

A Comparison of NL2GeoSQL Execution Accuracy Across Metadata Configurations: A Case Study of Crime Vulnerability Grid Database

Ki-Sung Song*

*Officer, Spatial Information Division, Seoul Metropolitan Government, Seoul, Korea

[Abstract]

Spatial data are continuously used in policy decision-making, but they are difficult for non-experts to use because spatial relationships and location-based operations must be considered together. Natural language spatial query technologies can serve as a technical alternative that enables users to directly specify required analytical conditions beyond predefined dashboards or prewritten query statements. This study compares the effects of metadata configurations on NL2GeoSQL execution accuracy using Seoul's crime-prevention decision support data. To this end, a PostGIS-based spatial database and an evaluation set of 30 Korean natural-language spatial queries were constructed, and SQL statements generated by two code-focused large language models under five metadata conditions were executed and evaluated. The results show that the SQL-to-text example condition achieved the highest execution accuracy and also performed relatively well in schema linking and spatial function use. This study demonstrates that metadata configuration affects GeoSQL generation performance and that SQL-to-text examples can be used as execution-oriented knowledge for GeoSQL generation.

▶ **Key words:** NL2GeoSQL, Metadata, Spatial Database, Large Language Model(LLM), Execution Accuracy

[요 약]

공간데이터는 정책 의사결정에 지속적으로 활용되지만, 공간관계와 위치 기반 연산을 함께 고려해야 하므로 비전문가가 활용하기 어렵다. 최근 논의되고 있는 자연어 공간질의 기술은 대시보드나 사전 작성된 조회 쿼리의 범위를 넘어 필요한 분석 조건을 직접 질의할 수 있게 하는 기술적 대안이 될 수 있다. 본 연구는 서울시 범죄예방 의사결정지원 데이터를 대상으로 메타데이터 구성 조건이 NL2GeoSQL 실행정확도에 미치는 영향을 비교하였다. 이를 위해 PostGIS 기반 공간DB와 30개 자연어 공간질의 평가셋을 구축하고, 다섯 가지 메타데이터 조건에서 대규모 언어모델이 생성한 SQL을 실행·평가하였다. 분석 결과, SQL-to-text 예시 조건에서 실행정확도가 가장 높았고, 스키마 연결과 공간함수 사용에서도 비교적 우수하였다. 본 연구는 메타데이터 구성 방식이 GeoSQL 생성 성능에 영향을 미치며, SQL-to-text 예시가 실행 지식으로 활용될 수 있음을 보여준다.

▶ **주제어:** 자연어 공간질의 변환, 메타데이터, 공간 데이터베이스, 대규모 언어모델, 실행정확도

-
- First Author: Ki-Sung Song, Corresponding Author: Ki-Sung Song
 - *Ki-Sung Song (ks.onemap@gmail.com), Spatial Information Division, Seoul Metropolitan Government
 - Received: 2026. 07. 14, Revised: 2026. 07. 27, Accepted: 2026. 08. 07.

I. Introduction

1. Background

최근 대규모 언어모델(Large Language Model, LLM)의 발전은 자연어 기반 데이터 질의의 가능성을 높이고 있다. 사용자가 SQL 작성 없이 자연어 질문만으로 데이터베이스에서 필요한 정보를 조회할 수 있다면, 데이터 활용의 기술적 장벽을 낮추고 다양한 분석 요구에 유연하게 대응할 수 있다. Text2SQL 연구에서도 LLM 기반 접근은 스키마 연결, 질의 분해, SQL 생성, 자기수정 등의 절차를 통해 실행 가능한 SQL 생성 성능을 높여 왔다[1]. 또한 GIS 분야에서도 자연어 지시를 공간데이터 수집, 처리, 분석 과정과 연결하려는 연구가 진행되면서, 비전문가가 공간분석 기능에 접근할 수 있는 가능성이 논의되고 있다[2].

이러한 가능성은 공공행정 분야의 공간 데이터베이스 활용에도 적용될 수 있다. 방법, 교통, 안전, 복지, 환경 등 정책 의사결정 과정에서는 격자, 행정경계, 보호구역과 같은 공간 데이터가 지속적으로 활용되지만, 공간 데이터는 일반 관계형 데이터보다 구조가 복잡하고 공간관계와 위치 기반 연산을 함께 고려해야 하기 때문에, 공공행정 실무자가 쉽게 접근하기 어렵다. 따라서 자연어 공간질의 기술은 공공행정 실무자가 정형화된 대시보드나 사전 작성된 조회 쿼리의 범위를 넘어 필요한 분석 조건을 직접 질의할 수 있게 하는 기술적 대안이 될 수 있다.

다만 PostgreSQL에 공간 데이터 처리 기능을 확장한 PostGIS 기반 NL2GeoSQL은 공간 데이터 유형, 공간함수 호출, 좌표계 해석을 포함하므로 일반 NL2SQL보다 생성과 실행 난이도가 높다[3]. 정확한 GeoSQL 생성을 위해서는 LLM이 데이터베이스 구조와 질의 맥락을 함께 해석해야 한다. Text2SQL 분야에서는 데이터베이스 내용 이해가 SQL 작성의 핵심 난점으로 논의되어 왔으며, 프로파일링, 질의 로그, SQL-to-text 예시 등은 스키마와 질의 맥락을 보완하는 메타데이터로 활용되어 왔다[4].

공간 데이터베이스에서는 일반 스키마 정보 외에도 geometry 타입, 좌표계, 공간함수, 공간관계와 같은 공간적 맥락이 GeoSQL 생성에 영향을 줄 수 있다. 따라서 자연어 공간질의 시스템을 실제 행정 데이터베이스에 적용하기 위해서는 단순히 LLM이 SQL을 생성할 수 있는지를 확인하는 수준을 넘어, 어떤 메타데이터를 제공할 때 실행 가능한 GeoSQL이 더 정확하게 생성되는지를 검증할 필요가 있다. 본 연구는 이러한 문제의식에서 출발하여, 실제 공공행정 환경에서 발생하는 자연어 공간질의의 특성을 반영하기 위해 서울시가 범죄예방 및 생활안전 정책 추진

과정에서 실제로 활용하고 있는 공간데이터를 기반으로 공간데이터베이스를 구축하고, 메타데이터 구성 조건에 따른 NL2GeoSQL 실행정확도 차이를 비교하고자 한다[5].

2. Objectives

본 연구의 목적은 PostGIS 기반 서울시 방법취약격자 데이터베이스를 대상으로, 메타데이터 구성 조건에 따라 LLM의 NL2GeoSQL 실행정확도가 어떻게 달라지는지를 정량적으로 비교하는 것이다. 이를 위해 방법취약격자, CCTV, 비상벨, 보안등, 어린이보호구역, 행정경계 데이터를 활용하여 공간 데이터베이스를 구축하고, 30개 자연어 공간질의 평가셋을 설계하였다. 또한 동일한 데이터베이스, 평가질의, 모델, 실행-평가 절차를 적용하여 메타데이터 구성 조건별 NL2GeoSQL 성능 차이를 분석하였다.

3. Contributions

본 연구의 기여는 다음 세 가지로 요약된다.

첫째, PostGIS 기반 자연어 공간질의를 대상으로 메타데이터 제공 방식에 따른 NL2GeoSQL 성능 비교 프레임워크를 제시한다.

둘째, 서울시 방법취약격자 데이터베이스와 30개 자연어 공간질의 평가셋을 이용하여 조건별 NL2GeoSQL 성능을 정량적으로 평가한다.

셋째, 생성 SQL의 오류유형을 분석하여 메타데이터가 LLM의 공간질의 해석과 GeoSQL 생성 과정에서 수행하는 역할을 실증적으로 제시한다.

II. Preliminaries

1. Natural Language Data Query and NL2SQL

자연어 기반 데이터 질의는 사용자의 자연어 질문을 데이터베이스 질의로 변환하여 필요한 결과를 조회하는 기술이다. 이 중 NL2SQL은 관계형 데이터베이스에 대한 비전문가의 접근성을 높이는 대표적 접근으로 발전해 왔으며, Spider와 BIRD 같은 벤치마크를 통해 복잡한 SQL, 교차 도메인 스키마, 실제 데이터베이스 기반 질의 성능을 평가하는 방향으로 확장되어 왔다[6, 7].

다만 NL2SQL 성능은 데이터베이스 스키마의 복잡성, 프롬프트 구성, 실행 환경, 평가 지표에 따라 달라질 수 있다. 특히 스키마 연결, 조인 경로, 조건 해석, 실행 전 검증은 실제 데이터베이스 적용에서 중요한 요소로 논의되어 왔으며, 관련 연구는 스키마 연결과 SQL 구조 생성을 분

리하거나 관련 스키마와 값을 선별하여 생성 성능을 높이는 방향으로 발전하고 있다[7-10]. 이러한 흐름은 자연어 질의 기술이 단순 SQL 생성 성능을 넘어, 데이터베이스 구조 이해와 실행 신뢰성 확보까지 함께 고려하는 방향으로 확장되고 있음을 보여준다.

2. Metadata for Text2SQL

Text2SQL에서 메타데이터는 자연어 표현과 데이터베이스 스키마 사이의 의미적 간극을 줄이는 정보 계층으로 활용된다. 특히 실제 데이터베이스 기반 Text2SQL 평가에서는 테이블·컬럼 구조뿐 아니라 데이터 값, 외부 지식, 업무 맥락이 SQL 생성 성능에 영향을 미치는 요소로 논의되어 왔다[7]. 관련 연구의 메타데이터는 구조 메타데이터, 의미 메타데이터, 통계 메타데이터, 사용 기반 메타데이터로 구분할 수 있다. 구조 메타데이터는 스키마와 제약조건, 의미 메타데이터는 테이블·컬럼 설명과 도메인 지식, 통계 메타데이터는 값 분포와 샘플, 사용 기반 메타데이터는 질의 로그와 SQL-to-text 예시를 포함한다. 이러한 메타데이터는 스키마 연결, 값 해석, 조인 경로 추론, 업무 로직 반영을 보완하는 정보로 논의되어 왔다[4, 11, 12].

최근 연구는 메타데이터의 효과가 단순히 제공 여부에 따라 결정되는 것이 아니라, 상세도와 구성 방식에 따라 달라질 수 있음을 보여준다. 특히 자동 생성 컬럼 설명 연구에서는 사람이 불필요하다고 판단한 세부 설명도 모델의 SQL 생성 정확도를 높일 수 있음이 보고되었다[12]. 또한 대규모 데이터베이스 환경에서는 모든 스키마와 값을 그대로 제공하기보다, 질의와 관련된 스키마와 값 정보를 선별하여 프롬프트에 구성하는 방식이 SQL 생성의 효율성과 실행 가능성을 높이는 전략으로 제시되고 있다[9].

3. NL2GeoSQL and Spatial Query Evaluation

NL2GeoSQL은 자연어 질문을 공간 데이터베이스에서 실행 가능한 GeoSQL로 변환하는 과제이다. GeoSQL은 공간 데이터 유형, 공간함수, 공간 인덱스, 공간 연산자, 좌표계, 위상관계 해석을 포함하므로 일반 NL2SQL보다 생성과 실행 과정이 복잡하다. 이에 따라 기존 평가 연구에서는 함수 이해, 매개변수 매칭, 스키마 연결, 구문 유효성, 실행 결과 검증이 주요 평가 요소로 제시되고 있다[3].

지리공간 질의응답 벤치마크 연구에서도 유사한 한계가 확인된다. GS-QA는 복합 공간술어, 수치형 결과, 다중 출처 추론에서 기존 LLM 기반 접근의 성능이 저하될 수 있음을 보였고, MapQA는 거리 계산과 다중 홑 공간 추론이 주요 병목임을 제시하였다[13, 14]. 이는 자연어 공간질의

평가에서 단순 정답 일치뿐 아니라 공간연산의 적절성, 공간함수 사용, 결과 단위의 정확성을 함께 고려해야 함을 시사한다.

4. Distinctiveness of This Study

기존 연구는 LLM 기반 NL2SQL 성능 개선, 데이터베이스 메타데이터 생성, 자연어 공간질의 평가체계 구축에 기여하였다. 그러나 PostGIS 기반 공간 데이터베이스에서 메타데이터 제공 방식이 NL2GeoSQL 실행정확도에 미치는 영향을 동일한 데이터베이스와 동일한 평가질의 조건에서 비교한 연구는 찾아보기 어렵다.

본 연구는 서울시 방법취약격자 데이터베이스를 대상으로 메타데이터 구성 조건별 NL2GeoSQL 실행정확도를 비교하여, 메타데이터가 LLM의 공간질의 해석과 GeoSQL 생성 과정에서 수행하는 역할을 실증적으로 확인한다.

III. Problem Definition & Design Goals

1. Problem Definition

본 연구는 자연어 공간질의를 입력으로 받아 PostGIS 데이터베이스에서 실행 가능한 GeoSQL을 생성하고, 그 실행 결과가 기존 SQL의 실행 결과와 일치하는지를 평가한다. 이 과정에서 해결해야 할 핵심 문제는 세 가지이다.

- (1) “고위험”, “부족”, “없음”, “어린이보호구역 포함”과 같은 자연어 업무 표현을 SQL 조건으로 변환해야 한다.
- (2) 질의 의도에 따라 파생 분석테이블과 원천 공간테이블 중 적절한 테이블을 선택해야 한다.
- (3) 중첩, 거리, 면적, 행정구역 집계와 같은 공간관계를 적절한 PostGIS 함수와 조인 구조로 변환해야 한다.

2. NL2GeoSQL Execution Framework

본 연구의 전체 실험 구조는 Fig. 1과 같다. 실행 절차는 한국어 자연어 평가질의 입력, 공통 질의 의미 정의와 조건별 메타데이터를 결합한 프롬프트 구성, LLM 기반 GeoSQL 생성, 생성 SQL 추출·정제, PostgreSQL/PostGIS 실행, 기존 SQL 실행 결과와의 비교평가, 오류유형 분석의 순서로 구성하였다. 30개의 한국어 자연어 평가질의에 대해 C0~C4 조건별 메타데이터 프롬프트를 구성하였으며, 모든 조건에는 동일한 평가질의, 공통 질의 의미 정의, SQL 생성 지침을 적용하였다. 공통 질의 의미 정의에는 ‘고위험’ 또는 ‘방법취약지수가 높은’을 `vuln_grade IN (4,5)`, ‘5등급’을 `vuln_grade=5`, ‘없음’을 해당 시설 수가 0인 경우, ‘부족’을

최종 분석격자 48,962개의 평균보다 작은 경우로 정의하였다. ‘모두 부족’은 관련 조건의 AND, ‘하나 이상 부족’은 OR 결합으로 해석하였다. 또한 평균값은 프롬프트에 제공된 값 또는 SQL 계산을 통해 처리하도록 하였으며, 평가대상 질의의 기준 SQL, 기준 결과, 정답값은 프롬프트에서 제외하였다. 조건 간 차이는 프롬프트에 제공되는 메타데이터 범위로 제한하였다. 생성된 SQL은 동일한 LLM 설정에서 생성한 뒤 PostgreSQL/PostGIS에서 실행하였으며, 사전에 작성한 기준 SQL 및 기준 실행 결과와 비교하여 SQL 실행 가능성, 실행정확도, 결과 일치도, 스키마 연결 정확도, 공간함수 출현 정확도를 평가하였다.

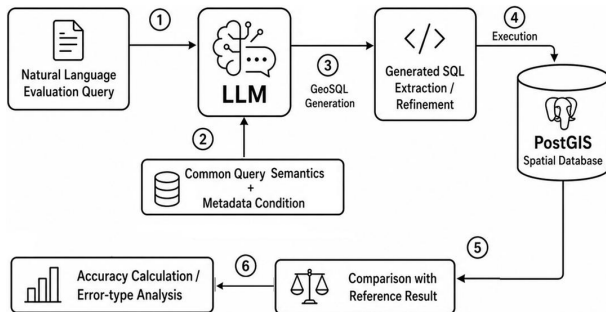


Fig. 1. NL2GeoSQL-based Spatial Analysis Workflow

3. Metadata Conditions

II장에서 논의한 기존연구를 바탕으로, 본 연구는 Text-to-SQL에서 활용되는 메타데이터를 구조 메타데이터, 통계·프로파일 메타데이터, 사용 기반 메타데이터로 정리하고, PostGIS 기반 GeoSQL의 특성을 반영하여 공간 메타데이터를 추가하였다. 구조 메타데이터는 테이블명, 컬럼명, 자료형 등 기본 스키마 정보를 의미하며, 공간 메타데이터는 geometry 컬럼, 좌표계, 공간 단위, 권장 PostGIS 함수 정보를 포함한다. 통계·프로파일 메타데이터는 컬럼 값의 분포와 주요 통계값을 제공하며, 사용 기반 메타데이터는 테이블 선택, 조인 경로, 공간함수 사용, 집계 단위 등 기존 질의 작성 패턴을 전달하는 SQL-to-text 예시를 의미한다.

본 연구의 C0~C4 조건은 이러한 메타데이터 유형을 실험적으로 비교하기 위해 단계적으로 구성하였다. C0은 구조 메타데이터 중 비공간 스키마만 제공한 기준조건이며, C1은 C0에 공간 메타데이터를 추가한 조건이다. C2는 C1에 통계·프로파일 메타데이터를 추가한 조건이고, C3은 C1에 사용 기반 메타데이터를 추가한 조건이다. C4는 C2와 C3를 결합하여 공간 메타데이터, 통계·프로파일 메타데이터, 사용 기반 메타데이터를 함께 제공한 통합 조건이

다. 따라서 C0~C4는 서로 독립적인 메타데이터 유형이 아니라, 기본 스키마 정보에서 출발하여 GeoSQL 생성에 필요한 메타데이터를 단계적으로 확장한 조건이다.

C3과 C4에 포함된 SQL-to-text 예시는 사용 기반 메타데이터로 구성하였다. SQL-to-text 예시는 후보 예시를 작성하고 PostGIS 실행 가능성을 검증한 후, 평가질의와 직접 중복되거나 의미적 중복이 높은 예시를 제외하여 최종 12개를 사용하였다. 최종 예시는 파생 분석테이블 질의와 원천 공간테이블 질의가 각각 6개씩 포함되도록 구성하였으며, 평균 비교, 다중조건 필터, 행정구역별 집계, 순위 산출, 거리 조건, Point-Polygon 공간조인 등 대표적인 SQL 작성 패턴을 포함하였다. 평가질의와 직접 중복되는 자연어 문장이나 기준 SQL은 포함하지 않았으며, 대상 시설, 조건값, 거리 기준, 집계 단위 또는 결과 단위가 달라지도록 구성하였다.

조건별 프롬프트의 평균 길이와 입력 토큰 수는 Table 1에 함께 제시하였다.

Table 1. Metadata Conditions

Cond.	Metadata scope	Purpose	Avg. length	Mean input tokens
C0	Non-spatial schema + common query semantics	Baseline under common controls	4,114	1,041
C1	C0 + spatial metadata	Effect of basic spatial metadata	5,902	1,651
C2	C1 + value profile	Effect of value distribution metadata	14,997	5,493
C3	C1 + SQL-to-text examples	Effect of query-pattern examples	30,027	12,370
C4	C1 + value profile + SQL-to-text examples	Effect of integrated metadata	39,143	16,212

Note. Mean input tokens were calculated from the usage.prompt_tokens values returned by Qwen2.5-Coder across 30 evaluation queries. Common Query Semantics and SQL generation rules were applied consistently across all conditions.

4. Design Goals

본 연구의 설계 목표는 Table 1에서 제시한 다섯 메타데이터 조건을 동일한 자연어 질의, 데이터베이스, 언어모델, 실행·평가 절차에서 비교하여 NL2GeoSQL 성능 차이를 확인하는 것이다.

(1) 공간 메타데이터 효과 평가: C0과 C1을 비교하여

공간 기본 메타데이터가 NL2GeoSQL 생성에 미치는 영향을 평가

- (2) 메타데이터 유형별 효과 평가: C2와 C3를 통해 값 분포 프로파일과 SQL-to-text 예시가 SQL 생성 결과에 미치는 영향을 비교[4,12].
- (3) 통합 메타데이터 효과 평가: 값 분포 프로파일과 SQL-to-text 예시를 결합한 C4가 실행정확도와 오류 유형에 미치는 영향을 분석
- (4) 실행 결과 중심 평가: 생성 SQL의 실행 가능성, 기준 결과와의 일치도, 공간함수 사용, 스키마 연결, 오류 유형을 기준으로 PostGIS 질의 실행 성능을 평가[3].

IV. Experimental Setup

1. Dataset and Evaluation Questions

본 연구는 서울시 방범취약격자와 관련 공간데이터를 PostgreSQL/PostGIS 데이터베이스로 구축하여 실험에 사용하였다. 좌표계는 EPSG:5186이며, 좌표 단위는 meter로 설정하였다. 주요 분석 단위는 100m 방범취약격자이며, 원천 격자 49,577개 중 행정구역 매칭이 누락된 615개를 제외한 48,962개 격자를 최종 분석 테이블로 구성하였다. 연구 자료는 2026년 4월 서울특별시에서 운영하는 범죄예방업무 지원서비스 데이터베이스에서 일괄 제공받았으며, 각 원천 자료의 제공기관과 자료명은 참고문헌에 제시하였다 [15-20]. 데이터 테이블의 구성은 Table 2와 같다.

평가질의는 작성된 초안을 서울특별시 범죄예방인프라 구축사업 담당자 및 범죄예방업무지원서비스 담당자 2인의 질의응답형 인터뷰를 거쳐 수정하였다. 검토 과정에서는 범죄예방 업무 활용 가능성, 질의 표현의 명확성을 확인하였다. 최종 평가질의는 30개의 한국어 자연어 공간질의로 구성하였으며, 파생 분석테이블을 중심으로 하는 질의 15개와 원천 공간테이블의 공간관계를 활용하는 질의 15개를 포함하였다. 또한 속성 필터, 집계, 행정구역 집계, 순위, 공간관계, 거리, 면적·중첩 비율 등 주요 NL2GeoSQL 질의 유형을 포함하도록 설계하였다. 각 평가질의에는 Table 3과 같이 기준 SQL과 기준 실행 결과를 작성하였으며, PostGIS에서 실행 가능성과 결과값을 확인한 후 실험에 사용하였다. 실제 입력 질의에는 “방범 취약지수가 높은 지역 중 CCTV, 비상벨, 보안등이 모두 평균보다 낮은 격자를 찾아줘”, “어린이보호구역 경계와 겹치는 고위험 방범취약격자를 찾아줘” 등이 포함되었다.

Table 2. Data Tables Used in the Experiment

Table	Geometry	Role
analysis.crime_grid_analysis	Multi-polygon	Final 100m grid analysis table
raw.crime_grid	Multi-polygon	Source crime-vulnerability grid
raw.cctv_point	Point	CCTV locations
raw.bell_point	Point	Emergency bell locations
raw.light_point	Point	Security light locations
raw.school_zone_polygon	Multi-polygon	School-zone polygons
raw.adm_dong_boundary	Multi-polygon	Administrative-dong boundaries
raw.gu_boundary	Multi-polygon	District boundaries

Table 3. Example of an Evaluation Question, Reference SQL, and Reference Execution Result

Natural language question	
Show me the high-risk grid cells that overlap with school zones.	
↓	
Reference SQL	
<pre>SELECT DISTINCT g.grid_id, g.vuln_grade FROM raw.crime_grid g JOIN analysis.crime_grid_analysis a ON a.grid_id = g.grid_id JOIN raw.school_zone_polygon z ON ST_Intersects(g.geom, z.geom) WHERE g.vuln_grade IN (4, 5) ORDER BY g.grid_id;</pre>	
↓	
Reference execution result	
Result unit	Distinct grid cells
Row count	3,906
Output columns	grid_id, vuln_grade
Result summary	Grade 4: 1,820 cells; Grade 5: 2,086 cells

Note. The complete 3,906-row result set was retained as the reference result; this table reports its schema and summary counts.

2. Model and Environment

본 연구는 Qwen2.5-Coder-32B-Instruct를 주 모델로 사용하고, 메타데이터 조건별 성능 변화의 모델 종속성을 점검하기 위해 Devstral-Small-2-24B-Instruct-2512를 추가 모델로 적용하였다. Qwen2.5-Coder는 코드 생성, 완성, 추론 및 수정에 특화된 code-specific LLM이며 [21], Devstral 모델은 코드베이스 탐색, 다중 파일 수정 및 도구 사용을 포함한 agentic software engineering 작업에 특화된 모델이다[22]. 생성 설정은 temperature 0, top_p 1, 최대 출력 토큰 1,024, context length

32,768로 고정하였으며, 세부 실험환경은 Table 4에 제시하였다.

Table 4. Test Environment

Category	Setup
LLM	Qwen2.5-Coder-32B-Instruct; Devstral-Small-2-24B-Instruct-2512
Backend	LM Studio local server
API type	OpenAI-compatible API
SQL execution timeout	60 seconds
Database	PostgreSQL/PostGIS
GPU	Intel Arc 140V
RAM	32 GB
CPU	Intel Core Ultra 7 258V

3. Comparison Axes

조건별 메타데이터 효과를 체계적으로 비교하기 위해, 앞 절에서 정의한 C0~C4 조건은 Table 5의 비교축에 따라 분석하였다.

Table 5. Comparison Axes

Cond.	Comparison	Interpretation
C0→C1	Spatial metadata effect	Effect of basic spatial metadata
C1→C2	Value profile effect	Additional effect of value distribution metadata
C1→C3	Example effect	Additional effect of SQL-to-text examples
C1→C4	Integrated metadata effect	Effect of integrated metadata
C2/C3→C4	Integration vs. component	Comparison between individual and integrated metadata

4. Evaluation Metrics

본 연구는 생성 SQL을 PostGIS에서 실행한 결과를 기준으로 NL2GeoSQL 성능을 평가하였다. 평가지표는 SQL 실행 가능성(SQL validity), 실행정확도(execution

accuracy), 결과 일치도(Result F1), 공간함수 사용 여부(Spatial function presence accuracy), 스키마 연결 정확도(Schema linking accuracy), 오류유형(Error type)으로 Table 6과 같이 구성하였다.

스키마 연결 정확도는 질의별로 요구되는 테이블과 컬럼의 사용 여부를 기준으로 산정하였다. 공간함수 출현 정확도는 질의별로 요구되는 'ST_*' 함수의 포함 여부를 기준으로 산정하였다. 이러한 지표는 GeoSQL-Eval에서 제시한 테이블, 필드, 함수 단위 평가 관점을 참고하였다[3]. 공간함수의 인자 구성과 공간조인 구조의 적절성은 독립 지표로 분리하지 않고 실행정확도와 오류유형 분석에서 함께 검토하였다.

V. Results and Discussion

1. Performance by Metadata Condition

C0~C4 조건별 주요 성능 요약은 Table 7과 같다.

먼저, 주 모델인 Qwen2.5-Coder에서 C1과 C2는 C0보다 SQL 실행 가능성은 높았지만 실행정확도는 낮게 나타났다. 이는 C1과 C2가 실행 가능한 GeoSQL 생성에는 도움을 주었으나, 올바른 공간조인 구조와 결과 단위까지 보장하지는 못했음을 의미한다.

가장 두드러진 성능 향상은 SQL-to-text 예시가 포함된 C3에서 나타났다. C3의 실행정확도는 43.3%로 C0~C2보다 높았고, 스키마 연결 정확도도 70.0%로 가장 높았다. 이는 SQL-to-text 예시가 단순한 추가 설명보다 자연어 질의가 어떤 테이블, 컬럼, 공간함수, 집계 단위로 변환되는지를 직접적으로 보여주는 역할을 했기 때문으로 해석된다. 즉, LLM은 값 분포나 공간 기본 메타데이터만 제공받는 경우보다, 실제 질의 작성 패턴을 포함한 예시를 제공받을 때 테이블 선택과 SQL 구조 형성에서 더 안정적인

Table 6. Evaluation Metrics

Metric	Description	Evaluation focus
SQL validity	Whether the generated SQL is executable in PostGIS	Syntax, table, column, function errors
Execution accuracy	Whether the generated result exactly matches the reference result	Final result correctness
Result F1	Overlap between standardized generated and reference results	Partial result agreement
Spatial function presence accuracy	Whether required ST_* functions are included	Spatial operation use
Schema linking accuracy	Whether required tables and columns are used	Table and column linking
Error type	Main cause of failure	Table/column, result-unit, execution, spatial-function, extra-column, and other errors

결과를 보였다.

통합 메타데이터 조건인 C4의 실행정확도는 36.7%로 C0~C2보다 높았으나 C3보다는 낮았으며, 스키마 연결 정확도도 C3보다 낮았다. 이는 값 프로파일과 SQL-to-text 예시를 함께 제공하더라도 항상 성능이 추가로 향상되는 것은 아니며, 메타데이터의 정보량 증가가 SQL 구조 형성에 일관되게 긍정적으로 작용하지 않을 수 있음을 보여준다. 다만 본 연구는 조건별 프롬프트 길이와 정보량을 통제된 실험이 아니므로, C3와 C4의 차이를 특정 요인의 효과로 단정하기는 어렵다.

결과 일치도는 C0에서 90.0%로 가장 높았으나, 실행정확도는 C3와 C4에서 가장 높았다. 이는 일부 생성 SQL이 기존 결과와 부분적으로 유사한 결과를 반환하더라도, 반환 컬럼, 결과 단위, 정렬 기준, 공간조건이 완전히 일치하지 않으면 실행 정확도에서는 오답으로 판정되기 때문이다. 공공행정 의사결정지원 환경에서는 결과 일부의 유사성보다 정확한 집계 단위와 조건 해석이 중요하므로, 결과 일치도와 실행정확도를 종합적으로 확인할 필요가 있음을 확인할 수 있다.

분석결과가 단일 LLM 모델에 한정되는지를 확인하기 위해 Devstral-Small-2-24B-Instruct-2512 모델을 추가 적용한 결과, 실행정확도는 C3(40.0%)가 C4(36.7%)보다 높게 나타나 Qwen2.5-Coder 모델의 결과와 유사한 경향을 보였다. 스키마 연결 정확도 역시 두 모델 모두 C3와 C4에서 가장 높은 수준을 보였다. 따라서 SQL-to-text 예시가 포함된 조건의 상대적 유용성이 실행정확도와 스키마 연결 측면에서 대체로 일관되게 나타났다.

Table 7. Performance by Metadata Condition

Model	Cond.	SQL Valid.	Exec. Acc.	Result F1	Spatial Func. Acc.	Schema Link. Acc.
Qwen	C0	76.7	20.0	90.0	21.4	36.7
	C1	96.7	13.3	86.0	28.6	33.3
	C2	100.0	13.3	87.9	28.6	33.3
	C3	100.0	43.3	86.9	50.0	70.0
	C4	96.7	36.7	86.4	57.1	63.3
Devs-tral	C0	86.7	6.7	82.6	28.6	33.3
	C1	73.3	13.3	92.1	42.9	36.7
	C2	93.3	16.7	83.4	21.4	36.7
	C3	96.7	40.0	85.5	42.9	66.7
	C4	83.3	36.7	82.9	42.9	66.7

Note. Values are percentages.

2. Error Type Analysis

Table 8은 Qwen2.5-Coder 모델을 기준으로 생성 결과의 주요 오류유형을 요약한 것이다. 오류는 주로 테이블 선택 오류(table error), 결과 단위 오류(result unit

error), 컬럼 선택 오류(column error)에 집중되었다.

원천 공간테이블과 행정경계 테이블을 직접 결합해야 하는 질의에서는 파생 분석테이블을 사용하거나 필요한 원천 테이블을 누락하는 사례가 나타났다. 또한 생성 SQL이 유사한 결과값을 반환하더라도, 격자, 자치구, 행정동, 시설물 등 질의에서 요구한 반환 단위를 맞추지 못해 실행 정확도 평가에서 오답으로 판정된 사례도 확인되었다.

컬럼 선택 오류는 모델이 유사한 의미의 컬럼을 혼동하거나 기존 SQL에 필요한 컬럼을 누락한 경우이다. 특히 방법시설 수, 방법취약등급, 행정구역명처럼 질의에 자주 등장하는 컬럼도 테이블 맥락에 따라 다르게 사용되어야 하므로, 단순한 컬럼명 제공만으로는 충분하지 않았다. 이러한 결과는 NL2GeoSQL 성능 개선을 위해 테이블 선택, 컬럼 의미, 공간연산 대상, 반환 단위를 함께 안내하는 메타데이터 구성이 필요함을 보여준다.

Table 8. Major Error Types

Error type	Count	Rate (%)
Table error	35	23.3
Result unit error	33	22.0
Column error	27	18.0
Execution error	6	4.0
Spatial operator error	5	3.3
Extra columns only	3	2.0
Others	3	2.0

Note. Rates are calculated based only on the 150 Qwen2.5-Coder generated SQL outputs. Others include syntax_error, numeric_error, and ranking_order_error.

VI. Conclusion

본 연구는 서울시 범죄예방 의사결정지원 데이터를 활용하여 메타데이터 구성 조건에 따른 NL2GeoSQL 성능 차이를 비교하였다. 분석 결과, SQL-to-text 예시 조건에서 실행정확도가 가장 높게 나타났으며, 통합 메타데이터 조건도 비공간 기준조건, 공간 기본 메타데이터 조건, 값 분포 프로파일 조건보다 높은 수준을 보였다. 또한 SQL-to-text 예시 조건은 스키마 연결 측면에서 우수하였고, 공간함수 사용에서도 비교적 높은 성능을 보였다. 이러한 경향은 추가 검증에 사용한 Devstral 모델에서도 유사하게 확인되었다. 이는 NL2GeoSQL 성능 향상을 위해 메타데이터의 양을 단순히 확대하는 방식보다, 자연어 질의가 SQL 구조로 변환되는 과정을 명확히 제시하는 방식이 중요할 수 있음을 보여준다.

오류유형 분석에서는 주로 테이블·컬럼 선택과 결과 단

위 구성에서 생성 실패가 발생하였다. 이 결과는 공공행정 공간질의에서 단순한 공간함수 목록이나 geometry 정보만으로는 충분하지 않으며, 분석 단위, 공간조인 경로, 집계 기준, 결과 반환 형식을 포함한 메타데이터 설계가 필요함을 시사한다. 또한 행정 의사결정지원 환경에서는 부분적인 결과 일치만으로 실무 활용성을 판단하기 어려우므로, 결과 일치도와 실행정확도를 함께 고려해야 한다.

본 연구의 의의는 실제 도시 행정 공간데이터와 업무 기반 자연어 질의를 대상으로, PostGIS 실행 결과를 기준으로 메타데이터 조건별 NL2GeoSQL 성능을 비교했다는 점에 있다. 또한 메타데이터 구성 방식이 SQL 실행 가능률, 실행정확도, 스키마 연결, 공간함수 사용에 서로 다른 영향을 미칠 수 있음을 실증하였다.

본 연구는 몇 가지 한계를 갖는다. 첫째, 실험은 서울시 범죄예방 의사결정지원 데이터를 대상으로 수행되어 다른 도시나 정책 분야의 공간데이터베이스에 대한 추가 검증이 요구된다. 둘째, 평가질의가 30개의 한국어 공간질의로 구성되어 모든 NL2GeoSQL 질의 유형을 포괄하기에는 범위가 제한된다. 셋째, 메타데이터 조건별 프롬프트 길이와 정보량이 동일하게 통제되지 않아 메타데이터 구성에 따른 효과와 프롬프트 정보량의 영향을 명확하게 구분하는데 한계가 있다. 넷째, 서로 다른 두 코드 특화 LLM을 적용하였으나, 모델의 수와 유형이 제한적이므로 다양한 LLM 환경에서 메타데이터 효과의 차이를 추가로 검증할 필요가 있다. 다섯째, 각 조건-질의 조합에서 1회 생성된 SQL을 평가하였으므로 동일 질의에 대한 반복 생성의 변동성과 안정성까지 확인하지는 못하였다.

향후 연구에서는 평가질의의 규모와 유형을 확대하고, 다양한 공간데이터베이스와 LLM 환경에서 추가 검증을 수행할 필요가 있다. 또한 조건별 토큰 길이와 반복 생성 조건을 통제하여 메타데이터 구성 효과, 프롬프트 정보량의 영향, 생성 SQL의 안정성을 함께 검증해야 한다. 나아가 테이블-컬럼 선택과 결과 단위 오류를 줄일 수 있는 공간질의 특화 메타데이터 설계를 구체화할 필요가 있다.

REFERENCES

- [1] M. Pourreza, and D. Rafiei, "DIN-SQL: Decomposed In-Context Learning of Text-to-SQL with Self-Correction," *Advances in Neural Information Processing Systems*, Vol. 36, pp. 36339-36348, New Orleans, LA, USA, Dec. 2023. DOI: 10.52202/075280-1577
- [2] Y. Zhang, C. Wei, Z. He, and W. Yu, "GeoGPT: An Assistant for Understanding and Processing Geospatial Tasks," *International Journal of Applied Earth Observation and Geoinformation*, Vol. 131, Art. no. 103976, July 2024. DOI: 10.1016/j.jag.2024.103976
- [3] S. Hou, H. Jiao, Z. Liu, L. Xie, G. Chen, S. Wu, X. Guan, and H. Wu, "GeoSQL-Eval: First Evaluation of LLMs on PostGIS-Based NL2GeoSQL Queries," *Expert Systems with Applications*, Vol. 320, Art. no. 132122, July 2026. DOI: 10.1016/j.eswa.2026.132122
- [4] V. Shkapenyuk, D. Srivastava, T. Johnson, and P. Ghane, "Automatic Metadata Extraction for Text-to-SQL," *arXiv preprint arXiv:2505.19988*, May 2025. DOI: 10.48550/arXiv.2505.19988
- [5] Seoul Metropolitan Government, "AI and 3D Spatial Information Combined 'S-Map Model' to Eliminate Blind Spots in Crime Prevention," *Seoul Metropolitan Government Press Release*, Oct. 30, 2025. <https://news.seoul.go.kr/gov/archives/572757>
- [6] T. Yu, R. Zhang, K. Yang, M. Yasunaga, D. Wang, Z. Li, J. Ma, I. Li, Q. Yao, S. Roman, Z. Zhang, and D. Radev, "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*, pp. 3911-3921, Brussels, Belgium, Oct.-Nov. 2018. DOI: 10.18653/v1/D18-1425
- [7] J. Li, B. Hui, G. Qu, J. Yang, B. Li, B. Li, B. Wang, B. Qin, R. Geng, N. Huo, X. Zhou, C. Ma, G. Li, K. C. C. Chang, F. Huang, R. Cheng, and Y. Li, "Can LLM Already Serve as A Database Interface? A Blg Bench for Large-Scale Database Grounded Text-to-SQLs," *Advances in Neural Information Processing Systems*, Vol. 36, pp. 42330-42357, New Orleans, LA, USA, Dec. 2023. DOI: 10.52202/075280-1835
- [8] H. Li, J. Zhang, C. Li, and H. Chen, "RESDSL: Decoupling Schema Linking and Skeleton Parsing for Text-to-SQL," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, No. 11, pp. 13067-13075, Washington, DC, USA, Feb. 2023. DOI: 10.1609/aaai.v37i11.26535
- [9] S. Taleai, M. Pourreza, Y.-C. Chang, A. Mirhoseini, and A. Saberi, "CHESS: Contextual Harnessing for Efficient SQL Synthesis," *arXiv preprint arXiv:2405.16755*, May 2024. DOI: 10.48550/arXiv.2405.16755
- [10] W.-B. Kim, and N. Moon, "NL2MCP: A New Framework for Natural Language Query Execution," *Journal of The Korea Society of Computer and Information*, Vol. 30, No. 10, pp. 43-52, Oct. 2025. DOI: 10.9708/jksci.2025.30.10.043
- [11] Y. Gao, and Z. Luo, "Automatic Database Description Generation for Text-to-SQL," *arXiv preprint arXiv:2502.20657*, Feb. 2025. DOI: 10.48550/arXiv.2502.20657
- [12] N. Wretblad, O. Holmström, E. Larsson, A. Wiksäter, O. Söderlund, H. Öhman, T. Pontén, M. Forsberg, M. Sörme, and F. Heintz, "Synthetic SQL Column Descriptions and Their Impact on Text-to-SQL Performance," *arXiv preprint*

- arXiv:2408.04691, Aug. 2024. DOI: 10.48550/arXiv.2408.04691
- [13] M. Saeedan, M. S. Rashid, A. Eldawy, and V. Hristidis, "GS-QA: A Benchmark for Geospatial Question Answering," arXiv preprint arXiv:2605.22811, May 2026. DOI: 10.48550/arXiv.2605.22811
- [14] Z. Li, M. Grossman, E. Qasemi, M. Kulkarni, M. Chen, and Y.-Y. Chiang, "Benchmarking Geospatial Question Answering with MapQA," Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems, pp. 1042-1045, Minneapolis, MN, USA, Nov. 2025. DOI: 10.1145/3748636.3764174
- [15] Seoul Metropolitan Government, "Crime Prevention Work Support Service Database," Internal research data, Apr. 2026.
- [16] Statistical Geographic Information Service (SGIS), "Census Administrative Boundary, 100-m Grid Boundary," 2024. Available: <https://link.mods.go.kr> (accessed Oct. 1, 2025).
- [17] Seoul Metropolitan Government, "Seoul Safe CCTV Linkage Status," Seoul Open Data Plaza, 2025. Available: <https://data.seoul.go.kr> (accessed Oct. 1, 2025).
- [18] Ministry of the Interior and Safety, "Emergency Call Box Location Information," LOCALDATA, 2025. Available: <https://www.localdata.go.kr> (accessed Oct. 1, 2025).
- [19] Seoul Metropolitan Government, "Lighting: Security Lights," S-Map OpenLab, 2025. Available: <https://openlab.eseoul.go.kr> (accessed Oct. 1, 2025).
- [20] Seoul Metropolitan Government, "Mobility Vulnerable Protection Zones," Smart Seoul Map, 2025. Available: <https://map.seoul.go.kr> (accessed Oct. 1, 2025).
- [21] B. Hui, J. Yang, Z. Cui, J. Yang, D. Liu, L. Zhang, T. Liu, J. Zhang, B. Yu, K. Lu, K. Dang, Y. Fan, Y. Zhang, A. Yang, R. Men, F. Huang, B. Zheng, Y. Miao, S. Quan, Y. Feng, X. Ren, X. Ren, J. Zhou, and J. Lin, "Qwen2.5-Coder Technical Report," arXiv preprint arXiv:2409.12186, Sept. 2024. DOI: 10.48550/arXiv.2409.12186
- [22] A. Rastogi et al., "Devstral: Fine-tuning Language Models for Coding Agent Applications," arXiv preprint arXiv:2509.25193, Aug. 2025. DOI: 10.48550/arXiv.2509.25193.

Authors



Ki-Sung Song received the B.S. and M.S. degrees from the Department of Geoinformatic Engineering, Inha University, Korea, in 2007 and 2012, respectively. He received the Ph.D. degree from the

Department of Urban Planning, Inha University, Korea, in 2025. Dr. Song is currently an Officer in the Spatial Information Division, Seoul Metropolitan Government, Seoul, Korea. His research interests include GeoAI, spatial databases, natural language spatial query processing, urban spatial analysis, and public-sector spatial decision-support systems.