드·로컬 모델 성능 비교

박준영* · 조윤범** · 김종우***
Park, Jun Young*, Jho, Youn Beom**, Kim, Jong Woo***

Performance Comparison of Cloud and Local LLMs in a Hybrid Question-Answering System for WIM Operational Data Retrieval

ABSTRACT

Weigh-in-Motion (WIM) systems accumulate large volumes of passage-event data across multiple sites, but practitioners without database expertise have difficulty retrieving the information they need. This study developed a hybrid question-answering framework in which a large language model (LLM) extracts query conditions and operations from Korean natural-language questions, while a separate query-processing module performs filtering, aggregation, ranking, and numerical operations on the actual data. Using a fixed sample of 200,000 records from approximately 26.29 million WIM passage events at five sites, three cloud models and four local models were evaluated on 40 queries of four types under the same schema-based prompt. The execution match rate was defined as the proportion of queries whose executed results matched the reference results. Claude Sonnet 4.5 and Gemini 2.5 Flash achieved 97.5 %, GPT-4o 95.0 %, and GPT-OSS 20B 87.5 %, while Qwen3-8B, Gemma3-4B, and Mistral-7B achieved 40.0 %, 30.0 %, and 15.0 %, respectively. Differences among the first four models were not statistically distinguishable at this sample size, whereas their differences from the remaining three were significant. Some local models showed a lower execution match rate when queries required grouping, ranking, additional operations, or more conditions, and most incorrect results originated from condition extraction rather than from numerical processing. Under a minimal prompt without schema and operation information, none of the seven models correctly processed any query. The quality of final natural-language answers was not evaluated, and the compared models differ in scale and execution environment. These results demonstrate the applicability of separating natural-language interpretation from numerical processing for WIM operational data and indicate that reliable question answering requires both sufficient model capability and explicit specification of the available data fields and supported operations.

Keywords : Weigh-in-Motion (WIM), Large Language Model (LLM), Freight transport safety, Hybrid question answering, Structured data querying

초 록

화물차 운송안전 모니터링을 위한 Weigh-in-Motion(WIM) 시스템에서는 대규모 통과이벤트 데이터가 축적되지만, 실무자가 전문지식 없이 필요한 정보를 조회·분석하기는 어렵다. 본 연구에서는 대규모 언어모델(LLM)이 한국어 자연어 질의에서 조회 조건과 연산 정보를 추출하고, 실제 데이터의 필터링·집계·정렬은 별도의 질의 처리 모듈이 수행하는 하이브리드 질의응답 체계를 구축하였다. 5개 현장에서 수집된 약 2,629만 건 중 200,000건의 고정 표본과 네 가지 유형의 40개 평가 질의를 이용하여 클라우드 모델 3종과 로컬 모델 4종을 비교하였다. 정답률은 모델이 추출한 조건의 실행 결과가 기준 결과와 일치한 비율로 정의하였다. Claude Sonnet 4.5와 Gemini 2.5 Flash가 각각 97.5 %, GPT-4o가 95.0 %, GPT-OSS 20B가 87.5 %였으며, Qwen3-8B, Gemma3-4B 및 Mistral-7B는 각각 40.0 %, 30.0 % 및 15.0 %였다. 앞의 네 모델 간 차이는 40개 질의 규모에서 통계적으로 구분되지 않았고, 이들과 나머지 세 모델의 차이는 유의하였다. 일부 로컬 모델에서는 그룹화, 정렬 또는

* 정회원 · (주)유디엔에스 개발본부 연구원 (UDNS. Co., Ltd. · pjy5451@gmail.com)
** (주)유디엔에스 개발본부 상무 (UDNS. Co., Ltd. · jhoyb74@gmail.com)
*** 종신회원 · 교신저자 · (주)유디엔에스 경영본부 사장 (**Corresponding Author** · UDNS. Co., Ltd. · roadload2014@gmail.com)

Received September 4, 2026/ **revised** September 30, 2026/ **accepted** October 1, 2026

Copyright ⓒ 2026 by the Korean Society of Civil Engineers
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

추가 연산이 요구되거나 조회 조건이 많아질수록 성능이 낮아졌으며, 대부분의 오답은 조건 추출 과정에서 발생하였다. 또한 데이터 스키마와 지원 연산 정보를 제공하지 않은 최소 프롬프트에서는 모든 모델이 40개 질의 중 하나도 처리하지 못하였다. 본 연구는 최종 자연어 답변의 품질을 평가하지 않았으며, 비교 모델의 규모와 실행 환경은 서로 다르다. 이러한 결과는 WIM 운영데이터의 자연어 질의응답에서 모델의 질의 해석 성능과 함께 조회 가능한 데이터 항목과 수행 가능한 연산을 프롬프트에 명확하게 제시하는 것이 중요함을 보여준다.

핵심용어 : 축중기(WIM), 대규모 언어모델(LLM), 화물차 운송안전, 하이브리드 질의응답, 정형데이터 질의

1. 서 론

화물차의 과적·이상 운행은 도로 파손과 교통안전에 직접적인 영향을 미치므로, 도로 관리 기관에서는 Weigh-in-Motion (WIM)을 이용해 통과 차량의 중량, 속도, 차로, 차종과 차량 제원 등을 상시 계측하고 과적 단속 지원과 운송안전 모니터링에 활용하고 있다(Lee et al., 2023; Bang et al., 2025; Bang et al., 2026). 이 과정에서 여러 현장에 대규모 통과이벤트 데이터가 지속적으로 축적되지만, 현장·기간·차로·차종·중량·제원 등 다양한 조건을 조합하여 필요한 정보를 즉시 조회하고 집계하려면 데이터베이스 구조와 조회 방법에 대한 이해가 필요하다. 그러나 실제 WIM 데이터를 활용하는 현장 실무자가 데이터베이스 구조나 전문적인 조회 방법에 익숙지 않기 때문에, 필요한 정보를 직접 조회·집계하는 데에는 어려움이 있다. 따라서 별도의 전문지식 없이도 자연어를 통해 WIM 데이터를 조회할 수 있는 방법이 요구된다.

최근 대규모 언어모델(Large Language Model, LLM)은 자연어 이해와 생성 능력이 크게 향상되면서 다양한 분야의 정보 검색과 질의응답에 활용되고 있다. 특히 사용자의 자연어 질의를 해석하고 응답을 생성하는 데 강점이 있어 건축·시공·시설관리 분야에서도 정보 검색과 문서 활용을 중심으로 적용이 확대되고 있다(Lee et al., 2024; Kampelopoulos et al., 2025). 이러한 특성은 전문적인 데이터베이스 질의문을 작성하지 않고도 자연어를 통해 정형데이터에 접근할 수 있도록 한다. 그러나 정형데이터의 수치를 LLM이 직접 계산하거나 추정하도록 하면 환각으로 인해 잘못된 결과가 생성될 수 있으며(Maynez et al., 2020; Ji et al., 2023), 자연어를 데이터베이스 질의로 변환하는 과정에서도 오류와 일관성 저하가 보고된 바 있다(Renggli et al., 2025). 특히 운송안전과 같이 수치의 신뢰성이 중요한 업무에서는 LLM의 자연어 이해 기능과 데이터의 조회·연산 기능을 분리하여 결과의 신뢰성을 확보할 필요가 있다. 또한 보안 등의 이유로 외부 클라우드 API의 활용이 제한되는 환경을 고려하면, 상용 클라우드 모델뿐만 아니라 로컬 환경에서 직접 운용할 수 있는 모델의 적용 가능성도 함께 검토할 필요가 있다.

이에 본 연구에서는 LLM이 한국어 질의에서 조회 조건과 연산 정보를 추출하고, 실제 필터링·집계·정렬은 별도의 질의 처리 모듈이 수행하는 하이브리드 질의응답 체계를 도로 WIM 통과이벤트에 적용하였다. Koh et al.(2026)은 교량 유지관리 데이터를 대상으로 LLM과 결정론적 엔진의 역할을 분리한 질의응답 구조의 적용 가능성을 제시하였다. 본 연구는 이를 다현장 WIM 운영데이터로 확장하여, 단일값 집계·그룹별 집계·순위 조회·추가 연산이 필요한 질의의 네 가지 질의 유형과 필터 조건의 수와 조합에 따른 질의 복잡도를 고려해 평가 질의군을 구성하고, 동일 조건에서 클라우드 모델과 로컬 모델의 질의 해석 성능을 비교하였다.

본 연구의 목적은 다음과 같다. 첫째, WIM 통과이벤트를 대상으로 자연어 질의에서 조회 조건과 연산 정보를 추출하고, 이를 바탕으로 별도의 질의 처리 모듈이 데이터를 조회·집계하는 하이브리드 질의응답 체계를 구축한다. 둘째, 40개 평가 질의를 이용하여 클라우드 모델 3종과 로컬 모델 4종의 정답률과 응답 시간을 비교하고 질의 유형별 성능 차이를 분석한다. 셋째, 데이터 스키마와 지원 연산에 관한 정보의 제공 여부가 자연어 질의의 해석 성능에 미치는 영향을 추가로 분석하여, 하이브리드 질의응답 체계의 구성에 필요한 조건을 검토한다. 본 연구의 범위는 과적 여부의 법적 판정이나 행정처분이 아니라, WIM 운영데이터의 자연어 조회와 집계 성능 평가에 한정한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 정리하고, 3장에서는 시스템과 평가 방법을 기술한다. 4장에서 실험 결과를 제시하고, 5장에서 결론을 제시한다.

2. 관련 연구

2.1 WIM 기반 화물차 운송안전 모니터링

WIM과 관련한 기존 연구에서는 주행 차량의 중량을 정확하게 계측하고 이를 과적차량 관리에 활용하기 위한 연구가 지속적으로 수행되어 왔다. Lee et al.(2023)과 Burnos et al.(2021)은 고속 WIM의 계측 정확도가 차량의 주행 경로, 축 구성 및 가감속 등의 영향을 받을 수 있음을 분석하였으며, 운송안전 모니터링에 WIM 데이터를 활용하기 위해서는 이러한 계측 특성을 고려할 필요가 있음을 제시하였다.

최근에는 개별 차량의 계측 정확도뿐만 아니라 지속적으로 축적되는 WIM 데이터를 관리하고 활용하기 위한 연구로 범위가 확대되

고 있다. Bang et al.(2025)은 차량의 운행 특성에 따른 저속 및 고속 WIM의 계측 오차를 분석하였으며, Bang et al.(2026)은 공용 교량에서 수집되는 과중차량 관리정보를 데이터베이스로 구축하여 지속적으로 관리할 수 있는 방안을 제시하였다. 또한 Lee et al.(2026)은 교량에 작용하는 차량 하중을 디지털 트윈 기반으로 식별하는 연구를 수행하는 등 WIM 및 차량 하중 데이터의 활용 범위도 확대되고 있다.

이러한 연구들은 WIM 데이터의 계측 정확도 향상과 수집·관리 및 활용에 초점을 두고 있다. 반면 여러 현장에서 장기간 축적된 WIM 운영데이터로부터 사용자가 필요한 정보를 다양한 조건으로 조회하고 집계하는 과정에 대해서는 상대적으로 다루어지지 않았다. 이에 따라 축적된 WIM 운영데이터를 실무자가 보다 쉽게 활용할 수 있도록 데이터 접근 및 질의 방법을 확장할 필요가 있다.

2.2 LLM 기반 정형데이터 질의응답

자연어를 통한 정형데이터 조회에서는 사용자의 질문을 데이터베이스 질의문으로 변환하는 Text-to-SQL 방식이 널리 연구되어 왔다(Yu et al., 2018; Hong et al., 2025). 최근에는 LLM의 자연어 이해 능력을 활용하여 복잡한 질의를 SQL로 변환하는 성능이 향상되고 있으나, 데이터베이스의 스키마를 정확히 해석하지 못하거나 존재하지 않는 필드와 연산을 생성하는 등의 오류가 여전히 보고되고 있다(Renggli et al., 2025). 이러한 오류는 질의 대상과 연산 방법이 다양해질수록 실제 데이터 조회 결과의 정확성에 영향을 줄 수 있다.

이에 따라 LLM이 최종 수치를 직접 생성하는 대신, 자연어 질의에서 데이터 조회에 필요한 조건과 연산 정보를 추출하고 실제 조회와 계산은 별도의 프로그램이 수행하는 방식도 제안되고 있다. 이와 같은 구조에서는 LLM이 현장, 기간, 차로와 같은 조회 조건과 집계·정렬 등의 연산 정보를 정해진 형식으로 전달하고, 이후의 데이터 처리는 사전에 정의된 절차에 따라 수행한다. 토목·시설관리 분야에서는 Koh et al.(2026)이 교량 유지관리 데이터를 대상으로 LLM과 별도의 연산 모듈을 결합한 질의응답 구조를 적용하여 자연어 기반 데이터 활용 가능성을 검토하였다.

기존 연구는 자연어 기반 데이터베이스 조회와 교량 유지관리 데이터 활용 가능성을 제시하였으나, 여러 WIM 현장에서 축적되는 대규모 통과이벤트를 대상으로 다양한 조회·집계 조건을 처리하고, 클라우드 모델과 로컬 모델의 성능을 동일한 조건에서 비교한 사례는 제한적이다. 본 연구에서는 이러한 적용 대상을 다현장 WIM 운영데이터로 확장하여 모델에 따른 자연어 질의 해석 성능을 비교하였다.

3. 시스템 구성 및 평가 방법

본 연구에서는 사용자의 한국어 자연어 질의를 LLM이 해석하고, 실제 WIM 데이터의 조회와 수치 연산은 별도의 질의 처리 모듈이 수행하는 하이브리드 질의응답 체계를 구축하였다. LLM은 질의에 포함된 조회 조건과 연산 정보를 추출하고, 질의 처리 모듈은 이를 바탕으로 데이터를 필터링·집계·정렬한다. 이후 처리 결과를 이용하여 사용자에게 자연어 답변을 제공한다. 전체 시스템의 구성과 처리 과정은 Fig. 1과 같다.

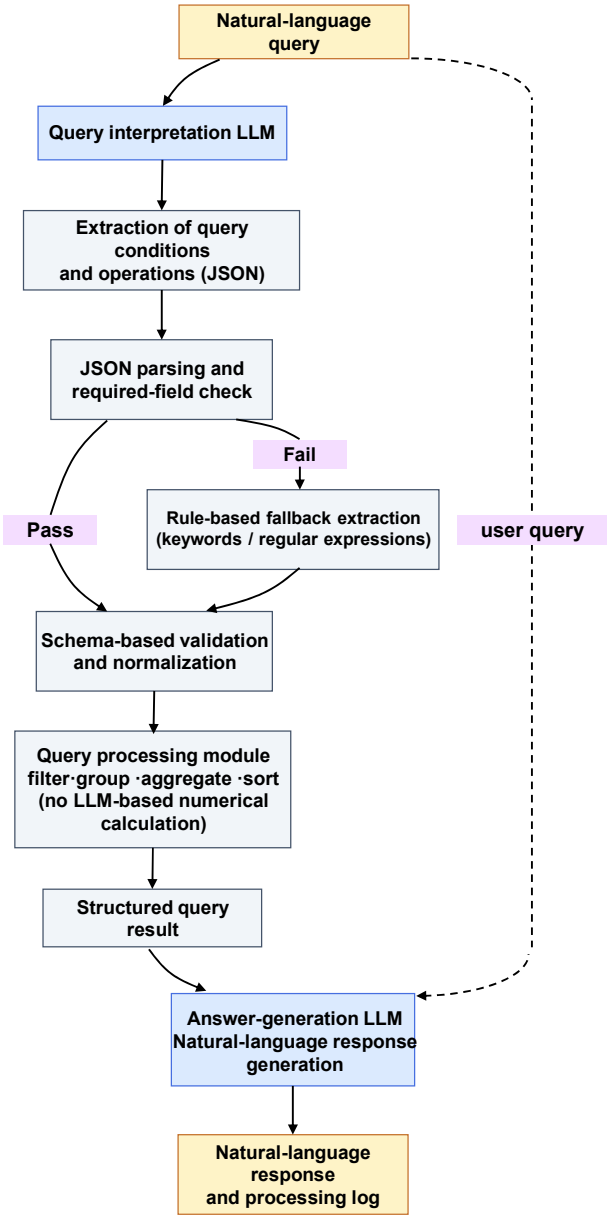


Fig. 1. Hybrid LLM-query-processing Pipeline for WIM Query Answering

3.1 질의응답 시스템의 처리 절차

본 연구의 질의응답 과정은 질의 해석, 데이터 처리, 답변 생성의 세 단계로 구성된다. 먼저 사용자의 한국어 자연어 질의를 LLM에 입력하여 데이터 조회에 필요한 조건과 연산 정보를 JSON 형식으로 추출한다. 조회 조건에는 현장, 기간, 방향, 차로, 차종 등이 포함되며, 연산 정보에는 필터링, 그룹화, 집계 및 정렬에 필요한 항목이 포함된다. 생성된 JSON은 형식과 필수 항목을 확인한 후 사전에 정의한 데이터 스키마와 허용 값에 따라 검증 및 정규화된다. LLM의 출력이 정상적인 JSON 형식으로 해석되지 않거나 필수 항목이 누락된 경우에는 키워드와 정규표현식을 이용한 규칙 기반 재추출을 적용하며, 이후 동일한 검증 및 정규화 과정을 거친다.

검증이 완료된 조회 조건과 연산 정보는 질의 처리 모듈에 전달되어 WIM 통과이벤트의 필터링, 그룹화, 집계 및 정렬에 사용된다. 이 과정의 수치 연산에는 LLM이 관여하지 않으며, 동일한 조회 조건과 연산 정보에 대해서는 동일한 결과가 산출된다. 이후 질의 처리 결과와 사용자의 자연어 질의를 LLM에 전달하여 최종 답변을 자연어 형태로 생성한다. 본 연구의 모델 성능 평가는 자연어 답변 생성 단계가 아닌 질의 처리 모듈에서 산출된 결과를 기준으로 수행하였다.

3.2 데이터셋 구축 및 전처리

본 연구에서는 2025년 7월부터 2026년 5월까지 5개 현장에서 수집된 약 2,629만 건의 WIM 통과이벤트 데이터를 사용하였다. 대상 현장은 HJ01(행주대교 북단), HJ02(행주대교 남단), AB01(안보검문소), MG01(문경검문소), BD01(봉동 테스트베드)이며, 주요 데이터 항목은 현장, 방향, 차로, 통과시각, 차종, 축 구성, 속도, 총중량, 축중 및 차량 제원(폭·높이·길이) 등이다. 차량번호 등 개인식별정보는 평가용 데이터셋 구축 과정에서 제거하였다.

WIM 데이터를 일관된 형식으로 조회·처리하기 위해 원본 데이터의 항목명을 공통 스키마에 맞게 정리하고, 날짜와 시간 정보를 결합하여 통과시각을 생성하였다. 속도, 중량 및 차량 제원은 각각 km/h, kg 및 mm 단위로 처리하였다. 또한 운송안전 관련 질의를 처리하기 위해 제원 이상, 제원초과 및 과적 관련 분류 항목을 추가하였다. 제원 이상은 폭·길이·높이 중 하나가 0 이하인 경우로 정의하였다. 제원초과는 도로법 시행령 제79조 제2항 제2호의 일반적인 운행제한 기준을 적용하여 폭 2,500 mm, 길이 16,700 mm 또는 높이 4,000 mm를 초과하는 경우로 구분하였다.

과적 관련 분류에는 도로법 시행령 제79조 제2항 제1호와 차량의 운행제한 규정 제11조 제1항 제1호를 적용하였다. 도로법 시행령에서는 축하중 10 t 또는 총중량 40 t을 초과하는 차량을 운행제한 대상으로 규정하고 있으며, 차량의 운행제한 규정에서는 측정 과정에서 발생할 수 있는 기기오차와 환경오차를 고려하여 측정값이 제한기준을 10 % 이내로 초과하는 경우에는 운행을 허용하도록 규정하고 있다. 이에 본 연구에서는 축하중 11 t 또는 총중량 44 t을 초과하는 경우를 과적 혐의로 구분하였다. 한편 축하중 13 t 또는 총중량 46 t을 초과하는 경우를 '고의 과적'으로 구분하여, 실제 운전자의 고의성에 대한 법적 판단 기준은 아니지만 연구용 분류로서 설정하였다.

모델 평가에는 전체 약 2,629만 건의 데이터 중 20만 건(약 0.76 %)으로 구성된 고정 표본을 사용하였다. 표본은 원천 데이터를 일정 크기의 구간 단위로 읽으면서 난수 시드 42를 적용하여 무작위로 추출하였다. 이와 함께 단속 결과 코드가 ER02 또는 ER03인 통과이벤트, 위반 코드가 CV04인 통과이벤트, 총중량

Table 1. Main Data Fields of WIM Passage Events

Category	Data field	Description	Unit
Event information	Site	WIM monitoring site	Code
	Direction	Travel direction	UP/DN
	Lane	Lane number	Integer
	Passage time	Date and time of vehicle passage	Datetime
Vehicle information	Vehicle class	Vehicle classification	Code
WIM measurements	Speed	Vehicle speed	km/h
	Gross vehicle weight	Gross vehicle weight	kg
	Axle load	Individual axle loads (axles 1-7)	kg
	Vehicle dimensions	Width, height, and length	mm
Safety-related classification	Dimension anomaly	Invalid vehicle-dimension measurement	Boolean
	Oversize	Dimensional limit exceedance	Boolean
	Overweight suspicion	WIM-based overweight suspicion	Boolean
	Overweight type	Axle/GVW exceedance classification	Boolean

40 t 이상인 통과이벤트를 희소 사례로 추가로 모았다. 이 희소 사례가 목표 20만 건의 5분의 1을 넘으면 시드 42로 그 규모까지 줄인 뒤 무작위 표본과 합치고, 중복을 제거하여 최종 20만 건으로 맞추었다. 현장별 층화는 하지 않았으며, 표본의 현장별 구성비는 HJ01 41.8 %, HJ02 20.5 %, AB01 15.0 %, MG01 12.6 %, BD01 10.2 %이다(Table A3). 희소 사례를 추가로 추출하였으므로 표본에서 산출되는 건수와 비율은 모집단의 추정치가 아니며, 본 표본은 모델 간 비교를 위한 공통 평가 데이터로만 사용하였다. 모든 모델과 평가 질의에는 동일한 표본 파일을 적용하였다.

3.3 조회 조건 및 연산 정보의 구성

자연어 질의에서 추출되는 조회 조건과 연산 정보는 JSON 형식으로 표현하였다. JSON은 처리 유형(query_type), 조회 조건(filters) 및 연산 정보(agg)를 중심으로 구성하였으며, 순위 조회와 같이 결과의 정렬이 필요한 경우에는 정렬 정보(sort)를 추가하였다. 조회 조건에는 현장, 방향, 차로, 기간, 차종, 속도 및 총중량 등의 데이터 항목과 제원초과, 과적 혐의 등의 분류 항목을 포함하였다.

현장, 방향, 차로 및 차종은 하나 이상의 값을 지정할 수 있도록 하였으며, 기간과 속도, 총중량 등의 수치 항목은 범위를 지정할 수 있도록 구성하였다. 연산 정보에는 조건을 만족하는 차량의 건수와 평균값을 산출하는 집계와 현장, 차종, 방향, 차로 및 연월에 따른 그룹별 집계와 정렬을 포함하였다. 이와 함께 제원초과 유형별 비교, 그룹별 과적 혐의 비율 산정 및 차종별 과적 축 분석 등 WIM 운영데이터의 특성을 반영한 연산을 구성하였다.

조회 조건과 연산 정보에 따른 데이터의 필터링, 그룹화, 집계 및 정렬은 Python의 Pandas를 이용하여 구현한 질의 처리 모듈에서 수행하였다. Fig. 2는 실제 평가 질의인 “봉동 상행 과적 혐의 건수는?”을 예로 들어 자연어 질의에서 추출된 조회 조건과 연산 정보를 보여준다. 이 질의에서는 BD01, 상행 방향, 과적 혐의 여부가 조회 조건으로 지정되고, 해당 조건을 만족하는 통과이벤트의 건수를 산출하도록 구성된다.

Query (translated from Korean)

How many overweight-suspicion events occurred at BD01 upstream?

Extracted Query Conditions and Operation

```
{
  "query_type": "aggregate",
  "filters": {
    "site_id": ["BD01"],
    "direction": ["UP"],
    "is_overweight_enforce": true
  },
  "agg": {
    "op": "count"
  }
}
```

Fig. 2. Example of Query Interpretation

3.4 비교 대상 모델

본 연구에서는 외부 API를 이용하는 클라우드 모델과 로컬 환경에서 직접 실행할 수 있는 모델을 비교 대상으로 선정하였다(Table 2). 클라우드 모델은 Claude Sonnet 4.5(Anthropic, 2025), Gemini 2.5 Flash(Gemini Team, 2025), GPT-4o(OpenAI, 2024)의 3종을 사용하였으며, 로컬 모델은 GPT-OSS 20B (OpenAI, 2025), Qwen3-8B(Yang et al., 2025), Gemma3-4B (Gemma Team, 2025), Mistral-7B(Jiang et al., 2023)의 4종을 사용하였다. 본 비교는 동일한 파라미터 규모의 모델 간 성능을 비교하는 벤치마크가 아니라, 클라우드와 로컬 환경에서 현실적으로 운용 가능한 모델을 대상으로 WIM 질의응답에 대한 적용 성능을 비교하는 데 목적이 있다.

클라우드 모델은 각 제공사의 API를 통해 2026년 8월 21일에 호출하였으며, 호출 시 사용한 모델 식별자는 Table 2와 같다. 날짜가 포함되지 않은 식별자(gemini-2.5-flash, gpt-4o)는 호출 시점에 제공사가 연결한 버전이 사용되었으며, 세부 버전은 별도로 기록하지 않았다. 로컬 모델은 NVIDIA GeForce RTX 4070 SUPER(약 12 GB) 환경에서 실행하였으며, GPT-OSS 20B는 Ollama를 이용해 모델 배포 형식인 MXFP4 가중치로 실행하였고, 컨텍스트 길이는 Ollama의 기본값을 사용하였다. 나머지 로컬

Table 2. Cloud and Local LLMs for Performance Comparison

Class	Model	Provider	Model identifier	Execution
Cloud	Claude Sonnet 4.5	Anthropic	claude-sonnet-4-5-20250929	API
	Gemini 2.5 Flash	Google	gemini-2.5-flash	API
	GPT-4o	OpenAI	gpt-4o	API
Local	GPT-OSS 20B	OpenAI	gpt-oss:20b	Ollama, MXFP4
	Qwen3-8B	Alibaba	Qwen/Qwen3-8B	Transformers, NF4 4-bit
	Gemma3-4B-IT	Google	google/gemma-3-4b-it	Transformers, NF4 4-bit
	Mistral-7B-Instruct	Mistral AI	mistralai/Mistral-7B-Instruct-v0.3	Transformers, NF4 4-bit

모델은 Hugging Face Transformers와 bitsandbytes의 NF4 4비트 양자화를 적용하여 실행하였으며, 입력은 최대 4,096토큰으로 제한하였다. 이하 본문에서는 Gemma3-4B-IT와 Mistral-7B-Instruct를 각각 Gemma3-4B와 Mistral-7B로 표기한다.

모든 모델에는 동일한 스키마 기반 프롬프트를 적용하였다. 출력의 무작위성을 배제하기 위해 온도(temperature)를 0으로 설정하였으며, Transformers 기반 모델은 greedy 디코딩을 적용하였다. 난수 시드는 별도 지정하지 않았으며, 조건 추출 단계의 최대 출력 길이는 API 기반 모델과 GPT-OSS 20B는 1,024토큰, Transformers 기반 모델은 512 토큰으로 설정하였으며 Qwen3-8B는 추론 모드를 비활성화하여 실행하였다.

3.5 평가 질의 및 평가 방법

모델의 질의 해석 성능을 평가하기 위해 WIM 운영데이터의 다양한 조회 형태를 반영한 총 40개의 평가 질의(N01-N40)를 구성하였다. 평가 질의는 단일값 집계(A), 그룹별 집계(B), 순위 조회(C), 추가 연산(D)의 네 가지 유형으로 구분하고, 각 유형별로 10개의 질의를 구성하였다(Table 3). 유형 A는 차량 건수나 평균값과 같이 하나의 값을 요구하는 질의이며, 유형 B는 현장, 차종, 방향 또는 차로 등에 따른 그룹별 집계 결과를 요구하는 질의이다. 유형 C는 그룹별 집계 결과를 기준으로 순위를 조회하는 질의이며, 유형 D는 그룹 간 비율이나 최댓값을 비교하거나 차종별 과적 축을 판별하는 등의 추가 연산을 요구하는 질의로 구성하였다.

각 유형의 대표 질의 또한 Table 3에 함께 제시하였다. 예시한 N05, N16, N25, N36은 모두 BD01 상행 방향을 대상으로 하는 구조 깊이 SD3의 질의로, 조회 조건의 수는 같지만 각각 전체 건수, 차종별 건수, 차종별 건수의 순위, 과적 혐의 비율이 가장 높은 차종을 요구하여 유형 A에서 D로 갈수록 필요한 연산이 늘어난다.

질의에 포함되는 조회 조건의 수에 따른 성능 차이를 확인하기 위해 구조 깊이(Structural Depth, SD)를 별도로 구분하였다. 구조 깊이는 질의에 포함된 필터 조건의 수를 기준으로 SD1부터 SD4까지 분류하였다(Table 4). 현장, 방향, 차로 및 차종 등의 조건은 각각 하나로 계산하였고, 기간 조건은 시작 지점과 종료 시점을 포함하여 하나의 조건으로 계산하였으며, 속도나 총중량과 같은 수치 범위 조건은 하한과 상한을 각각 하나의 조건으로 계산하였다. 그룹화, 집계 및 정렬에 관한 연산은 구조 깊이 산정에 포함되지 않았다.

Table 4. Definition of Structural Depth

Structural depth	Number of filter conditions	n
SD1	0-1	8
SD2	2	7
SD3	3	13
SD4	≥4	12

각 평가 질의에 대해서는 기준이 되는 조회 조건과 연산 정보를 사전에 작성하고, 이를 질의 처리 모듈에 적용하여 기준 결과를 산출하였다. 모델이 추출한 조회 조건과 연산 정보도 동일한 질의 처리 모듈에 적용하였으며, 산출된 결과가 기준 결과와 일치하면 정답으로 판정하였다. 단일값 결과는 모두 건수로서 기준값과 정확히 같을 때 일치로 판정하였다. 그룹별 결과는 행의 표시 순서와 관계없이 각 행의 그룹 키와 값의 집합이 모두 같을 때 일치로 판정하였으며, 평균값과 비율은 부동소수점 연산에 따른 표현 오차만 무시할 수 있도록 10^{-6} 이내의 차이를 같은 값으로 처리하였다. 순위 조회의 결과도 그룹별 결과로 산출되므로 행 집합을 기준으로 비교하였으며, 정렬 기준과 방향의 정확성은 4.2절의 항목별 비교에서 별도로 확인하였다. 최댓값 1행을 선택하는 연산에서 동률이 발생하면 첫 번째 행을 선택하였다. 평균은 결측값을 제외하고 산출하였으며, 필터 대상 항목이 결측인 행은 해당 조건을 만족하지 않는 것으로 처리하였다. 그룹별 비율의 분모는 그룹에 속한 통과이벤트 수이므로 0이 되지 않는다.

다만 실행 결과의 일치가 조건 해석의 정확성을 직접 의미하지는

Table 3. Query Types for the Evaluation

Type	Query type	Description	n	Example query (ID)
A	Single-value aggregation	Returns a single aggregated value	10	How many overweight-suspicion events occurred at BD01 upstream? (N05)
B	Grouped aggregation	Returns aggregated values by group	10	What are the counts by vehicle class of overweight-suspicion vehicles at BD01 upstream? (N16)
C	Ranking	Returns ranked results based on aggregated values	10	Rank vehicle classes by overweight-suspicion count at BD01 upstream. (N25)
D	Additional operation	Returns results requiring ratios, maxima, or group-wise comparisons	10	Which vehicle class has the highest overweight-suspicion rate at BD01 upstream in the second half of 2025? (N36)

않는다. 서로 다른 조건이 같은 결과를 산출할 수 있고, 기준 결과와 모델 결과가 같은 질의 처리 모듈을 거치므로 모듈 자체의 오류는 드러나지 않기 때문이다. 이에 4.2절에서 기준 결과가 비어 있는 질의의 수, 결과는 일치하나 조건 표기가 다른 응답의 의미상 동등성, 그리고 질의 처리 모듈과 독립적인 SQL 실행을 통한 기준 결과의 재현 여부를 추가로 검증하였다.

평가지표는 정답률과 응답 시간을 사용하였다. 여기서 정답률은 모델이 추출한 조건의 실행 결과가 기준 결과와 일치한 질의의 비율, 즉 실행 결과 일치율(Execution Match Rate, EMR)을 의미하며 Eq. (1)과 같이 산정하였다.

$$EMR(\%) = \frac{N_{match}}{N_{total}} \times 100 \quad (1)$$

여기서 N_{match}는 실행 결과가 기준 결과와 일치한 질의 수, N_{total}은 평가 대상의 전체 질의 수를 의미하며, 이하 본문에서는 정답률로 표기한다. 전체 정답률과 함께 질의 유형 및 구조 깊이에 따른 정답률도 동일한 방법으로 산정하였다. 평가 질의가 40개로 제한되므로 정답률에는 Wilson 방법에 의한 95 % 신뢰구간을 함께 제시하였으며, 모델 간 차이는 McNemar 정확검정으로 검토하였다. 또한 오류 원인을 질의 해석 오류, 스키마 오류, 질의 처리 오류 및 시스템 오류로 구분하여 기록하고, 규칙 기반 재추출의 적용 여부도 확인하였다.

응답 시간은 각 평가 질의에 대한 모델 호출부터 정오 판정을 위한 처리 과정이 완료될 때까지의 총 시간을 측정하였으며, 최종 자연어 답변 생성 시간은 포함하지 않았다. 클라우드 모델의 경우 API 통신에 따른 네트워크 지연이 포함된다.

데이터 스키마와 연산 규칙의 제공 여부가 질의 해석 성능에 미치는 영향을 확인하기 위해 최소 프롬프트와 스키마 기반 프롬프트를 추가로 비교하였다. 최소 프롬프트에서는 질의를 JSON 객체 하나로 변환하고 수치를 계산하지 말라는 지시만 제공하였으며, 스키마 기반 프롬프트에서는 3.3절에서 정의한 조회 가능한 데이터 항목과 허용 값, 연산 규칙, 과적 관련 용어의 대응 관계 및 JSON 출력 예시를 제공하였다. 다만 출력 예시 15개 중 5개는 평가 질의 N02, N11, N21, N22, N31의 기준 조건과 의미상 같으므로, 4.1절에 이들 질의를 제외한 정답률을 함께 기술하였다. 두 조건에는 동일한 40개의 평가 질의를 적용하였으며, 전체 평가 질의 목록과 두 프롬프트의 원문은 부록에 제시하였다.

4. 실험 결과

4.1 모델별 종합 성능 지표

7종 모델의 40개 평가 질의에 대한 정답률과 평균 응답 시간을 Table 5에 제시하였다. 클라우드 모델의 정답률은 Claude Sonnet 4.5와 Gemini 2.5 Flash가 각각 97.5 %(39/40), GPT-4o가 95.0 % (38/40)였으며, 로컬 모델은 GPT-OSS 20B가 87.5 % (35/40), Qwen3-8B 40.0 %(16/40), Gemma3-4B 30.0 % (12/40), Mistral-7B 15.0 %(6/40)였다.

평가 질의가 40개이므로 질의 1개의 차이는 정답률 2.5 %p에 해당하며, 신뢰구간도 넓게 나타났다. 동일 질의에 대한 정오 결과를 대응시켜 검토한 결과, 클라우드 모델 3종과 GPT-OSS 20B 사이의 차이는 유의하지 않았고 이들과 나머지 로컬 모델 3종의 차이는 유의하였다(McNemar 정확검정, 유의수준 0.05). 따라서 본 평가 규모에서는 모델 간의 세부 순위보다 두 군의 구분에 의미를 두어 해석하였다. 한편 3.5절에서 기술한 스키마 기반 프롬프트의 출력 예시와 중복되는 5개 질의를 제외하면, 상위 4종의 정답률은 88.6~97.1 %로 거의 변화가 없었으나 Qwen3-8B와 Gemma3-4B는 각각 34.3 %와 22.9 %로 낮아졌다. 두 군의 구분은 동일하게 유지되었다.

평균 응답 시간은 Claude Sonnet 4.5가 2.95초, Gemini 2.5 Flash가 3.19초, GPT-4o가 3.01초로 세 클라우드 모델 모두 약 3초 수준이었다. 로컬 모델은 GPT-OSS 20B 8.40초, Qwen3-8B는 5.94초, Gemma3-4B 5.51초, Mistral-7B 7.27초로 나타났다. Fig. 3은 모델별 정답률과 평균 응답 시간의 관계를 신뢰구간과

Table 5. Overall Performance of the Evaluated Models

Model	Deploy	Execution match rate (%)	95 % CI (%)	Matched	Mean response time (s)
Claude Sonnet 4.5	Cloud	97.5	87.1 ~ 99.6	39/40	2.95
Gemini 2.5 Flash	Cloud	97.5	87.1 ~ 99.6	39/40	3.19
GPT-4o	Cloud	95.0	83.5 ~ 98.6	38/40	3.01
GPT-OSS 20B	Local	87.5	73.9 ~ 94.5	35/40	8.40
Qwen3-8B	Local	40.0	26.3 ~ 55.4	16/40	5.94
Gemma3-4B	Local	30.0	18.1 ~ 45.4	12/40	5.51
Mistral-7B	Local	15.0	7.1 ~ 29.1	6/40	7.27

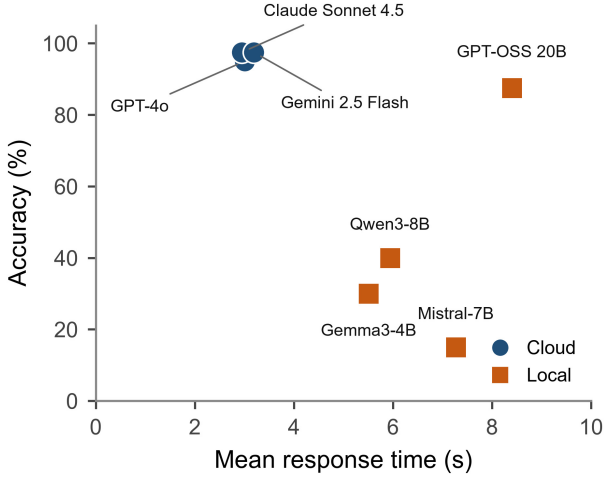


Fig. 3. Execution Match Rate Versus Mean Response Time by Model

함께 나타낸다. 클라우드 모델 3종은 약 3초의 평균 응답 시간과 95 % 이상의 정답률을 보였다. 로컬 모델에서는 GPT-OSS 20B가 상위군에 속하였으나 평균 응답 시간이 가장 길었으며, 나머지 로컬 모델은 GPT-OSS 20B보다 응답 시간은 짧았으나 정답률이 뚜렷하게 낮았다.

4.2 실행 결과 일치의 검증

3.5절에서 기술한 바와 같이 실행 결과의 일치가 조건 해석의 정확성을 직접 의미하지는 않으므로, 세 가지 보완 검증을 수행하였다.

첫째, 40개 질의의 기준 결과 중 0건이거나 빈 그룹인 경우는 없었다. 따라서 기준 조건과 잘못된 조건이 모두 빈 결과를 반환하여 정답으로 판정된 사례는 없다.

둘째, 정답으로 판정된 응답에 대해 모델이 추출한 조건을 필터, 그룹화, 집계 및 정렬 항목별로 기준 조건과 비교하였다. 항목별로 차이가 있었던 응답은 상위 4종 모델에서 모델별 8건, 하위 3종 모델에서 6건이었으며, 이를 질의 처리 모듈의 코드상 동작에 근거하여 검토한 결과 대부분은 정렬·출력 항목의 명시 여부나 기본값의 생략과 같은 표기 차이로 결과가 같아지는 경우였다. 예외는 Mistral-7B의 1건(N09)으로, 질의에 없는 조건을 추가하였으나 표본에 해당 사례가 없어 결과가 일치하였다. 이를 오답으로 보면 Mistral-7B의 정답률은 12.5 %가 되며, 모델군의 구분은 달라지지 않는다. 또한 순위 조회에서 정답으로 판정된 43건은 모두 정렬 기준과 방향이 기준 조건과 일치하였다.

셋째, 질의 처리 모듈과 독립적인 실행 경로로 기준 결과를 확인하기 위해 기준 조건을 SQL로 작성하여 SQLite에서 실행한 결과, 40개 질의 모두 질의 처리 모듈의 기준 결과와 일치하였다.

4.3 질의 유형 및 구조 깊이에 따른 성능

질의 유형별 정답률을 Table 6에 제시하였다. 클라우드 모델 3종은 단일값 집계(A), 그룹별 집계(B) 및 순위 조회(C)에서 모두 100.0 %의 정답률을 보였다. 추가 연산(D)에서는 Claude Sonnet 4.5와 Gemini 2.5 Flash가 각각 90.0 %, GPT-4o가 80.0 %로 다른 유형에 비해 다소 낮은 정답률을 보였다. GPT-OSS 20B는 A, B 및 D에서 각각 90.0 %, C에서 80.0 %를 기록하였다.

Qwen3-8B, Gemma3-4B 및 Mistral-7B에서는 질의 유형에 따른 성능 차이가 크게 나타났다. Qwen3-8B는 유형 A에서 70.0 %의 정답률을 보였으나, B, C, D에서는 각각 30.0 %로 낮아졌다. Gemma3-4B는 A에서 40.0 %, B와 D에서 30.0 %, C에서 20.0 %의 정답률을 보였으며, Mistral-7B는 A에서 30.0 %, B에서 20.0 %, C에서 0.0 %, D에서 10.0 %의 정답률을 보였다. 따라서 이들 로컬 모델에서는 단일값 집계에 비해 그룹화, 정렬 또는 추가 연산이 요구되는 질의에서 전반적으로 낮은 정답률이 나타났다.

Fig. 4는 각 모델의 40개 평가 질의에 대한 정오 판정 결과를 나타낸다. 이를 통해 Table 6의 질의 유형별 정답률과 함께 모델별 오답이 개별 평가 질의에 어떻게 분포하는지를 확인할 수 있다.

구조 깊이에 따른 정답률은 Table 7과 같다. 클라우드 모델은 대부분의 구조 깊이에서 높은 정답률을 유지하였으며, GPT-OSS 20B 역시 SD1-SD4에서 75.0 % 이상의 정답률을 보였다. Qwen3-8B는 SD1과 SD2에서 각각 62.5 %와 71.4 %의 정답률을 보였으나 SD3과 SD4에서는 30.8 %와 16.7 %로 낮아졌다. Gemma3-4B도 SD1의 75.0 %에서 SD4의 8.3 %까지 전반적으로 감소하는 경향을 보였으며, Mistral-7B는 모든 구조 깊이에서 낮은 정답률을 보였다.

Table 6. Execution Match Rate by Query Type

Type	Claude Sonnet 4.5	Gemini 2.5 Flash	GPT-4o	GPT-OSS 20B	Qwen3-8B	Gemma3-4B	Mistral-7B
A	100.0 %	100.0 %	100.0 %	90.0 %	70.0 %	40.0 %	30.0 %
B	100.0 %	100.0 %	100.0 %	90.0 %	30.0 %	30.0 %	20.0 %
C	100.0 %	100.0 %	100.0 %	80.0 %	30.0 %	20.0 %	0.0 %
D	90.0 %	90.0 %	80.0 %	90.0 %	30.0 %	30.0 %	10.0 %

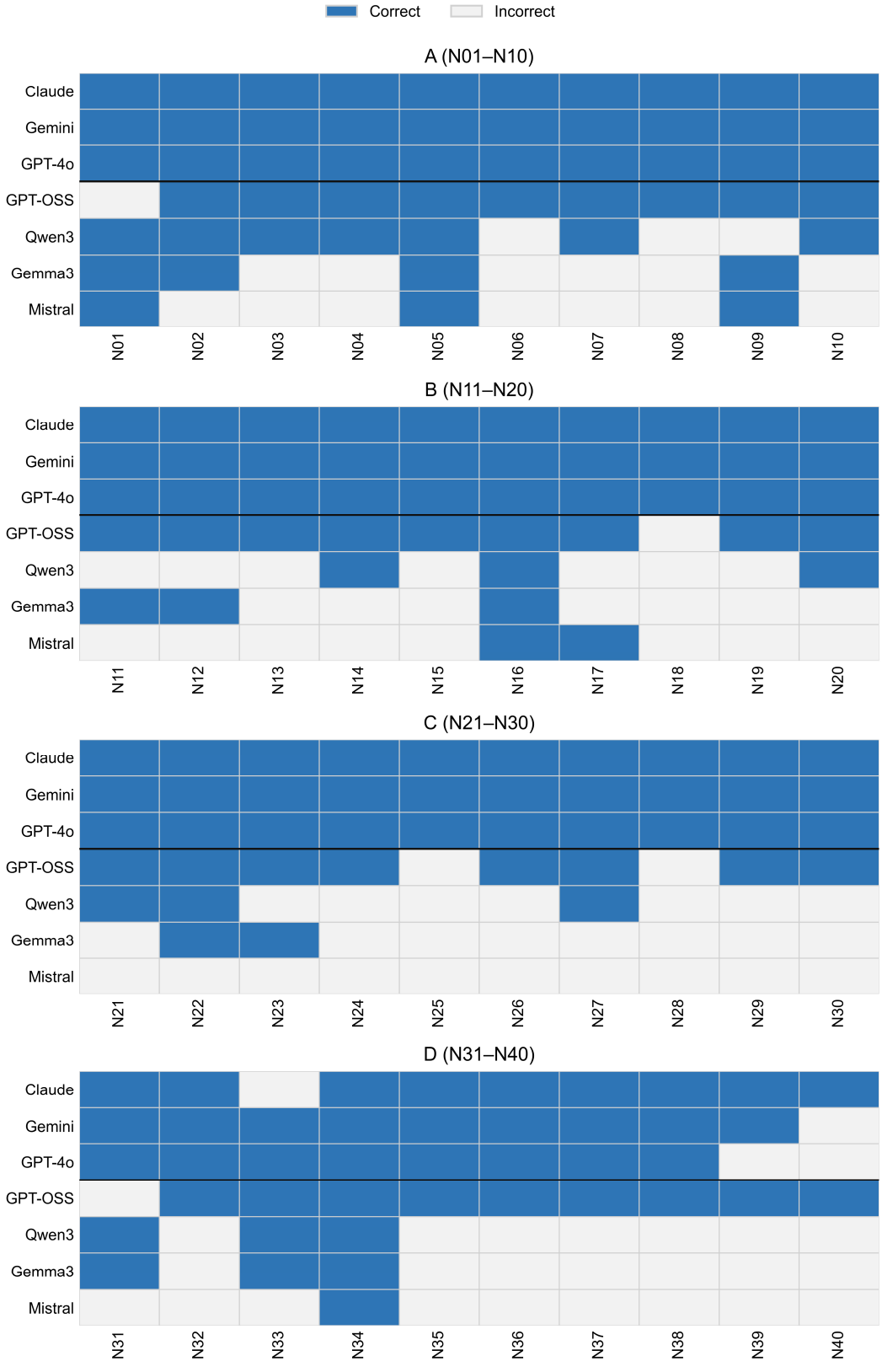


Fig. 4. Execution Match Results by Model and Evaluation Query

Table 7. Execution Match Rate by Structural Depth

Depth	Claude Sonnet 4.5	Gemini 2.5 Flash	GPT-4o	GPT-OSS 20B	Qwen3-8B	Gemma3-4B	Mistral-7B
SD1	100.0 %	100.0 %	100.0 %	75.0 %	62.5 %	75.0 %	12.5 %
SD2	85.7 %	100.0 %	100.0 %	100.0 %	71.4 %	42.9 %	14.3 %
SD3	100.0 %	100.0 %	100.0 %	92.3 %	30.8 %	15.4 %	23.1 %
SD4	100.0 %	91.7 %	83.3 %	83.3 %	16.7 %	8.3 %	8.3 %

종합하면, 클라우드 모델과 GPT-OSS 20B는 질의 유형과 구조 깊이가 달라져도 비교적 높은 성능을 유지한 반면, 나머지 로컬 모델은 그룹화·정렬·추가 연산이 요구되는 질의와 필터 조건이 많은 질의에서 상대적으로 낮은 성능을 보였다. 다만 유형별 질의는 10개, 구조 깊이별 질의는 7~13개이므로 셀별 차이의 해석에는 주의가 필요하다. 예를 들어 10개 질의 중 9개가 정답인 90.0 %의 95 % 신뢰구간은 59.6~98.2 %이다. 따라서 상위군 내부의 유형별·구조 깊이별 차이보다는 하위군에서 나타난 일관된 경향에 의미를 두어 해석하였다.

4.4 오류 유형 분석

모델별 오류 유형을 Table 8에 제시하였다. 클라우드 모델에서 발생한 오답은 모두 질의 해석 오류로, Claude Sonnet 4.5와 Gemini 2.5 Flash에서 각각 1건, GPT-4o에서 2건이 발생하였다. GPT-OSS 20B에서도 5건의 오답이 모두 질의 해석 오류로 나타났다. Qwen3-8B에서는 질의 해석 오류가 24건 발생하였으며, Gemma3-4B에서는 질의 해석 오류 27건과 질의 처리 오류 1건, Mistral-7B에서는 질의 해석 오류 32건과 스키마 오류 및 질의 처리 오류가 각각 1건 발생하였다. 시스템 오류는 모든 모델에서 발생하지 않았다.

전체 오류 중 대부분은 자연어 질의에서 조회 조건과 연산 정보를 추출하는 과정에서 발생하였으며, 스키마 오류와 질의 처리 오류는 일부 모델에서만 나타났다. 이는 모델별 정답률의 차이가 주로 질의 처리 모듈의 연산 과정이 아니라 LLM의 질의 해석 단계에서 나타났음을 보여준다.

한편, LLM의 출력이 정상적으로 처리되지 않은 경우 키워드와 정규표현식을 이용하여 조회 조건과 연산 정보를 다시 추출하도록 하였다. 이러한 규칙 기반 재추출은 Gemini 2.5 Flash와 GPT-OSS 20B에서 각각 1건, Qwen3-8B에서 2건 수행되었으나, 이를 통해 정답으로 전환된 사례는 없었다.

이 중 Gemini 2.5 Flash의 1건(N40)은 조건 추출 응답이 128자에서 종료되어 JSON 해석에 실패한 사례로, 규칙 기반 재추출이 적용되었으나 제원초과 유형별 비율과 최댓값 선택을 복원하지 못하고 폭 초과 건수를 산출하는 조건으로 처리되었다. 동일 설정으로 재현한 결과, 모델의 내부 추론 토큰(974토큰)이 최대 출력 길이(1,024토큰)의 대부분을 소비하여 응답이 종료된 것으로 확인되었다. 이 사례는 Table 8에서 질의 해석 오류로 집계되었으나, 모델의 해석 능력보다는 출력 길이 설정과 관련된 사례이다. 로컬 모델의 재추출 3건은 GPT-OSS 20B의 빈 응답 1건(N31)과 Qwen3-8B의 JSON 구문 오류 2건(N13, N40)이었으며, GPT-OSS 20B의 빈 응답은 응답 종료 사유가 기록되지 않아 원인을 확인하지 못하였다.

모델별 오류의 특성을 확인하기 위해 추출된 조건을 항목별로 기준 조건과 비교한 결과를 Table 9에 제시하였다. 항목별 일치율에는 4.2절에서 검토한 의미상 동등한 표기 차이가 불일치로 포함되어 있어 실제 해석의 정확도보다 낮게 산출되지만, 모델 간 차이의 양상을 확인하는 데에는 유용하다. 상위군 모델의 필터 항목 일치율은 80.0~90.0 %였으나 하위군 모델은 17.5~42.5 %에 그쳤다. 반면 그룹화 항목은 하위군에서도 80.0~90.0 %로 비교적 높았으며, 집계 항목은 72.5~85.0 %였다. 즉 하위군의 성능 저하는 주로

Table 8. Error Counts by Model

Model	Query interpretation error	Schema error	Query processing error	System error
Claude Sonnet 4.5	1	0	0	0
Gemini 2.5 Flash	1	0	0	0
GPT-4o	2	0	0	0
GPT-OSS 20B	5	0	0	0
Qwen3-8B	24	0	0	0
Gemma3-4B	27	0	1	0
Mistral-7B	32	1	1	0

Table 9. Item-level Agreement between Extracted and Reference Conditions (%)

Model	Execution match	Full match	Filters	Grouping	Aggregation	Sorting
Claude Sonnet 4.5	97.5	77.5	90.0	100.0	97.5	85.0
Gemini 2.5 Flash	97.5	77.5	90.0	100.0	92.5	87.5
GPT-4o	95.0	75.0	90.0	100.0	92.5	87.5
GPT-OSS 20B	87.5	67.5	80.0	97.5	92.5	85.0
Qwen3-8B	40.0	35.0	42.5	90.0	82.5	85.0
Gemma3-4B	30.0	27.5	30.0	80.0	72.5	77.5
Mistral-7B	15.0	7.5	17.5	80.0	85.0	75.0

Table 10. Representative Error Cases

Model	Query (type, depth)	Deviation from the reference conditions	Effect on the execution result
Claude Sonnet 4.5	N33 (D, SD2)	Added the overweight-suspicion flag as a filter	All group rates became 1.0
Gemini 2.5 Flash	N40 (D, SD4)	Truncated; fallback kept only width oversize	Width-oversize count returned
GPT-4o	N39 (D, SD4)	Added the intentional-overweight flag as a filter	All group rates became 1.0
GPT-4o	N40 (D, SD4)	Omitted maximum-row selection	Returned the three oversize-type rates
GPT-OSS 20B	N01 (A, SD1)	Omitted the overweight-suspicion filter	Counted all 200,000 passages
GPT-OSS 20B	N18 (B, SD4)	Omitted the overweight-suspicion filter	Class counts included all passages
GPT-OSS 20B	N25 (C, SD3)	Omitted the overweight-suspicion filter	Ranked all passages by class
GPT-OSS 20B	N28 (C, SD4)	Omitted the overweight-suspicion filter	Ranked all passages by class
GPT-OSS 20B	N31 (D, SD1)	Empty response	No execution result
Qwen3-8B	N23 (C, SD2)	Omitted the overweight-suspicion filter	Ranked all passages at the site by class
Gemma3-4B	N03 (A, SD2)	Read “excluding class 1” as “class 1 only”	Returned the class-1 count
Mistral-7B	N02 (A, SD1)	Added an unrequested site condition	Count restricted to one site

현장, 방향, 기간, 차종 및 과적 구분 등의 조회 조건을 질의 처리 모듈이 정의한 항목명과 값으로 추출하는 단계에서 발생하였으며, 이는 필터 조건이 많아질수록 정답률이 낮아진 Table 7의 경향과도 부합한다.

Table 10은 상위 4종 모델의 오답과 하위 3종 모델의 대표 오류를 정리한 것이다. GPT-OSS 20B는 순위 조회 2건(N25, N28)에서 과적 혐의 조건을 누락하여 전체 통과 차량을 기준으로 차종별 순위를 산출하였다. Claude Sonnet 4.5(N33)와 GPT-4o (N39)는 비율 연산의 분자에 해당하는 과적 구분을 필터에도 지정하여 분모가 과적 차량으로 한정되었고, 그 결과 모든 그룹의 비율이 1.0으로 산출되었다. GPT-4o(N40)는 연산 종류와 대상 항목은 정확히 추출하였으나 최댓값 1행의 선택을 누락하여 세 유형의 비율을 모두 반환하였다. 하위 3종에서는 조회 조건의 추출 오류가 두드러졌으며, Qwen3-8B와 Gemma3-4B는 조건을 누락하는 경우가, Mistral-7B는 질의에 없는 조건을 추가하는 경우가 많았다.

4.5 프롬프트 구성에 따른 실행 결과 일치율

데이터 스키마와 지원 연산에 관한 구체적인 정보를 제공하지 않은 최소 프롬프트를 적용한 결과, 7종 모델 모두 40개 평가 질의에서 정답으로 판정된 질의가 없었다. 당초 실험에서는 일부 API 호출이 사용량 제한으로 응답을 받지 못하였으므로, 7종 모델 모두 동일한 조건으로 재실행한 결과를 제시한다. 재실행에서는 최대 출력 길이를 스키마 기반 프롬프트 조건보다 큰 4,096토큰으로 설정하였으며, 출력 길이 제한이나 사용량 제한으로 응답이 종료된 사례는 없었다.

최소 프롬프트의 불일치 280건을 규칙 기반으로 분류한 결과, 256건(91.4 %)은 모델이 JSON을 생성하였으나 질의 처리 모듈의 항목명과 값 체계를 따르지 않아 조건이 기준과 달라진 경우였다. 예를 들어 과적 혐의 건수를 묻는 질의(N01)에서 모델은 과적 혐의를 'overload_suspicion'이라는 임의의 항목명으로 표현하였으며, 하행 방향을 스키마의 'DN'이 아닌 'down'으로 표기한 사례도 있었다. 이외에 생성된 조건을 질의 처리 모듈이 실행하지 못한 경우가 18건, 허용되지 않은 항목이나 값으로 스키마 검증을 통과하지 못한 경우가 5건, JSON 형식 오류가 1건이었다. 이러한 사례에서 보듯이 모델이 질의의 의미를 일부 반영하더라도 질의 처리 모듈이 해석할 수 있는 항목명과 값 체계로 표현하지 못하면 조회

결과를 얻을 수 없다.

반면 스키마 기반 프롬프트를 적용한 경우에는 4.1절에 제시한 바와 같이 모델에 따라 15.0~97.5 %의 정답률을 보였다. 이러한 결과는 하이브리드 체계에서 프롬프트가 모델의 출력과 질의 처리 모듈을 연결하는 인터페이스의 역할을 하며, 조회 가능한 데이터 항목과 수행 가능한 연산을 명확하게 제시하는 것이 필수적임을 보여준다. 다만 동일한 스키마 기반 프롬프트를 적용한 경우에도 모델별 정답률에 큰 차이가 나타나, 프롬프트 구성만으로 모델 간 질의 해석 성능의 차이가 해소되지는 않았다.

5. 결 론

본 연구에서는 도로 WIM 운영데이터를 대상으로 LLM이 자연어 질의에서 조회 조건과 연산 정보를 추출하고, 실제 데이터의 필터링·집계·정렬은 별도의 질의 처리 모듈이 수행하는 하이브리드 질의응답 체계를 구축하였다. 5개 현장의 WIM 통과이벤트 20만 건 표본과 40개의 평가 질의를 이용하여 클라우드 모델 3종과 로컬 모델 4종의 실행 결과 일치율을 비교하였으며, 주요 결과는 다음과 같다.

첫째, 스키마 기반 프롬프트에서 클라우드 모델 3종과 GPT-OSS 20B는 87.5~97.5 %의 실행 결과 일치율을 보였고, Qwen3-8B, Gemma3-4B 및 Mistral-7B는 15.0~40.0 %였다. 두 군은 통계적으로 뚜렷하게 구분되었으나 각 군 내부의 차이는 40개 질의 규모에서 구분되지 않았으므로, 본 결과는 모델 간의 세부 순위보다 적용 가능한 모델군을 구분하는 근거로 해석하는 것이 적절하다. 로컬 환경에서 운용한 GPT-OSS 20B가 클라우드 모델과 같은 상위군에 속한 점은 외부 API의 사용이 제한되는 환경에서도 하이브리드 체계를 적용할 수 있는 가능성을 보여준다. 다만 GPT-OSS 20B의 평균 응답 시간은 8.40초로 클라우드 모델의 약 3초보다 길었다.

둘째, 실행 결과 일치의 타당성을 검토한 결과, 기준 결과가 비어 있는 질의는 없었으며, 상위 4종 모델에서 결과는 일치하였으나 조건 표기가 달랐던 32건 중 해석이 실질적으로 다른 사례는 없었다. 하위군에서는 해석이 다르면서 결과만 일치한 사례가 1건 확인되었으나 군의 구분에는 영향이 없었다. 또한 질의 처리 모듈과 독립적인 SQL 실행으로 40개 질의의 기준 결과를 모두 재현하였다. 이는 본 연구의 실행 결과 일치율이 상위군 모델의 조건 추출 정확성을 과대평가하지 않았음을 뒷받침한다.

셋째, 모델 간 성능 차이는 주로 조건 추출 단계에서 나타났다. 하위군 모델은 그룹화 항목은 비교적 정확하게 추출하였으나 필터 조건의 일치율이 17.5~42.5 %에 그쳤으며, 필터 조건이 많아질수록 일치율이 낮아졌다. 오류의 양상도 모델에 따라 달라, Qwen3-8B와 Gemma3-4B는 조건을 누락하는 경우가, Mistral-7B는 질의에 없는 조건을 추가하는 경우가 많았다. 상위군 모델의 오답은 소수였지만, 비율 연산에서 분자 조건을 필터에 중복 지정하여 분모가 달라지거나, 순위 조회에서 핵심 조건을 누락하거나, 최댓값 선택을 생략하는 등 연산의 의미와 조건 간의 관계를 잘못 해석한 경우였다. 또한 내부 추론을 수행하는 모델에서는 추론 과정이 출력 한도를 소비하여 복잡한 질의의 응답이 도중에 종료되는 사례가 확인되었으므로, 모델의 특성에 맞는 출력 길이 설정이 필요하다.

넷째, 데이터 스키마와 지원 연산 정보를 제공하지 않은 최소 프롬프트에서는 7종 모델 모두 실행 결과가 일치한 질의가 없었다. 불일치의 대부분은 모델이 질의 처리 모듈과 다른 항목명과 값 체계를 사용하여 발생하였다. 따라서 하이브리드 체계에서 프롬프트는 모델의 출력과 질의 처리 모듈을 연결하는 인터페이스로서, 조회 가능한 항목과 연산을 명시적으로 제시하는 것이 필수적이다. 다만 동일한 프롬프트에서도 모델별 차이가 크게 나타나 프롬프트 구성만으로 모델 간 격차가 해소되지는 않았다.

본 연구의 평가는 5개 현장과 사전에 구성한 40개 질의를 대상으로 하였으므로 신뢰구간이 넓으며, 실제 사용 과정의 구어체 표현, 오타 및 연속적인 대화 형태의 질의는 포함하지 않았다. 평가 표본은 희소 사례를 추가로 추출하여 구성하였으므로 산출된 건수와 비율은 모집단의 추정치가 아니다. 스키마 기반 프롬프트의 출력 예시 일부는 평가 질의의 기준 조건과 같았으며, 이를 제외한 분석에서 결론은 유지되었으나 하위군 일부 모델의 일치율은 낮아졌다. 또한 비교 모델의 파라미터 규모와 실행 환경이 서로 다르며, 최종 자연어 답변의 품질, 운영 비용, 보안성 및 유지관리성은 평가하지 않았다. 따라서 본 연구의 결과는 동일한 규모의 모델 간 성능 벤치마크나 특정 모델의 실제 운영 적합성에 대한 종합적인 평가로 해석하기에는 한계가 있다.

향후 연구로는 다음을 제안한다. 첫째, 실제 WIM 운영자의 질의 표현과 연속적인 대화를 반영하여 평가 질의를 확대함으로써 모델 간 차이를 보다 정밀하게 구분할 필요가 있다. 둘째, 소형 로컬 모델의 조건 추출 정확도를 높이기 위해 스키마에 따라 출력 형식을 제한하는 구조화 출력, 도메인 예시의 확충 및 미세조정의 효과를 검토할 필요가 있다. 셋째, 응답의 종료 사유와 토큰 사용량을 기록하고 출력이 절단된 경우 재요청하는 등 운영 안정성을 높이는 한편, 규칙 기반 재추출을 복합 연산까지 확장할 필요가 있다. 넷째, 최종 자연어 답변이 질의 처리 결과의 수치를 정확하게 인용하는지를 평가하고, 비용, 보안 및 응답 시간을 포함하여 클라우드 모델과 로컬 모델의 적용 조건을 종합적으로 검토할 필요가 있다. 본 연구에서 제안한 체계는 WIM 운영데이터의 조회와 분석을 지원하기 위한 것으로, 과적 여부에 대한 법적 판정이나 행정처분을 대체하는 것을 목적으로 하지 않는다.

Acknowledgements

This research was funded by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure, and Transport under grant number [RS-2022-00142239].

ORCID

Junyoung Park https://orcid.org/0000-0002-3527-8534
Younbeom Jho https://orcid.org/0000-0003-2638-9467
Jongwoo Kim https://orcid.org/0000-0001-8175-5816

References

Anthropic (2025). Claude Sonnet 4.5 system card, Anthropic.

Bang, G. H., Lee, J. H., Jeon, S. G., and Heo, G. H. (2026). "Extraction of foundational data for database construction in overloaded vehicle monitoring." *KSCE Journal of Civil and Environmental Engineering Research,* Vol. 46, No. 1, pp. 29-38, https://doi.org/10.12652/Ksce.2026.46.1.0029 (in Korean).

Bang, G. H., Lee, J. H., Lee, Y. J., Seo, S. G., and Heo, G. H. (2025). "An analysis of load measurement accuracy of EnBWIM system according to vehicle operation characteristics." *KSCE Journal of Civil and Environmental Engineering Research*, Vol. 45, No. 1, pp. 43-49, https://doi.org/10.12652/Ksce.2025.45.1.0043 (in Korean).

Burnos, P., Gajda, J., Sroka, R., Wasilewska, M., and Dolega, C. (2021). "High accuracy weigh-in-motion systems for direct enforcement." *Sensors,* Vol. 21, No. 23, 8046, https://doi.org/10.3390/s21238046.

Gemini Team, Google (2025). Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, Technical report, Google DeepMind.

Gemma Team, Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Rame, A., Riviere, M., et al. (2025). Gemma 3 technical report, arXiv preprint, arXiv:2503.19786.

Hong, Z., Yuan, Z., Zhang, Q., Chen, H., Dong, J., Huang, F., and Huang, X. (2025). "Next-generation database interfaces: A survey of LLM-based Text-to-SQL." *IEEE Transactions on Knowledge and Data Engineering,* Vol. 37, No. 12, pp. 7328-7345, https://doi.org/10.1109/TKDE.2025.3609486.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Chen, D., Dai, W., Chan, H. S., Madotto, A., and Fung, P. (2023). "Survey of hallucination in natural language generation." *ACM Computing Surveys*, Vol. 55, No. 12, pp. 1-38, https://doi.org/10.1145/3571730.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7B, arXiv preprint, arXiv:2310.06825.

Kampelopoulos, D., Tsanousa, A., Vrochidis, S., and Kompatsiaris, I. (2025). "A review of LLMs and their applications in the architecture, engineering and construction industry." *Artificial Intelligence Review*, Vol. 58, Article 250, https://doi.org/10.1007/s10462-025-11241-7.

Koh, E., Sun, J. W., and Park, K. H. (2026). "Performance validation of a hybrid LLM-based QA system for bridge management." *KSCE Journal of Civil and Environmental Engineering Research,* Vol. 46, No. 3, pp. 259-268, https://doi.org/10.12652/Ksce.2026.46.3.0259 (in Korean).

Lee, D., Lee, J., Cho, J. W., Kim, J. W., and Kim, T. S. (2023). "Analysis of secondary influencing factors for the WIM system used in overload enforcement." *International Journal of Highway Engineering,* Vol. 25, No. 4, pp. 25-35, https://doi.org/10.7855/IJHE.2023.25.4.025 (in Korean).

Lee, J., Jung, W., and Baek, S. (2024). "In-house knowledge management using a large language model: Focusing on technical specification documents review." *Applied Sciences,* Vol. 14, No. 5, 2096, https://doi.org/10.3390/app14052096.

Lee, Y., Kim, C., Lee, J., Bang, G., and Heo, G. (2026). "Dual-stage optimization of a digital twin for bridge load identification." *KSCE Journal of Civil and Environmental Engineering Research,* Vol. 46, No. 1, pp. 11-20, https://doi.org/10.12652/Ksce.2026.46.1.0011 (in Korean).

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). "On faithfulness and factuality in abstractive summarization." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics,* pp. 1906-1919, https://doi.org/10.18653/v1/2020.acl-main.173.

OpenAI (2024). GPT-4o system card, arXiv preprint, arXiv:2410.21276.

OpenAI (2025). gpt-oss-120b & gpt-oss-20b model card, arXiv preprint, arXiv:2508.10925.

Renggli, C., Ilyas, I. F., and Rekatsinas, T. (2025). Fundamental challenges in evaluating Text2SQL solutions and detecting their limitations, arXiv preprint, arXiv:2501.18197.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. (2025). Qwen3 technical report, arXiv preprint, arXiv:2505.09388.

Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., Ma, J., Li, I., Yao, Q., Roman, S., et al. (2018). "Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and Text-to-SQL task." Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, pp. 3911-3921, https://doi.org/10.18653/v1/D18-1425.

Appendix

Table A1. Evaluation Query List

ID	Type	SD	Query (translated from Korean)
N01	A	SD1	How many overweight-suspicion vehicles are in the full dataset?
N02	A	SD1	How many vehicles have a dimension-sensor anomaly (non-positive dimensions)?
N03	A	SD2	At MG01, how many passages remain after excluding class 1?
N04	A	SD2	How many passages have GVW of at least 35 t and less than 40 t?
N05	A	SD3	How many overweight-suspicion events occurred at BD01 upstream?
N06	A	SD3	How many passages were recorded at MG01 lane 2 in December 2025?
N07	A	SD3	At MG01, how many overweight-suspicion vehicles passed below 80 km/h?
N08	A	SD4	How many class-3 vehicles passed AB01 downstream in January 2026?
N09	A	SD4	How many intentional-overweight events occurred at BD01 upstream in October 2025?
N10	A	SD4	At HJ01 lane 1, how many events have GVW of at least 40 t and width oversize?
N11	B	SD1	Report passage counts by site.
N12	B	SD1	What is the mean speed by site?
N13	B	SD3	By vehicle class, how many dimension-sensor anomalies occurred at BD01 in October 2025?
N14	B	SD2	What is the mean speed by lane for MG01 upstream?
N15	B	SD3	What are the passage counts by vehicle class at MG01 upstream in October 2025?
N16	B	SD3	What are the counts by vehicle class of overweight-suspicion vehicles at BD01 upstream?
N17	B	SD3	What are the monthly passage counts of class-5-or-higher vehicles at AB01 downstream?
N18	B	SD4	By vehicle class, what are the overweight-suspicion counts at HJ02 downstream in the second half of 2025?
N19	B	SD4	By lane, how many axle-only overweight events occurred at BD01 upstream in October 2025?
N20	B	SD4	What is the monthly mean GVW of class-6 vehicles at MG01 upstream in 2026 Q1?
N21	C	SD1	Rank sites by overweight-suspicion count.
N22	C	SD1	Rank oversize types (width, height, length) by count.
N23	C	SD2	Rank vehicle classes by overweight-suspicion count at BD01.
N24	C	SD2	Rank sites by intentional-overweight count in October 2025.
N25	C	SD3	Rank vehicle classes by overweight-suspicion count at BD01 upstream.
N26	C	SD3	Rank sites by axle-only overweight count for upstream traffic in October 2025.
N27	C	SD3	Rank lanes by passage count at MG01 upstream in December 2025.
N28	C	SD4	Rank vehicle classes by overweight-suspicion count at BD01 upstream in October 2025.
N29	C	SD4	Rank sites by count of downstream vehicles with GVW of at least 40 t and class 5 or higher in the second half of 2025.
N30	C	SD4	Rank vehicle classes by width-oversize count at AB01 downstream in January 2026.
N31	D	SD1	For each vehicle class, which axle most often exceeds the limit?
N32	D	SD1	What is the overweight-suspicion rate by site (suspicion count / passage count)?
N33	D	SD2	What is the overweight-suspicion rate by vehicle class for upstream traffic in October 2025?
N34	D	SD2	For axle-overweight vehicles at BD01, which axle is exceeded most often by class?
N35	D	SD3	For each class, which axle is exceeded most often at MG01 upstream in October 2025?
N36	D	SD3	Which vehicle class has the highest overweight-suspicion rate at BD01 upstream in the second half of 2025?
N37	D	SD3	What is the overweight-suspicion rate by lane at MG01 downstream in December 2025?
N38	D	SD4	Which axle is exceeded most often among class-6 vehicles at BD01 upstream in October 2025?
N39	D	SD4	Rank classes 5 and higher by intentional-overweight rate at HJ01 upstream in the second half of 2025.
N40	D	SD4	Among width, height, and length oversize at AB01 downstream lane 1 in 2026 Q1, which type is most frequent and what is its rate?

Table A2. Prompt Texts

Prompt	Text
Minimal Prompt	SYSTEM: | 당신은 질의응답용 의도 추출기입니다. 사용자 질의를 JSON 객체 하나로만 변환하세요. 수치를 계산하거나 추정하지 마세요. JSON만 출력하세요 (설명·마크다운 금지). USER_TEMPLATE: | 질의: {query}
Schema-based Prompt	SYSTEM: | 당신은 화물차 운송안전·WIM 운영데이터 질의 의도 추출기입니다. 자연어 질의를 JSON 객체 하나로만 변환하세요. 수치 계산·단속 판정 금지. JSON만 출력(설명/마크다운 금지). 필수 규칙: 1) site_id, lane, direction, date_range, vehicle_class, vehicle_class_exclude, vehicle_class_min, speed_kmh, gvw_kg, is_dim_anomaly, is_oversize, is_oversize_width, is_oversize_height, is_oversize_length, is_overweight_enforce, is_overweight_intentional, is_overweight_axle_only, is_overweight_gvw_only, is_overweight_both, is_axle_overweight 는 반드시 filters 객체 안에만 둡니다. 최상단에 두지 마세요. 2) site_id / direction / lane / vehicle_class 는 항상 배열입니다. 예: "site_id": ["MG01"], "direction": ["UP"], "lane": [2], "vehicle_class": ["5","6"] 3) direction 값은 오직 "UP" 또는 "DN" (상행=UP, 하행=DN). 4) 건수 질의는 agg.op="count". 평균은 agg.op="mean" 과 column. 플래그·수치 합계는 agg.op="sum" 과 column (예: is_dim_anomaly). 5) 월 구간 필터는 date_range 객체: {"start":"YYYY-MM-01","end":"YYYY-MM-말일"}. date_range는 필터이고, 월별 그룹화 키는 year_month이다(서로 다름). 6) "미만" 속도는 speed_kmh.max_exclusive 사용 (max 말고). "미만" 총중량은 gvw_kg.max_exclusive 사용. 7) 그룹·순위: - group_by 유효 키: site_id, vehicle_class, direction, lane, year_month, label. - 사이트/차종/방향/차로/월별 건수·평균: agg.group_by + sort.by="value" (필요 시). - 질문에 "상위 N"/"Top N"이 **있을 때만** top_k=N. 없으면 top_k를 넣지 마세요(전체 그룹 반환). - 행 단위 상위(가장 무거운 차량 등): sort.by="gvw_kg", ascending=false, top_k. - 제원초과를 너비·높이·길이로 나눠 건수 비교: agg.op="count_labeled_flags", flags=[{"label":"width","column":"is_oversize_width"},{"label":"height","column":"is_oversize_height"},{"label":"length","column":"is_oversize_length"}] - 제원초과를 너비·높이·길이로 나눠 비율 비교: agg.op="ratio_labeled_flags", flags는 count_labeled_flags와 동일. 행마다 label, value(=분자/분모), numerator, denominator 반환. - 그룹별 플래그 비율: agg.op="ratio", agg.flag(예: is_overweight_enforce), agg.group_by. 분자=flag true 건수, 분모=그룹 행 수. 행마다 value, numerator, denominator 반환. - 그룹 결과에서 최댓값 1행만 필요할 때: agg.reduce="argmax" (value 기준). - 차종별 과적 축 번호: filters.is_axle_overweight=true, agg.op="dominant_axle_by_class" 8) 제원 센서이상/치수≤0 → is_dim_anomaly=true (is_oversize 아님). 제원초과(폭·높이·길이 법정) → is_oversize=true. 너비만/높이만/길이만 → is_oversize_width / is_oversize_height / is_oversize_length. 9) 중량 플래그 용어(중요): - "혐의차량", "과적 혐의", "운행제한 위반 혐의", "과적판정" → is_overweight_enforce (축>11000kg 또는 총중>44000kg). intentional 로 쓰지 마세요. - "고의과적"만 → is_overweight_intentional (축>13000 또는 총중>46000). - 축중만/총중만/모두초과 → is_overweight_axle_only / is_overweight_gvw_only / is_overweight_both - 1종 제외 → vehicle_class_exclude=["1"] - 4종 이상 → vehicle_class_min=4 site_id: HJ01(행주북단), HJ02(행주남단), AB01(안보), MG01(문경), BD01(봉동) CV/ER/drive_code 필드는 사용하지 마세요.

Prompt	Text
	예시: {"query_type":"facility","filters":{"site_id":["MG01"]},"agg":{"op":"count"}} {"query_type":"aggregate","filters":{"direction":["UP"]},"agg":{"op":"count"}} {"query_type":"filter","filters":{"site_id":["HJ01"],"direction":["DN"]},"agg":{"op":"count"}} {"query_type":"aggregate","filters":{"site_id":["MG01"]},"agg":{"op":"mean","column":"speed_kmh"}} {"query_type":"ranking","filters":{"is_overweight_enforce":true},"sort":{"by":"gvw_kg","ascending":false},"top_k":5,"select":["event_id","site_id","gvw_kg"]} {"query_type":"data_quality","filters":{"is_dim_anomaly":true},"agg":{"op":"count"}} {"query_type":"aggregate","filters":{},"agg":{"op":"count","group_by":["site_id"]},"sort":{"by":"value","ascending":false}} {"query_type":"ranking","filters":{"is_overweight_enforce":true},"agg":{"op":"count","group_by":["site_id"]},"sort":{"by":"value","ascending":false},"select":["site_id","value"]} {"query_type":"ranking","filters":{},"agg":{"op":"count_labeled_flags","flags":[{"label":"width","column":"is_oversize_width"},{"label":"height","column":"is_oversize_height"},{"label":"length","column":"is_oversize_length"}]},"sort":{"by":"value","ascending":false},"select":["label","value"]} {"query_type":"ranking","filters":{"is_axle_overweight":true},"agg":{"op":"dominant_axle_by_class","axle_limit_kg":11000},"select":["vehicle_class","axle","value"]} {"query_type":"aggregate","filters":{"site_id":["HJ02"]},"agg":{"op":"sum","column":"is_dim_anomaly","group_by":["lane"]}} {"query_type":"aggregate","filters":{"site_id":["HJ01"],"direction":["UP"]},"agg":{"op":"count","group_by":["year_month"]}} {"query_type":"ranking","filters":{"direction":["DN"]},"agg":{"op":"ratio","flag":"is_overweight_enforce","group_by":["site_id"]},"sort":{"by":"value","ascending":false},"select":["site_id","value","numerator","denominator"]} {"query_type":"ranking","filters":{},"agg":{"op":"ratio_labeled_flags","flags":[{"label":"width","column":"is_oversize_width"},{"label":"height","column":"is_oversize_height"},{"label":"length","column":"is_oversize_length"}]},"sort":{"by":"value","ascending":false},"select":["label","value","numerator","denominator"]} {"query_type":"ranking","filters":{"direction":["UP"]},"agg":{"op":"ratio","flag":"is_overweight_enforce","group_by":["direction"],"reduce":"argmax"},"select":["direction","value","numerator","denominator"]} USER_TEMPLATE: | filters 키만 사용: site_id[], lane[], direction[UP|DN], date_range{start,end}, vehicle_class[], vehicle_class_exclude[], vehicle_class_min, speed_kmh{min,max,max_exclusive}, gvw_kg{min,max,max_exclusive}, is_dim_anomaly, is_oversize, is_oversize_width, is_oversize_height, is_oversize_length, is_overweight_enforce, is_overweight_intentional, is_overweight_axle_only, is_overweight_gvw_only, is_overweight_both, is_axle_overweight agg: {op, column?, group_by[]?, flags[]?, flag?, reduce?, axle_limit_kg?} group_by 유효 키: site_id, vehicle_class, direction, lane, year_month, label (date_range=필터 기간, year_month=월 그룹 키) ranking: sort{by, ascending}, top_k(상위N일 때만), select[] 용어: 혐의/과적혐의=is_overweight_enforce, 고의과적=is_overweight_intentional 질의: {query}

Table A3. Comparison of the Evaluation Sample (n=200,000)

Axis	Category	Records	Share (%)
Site	HJ01	83,601	41.8
	HJ02	40,946	20.5
	AB01	29,931	15.0
	MG01	25,117	12.6
	BD01	20,405	10.2
Month	2025-07	4,589	2.3
	2025-08	12,891	6.4
	2025-09	13,385	6.7
	2025-10	20,367	10.2
	2025-11	20,344	10.2
	2025-12	16,482	8.2
	2026-01	19,463	9.7
	2026-02	17,707	8.9
	2026-03	20,702	10.4
	2026-04	23,142	11.6
	2026-05	23,064	11.5