SQLStructEval: Structural Evaluation of LLM Text-to-SQL Generation

Anonymous ACL submission

Abstract

Despite strong performance on Text-to-SQL benchmarks, it remains unclear whether LLM-generated SQL programs are structurally reliable. In this work, we investigate the structural behavior of LLM-generated SQL queries and introduce SQLStructEval, a framework for analyzing program structures through canonical abstract syntax tree (AST) representations. Our experiments across multiple Text-to-SQL benchmarks show that the evaluated LLMs often produce structurally diverse queries for the same input, even when execution results are correct, and that such variance is frequently triggered by surface-level input changes such as paraphrases or schema presentation. As a case study, we further show that a compile-style pipeline that generates structured intermediate representations before deterministic compilation improves both execution accuracy and structural consistency in this setting. These findings suggest that structural reliability provides a useful complementary dimension for evaluating LLM-based program generation systems. Our code is available at https://anonymous.4open.science/r/StructEval-2435.

1 Introduction

LLMs have achieved strong performance in program generation tasks such as code synthesis and Text-to-SQL translation (Chen et al., 2021; Austin et al., 2021). These systems are commonly evaluated by functional correctness: unit tests for code generation and execution accuracy for Text-to-SQL (Chen et al., 2021; Yu et al., 2018). Such metrics are essential, but they treat all execution-correct programs as equivalent.

Programs with identical outputs can differ substantially in structure, for example through alternative joins, nested queries, or redundant operators (Yu et al., 2018; Zhong et al., 2020). These differences are not always errors, but they matter when generated SQL is inspected, audited, cached, optimized, or reused after database changes. They also remain invisible to execution-based evaluation, even though LLM generations can vary structurally under repeated sampling (Wang et al., 2023; Rajkumar et al., 2022).

We study this gap through SQLStructEval, a structure-aware framework that canonicalizes SQL queries into AST-based representations and measures structural consistency, diversity, reference alignment, and robustness under semantically invariant perturbations. We find that the evaluated LLMs often generate structurally diverse SQL for the same input, including among execution-correct outputs, and that paraphrases or schema presentation changes can shift the dominant query structure. As a case study, we further evaluate a compile-style pipeline that generates structured intermediate representations before deterministic compilation; the full pipeline improves execution accuracy and structural consistency in this setting.

Our contributions are threefold: 1) we identify structural instability as an underexplored reliability issue in Text-to-SQL, including among execution-correct generations; 2) we propose SQLStructEval, a framework for measuring structural consistency, diversity, reference alignment, and perturbation robustness via canonicalized AST representations; and 3) we show that execution accuracy can hide structural instability and provide a full-pipeline case study suggesting that explicit structural generation is a promising mitigation direction, while leaving component-level attribution to future work.

2 Related Work

LLM Generation Instability The instability of LLM outputs has received increasing attention. Prior work shows that LLMs are sensitive to stochastic sampling and may produce inconsis-

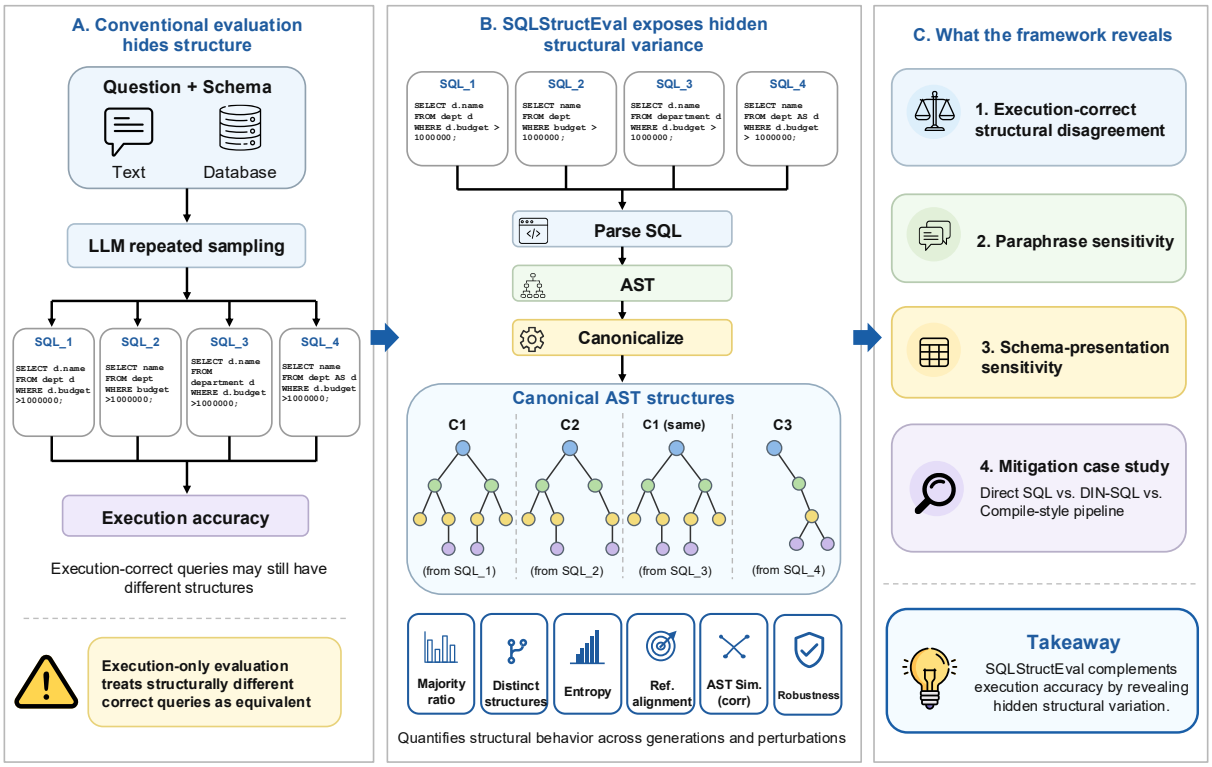


Figure 1: Overview of direct SQL generation and SQLSTRUCTEVAL. SQLSTRUCTEVAL canonicalizes generated SQL into AST-based structures, enabling explicit analysis of structural consistency, diversity, and robustness.

tent outputs under identical prompts (Wang et al., 2023), with early reasoning errors propagating to later steps (Zhang et al., 2024). Models may also over-deliberate during decoding, favoring incorrect solutions despite initially correct reasoning paths (Wang et al., 2025).

To improve robustness in program generation, prior approaches have explored schema linking, self-correction, and execution-based verification (Pourreza and Rafiei, 2023; Ni et al., 2023). However, self-correction remains unreliable without external feedback (Huang et al., 2024), and self-consistency methods require multiple sampling runs with high computational cost (Zhu et al., 2024). Recent work also highlights structural or algorithmic variance across generations (Song et al., 2025; Rajput et al., 2025), motivating stability-oriented approaches such as instance-level randomization (Li et al., 2025).

However, existing studies primarily focus on improving prediction accuracy or reducing output variance, without explicitly analyzing the structural properties of generated programs. In contrast, our work evaluates the structural reliability of LLM-generated SQL queries through a structure-aware framework based on canonicalized AST representations, and further uses structured intermediate representations as a case study for improving structural stability and generation accuracy.

Evaluation for Program Generation Evaluating program generation remains challenging. Early approaches relied on surface-level metrics such as exact match or BLEU (Rajpurkar et al., 2016; Papineni et al., 2002), which are inadequate for program synthesis since semantically equivalent programs may differ in form (Kulal et al., 2019).

Modern benchmarks instead adopt execution-based evaluation. In Text-to-SQL, execution accuracy measures whether predicted queries return the same result as the reference (Zhong et al., 2020), while code generation benchmarks use unit tests (Chen et al., 2021; Austin et al., 2021). Although these metrics better capture functional correctness (Roziere et al., 2020), recent work shows that structurally different programs can produce identical outputs (Liu et al., 2023; Kim et al., 2025). This limitation motivates the need for structure-aware evaluation.

Program Equivalence and SQL Canonicalization A closely related line of work studies program equivalence and canonicalization, aiming to

determine whether two programs are semantically identical despite syntactic differences. In the context of Text-to-SQL, prior work has explored query normalization, equivalence checking, and semantic parsing techniques to address spurious mismatches between predicted and reference queries (Zhong et al., 2020; Guo et al., 2019). However, these approaches primarily focus on improving evaluation fairness or matching correctness, rather than analyzing the variability of generated program structures across multiple generations. In contrast, our work leverages canonicalized representations not to collapse equivalent programs into a single label, but to explicitly quantify structural diversity and instability in LLM outputs.

Structured and Constrained Generation Prior work has explored structured generation and constrained decoding to improve program validity. Methods such as Poesia et al. (2022) and Scholak et al. (2021) enforce structural constraints during decoding, while grammar-based approaches ensure adherence to predefined syntax rules (Geng et al., 2023; Raspanti et al., 2025; Zhang et al., 2025).

However, these approaches focus on syntactic correctness and do not address structural variability across multiple valid generations (Raspanti et al., 2025; Albinhassan et al., 2025). More recent work incorporates logical constraints during decoding (Ma and Hu, 2025), yet structural variability remains unresolved. *In contrast, our work evaluates structural reliability post hoc and explores a compile-style generation paradigm that produces structured intermediate representations before compiling them into executable programs.*

3 SQLStructEval

In this section, we present SQLStructEval, a structure-aware framework for analyzing the structural reliability of LLM-generated programs. Given multiple generations for the same input, SQLStructEval represents programs as canonicalized ASTs and quantifies their structural consistency, diversity, and robustness. We also examine a compile-style generation paradigm that leverages structured intermediate representations to improve program reliability.

3.1 Problem Setup

We study the problem of program generation with LLMs, focusing on the Text-to-SQL task as a controlled and well-defined environment. Given a natural language question x and a database schema S, the goal is to generate an SQL query y that correctly answers the question.

Unlike conventional evaluation settings that consider only a single generated program, we analyze the structural behavior of LLMs across multiple generations. For each input question x, the model produces a set of candidate programs

$$Y = \{y_1, y_2, ..., y_k\}, \quad (1)$$

where k denotes the number of sampled outputs. Each query is converted into its corresponding AST representation, which captures the hierarchical structure of the program. These representations enable SQLStructEval to systematically analyze how consistently the model constructs programs across generations.

3.2 Structural Representation

We represent each generated SQL query as an abstract syntax tree (AST).[1] An AST encodes the hierarchical structure of a program, where nodes correspond to SQL operators such as `SELECT`, `WHERE`, `JOIN`, and `GROUP BY`, and edges represent syntactic relationships among these components.

Formally, for each generated query y_i, we obtain an AST representation

$$T_i = \text{AST}(y_i). \quad (2)$$

Compared with raw SQL text, AST representations abstract away superficial differences such as formatting or alias naming and focus on the underlying program structure. As a result, queries that are structurally equivalent but textually different can be mapped to comparable structural representations.

These representations enable SQLStructEval to compare programs generated for the same input and quantify their structural similarity, diversity, and robustness across generations.

3.3 Structural Evaluation Measures

Based on canonical AST representations, SQLStructEval introduces a set of evaluation measures to characterize the structural behavior of LLM-generated programs. These measures are computed over multiple generations for the same input and capture complementary aspects of model behavior, including structural consistency, diversity,

[1] We parse the SQL queries into ASTs using the `sqlglot` library: https://sqlglot.com/.

alignment with reference programs, and robustness under input perturbations.

For an input question x, suppose the model generates N candidate programs $\{p_i\}_{i=1}^{N}$. After SQL parsing and canonicalization, each program is mapped to either a canonical AST structure $a_i \in \mathcal{A}$ or a parsing failure $a_i = \perp$. Let $\mathcal{I} = \{i \mid a_i \neq \perp\}$ and $M = |\mathcal{I}|$ denote the successfully parsed generations. Among $\{a_i : i \in \mathcal{I}\}$, suppose there are K distinct canonical structures $\{s_k\}_{k=1}^{K}$, with counts $n_k = |\{i \in \mathcal{I} : a_i = s_k\}|$ and empirical frequencies $p_k = n_k/M$.

Structural consistency. The majority-structure ratio

$$\mathrm{Cons}(x) = \max_{1 \le k \le K} p_k \tag{3}$$

measures the proportion of generations that share the dominant AST structure; we refer to it as the *majority ratio* in tables. We additionally report AST Sim on selected subsets, such as execution-correct generations, using the majority-structure ratio within that subset. For Experiment 2, we define two diagnostic mismatch indicators: *Exec-Corr Struct-Diff*, which marks inputs with multiple execution-correct generations but low structural agreement, and *High-Acc Low-Struct*, which marks inputs with high sample-level execution accuracy but more than one execution-correct structure. Full threshold definitions and parse-failure conventions are provided in Appendix A.

Structural diversity. We measure diversity by the distinct structure count $\mathrm{Div}(x) = K$ and dispersion via entropy

$$\mathrm{H}(x) = -\sum_{k=1}^{K} p_k \log_2 p_k. \tag{4}$$

Annotated-reference structure alignment. The reference-alignment rate

$$\mathrm{RefAlign}(x) = \frac{1}{M} \sum_{i \in \mathcal{I}} \mathbf{1}[a_i = a^*] \tag{5}$$

is the fraction of successfully parsed generations whose canonical AST matches the annotated reference SQL (a^*). This is a diagnostic measure, not an assumption that the annotated SQL is the only valid structure.

Structural robustness. For an original input $x^{(0)}$ and T perturbed variants $\{x^{(t)}\}_{t=1}^{T}$, let s_t^* be the majority AST structure among generations for $x^{(t)}$. Cross-variant structural agreement is

$$\mathrm{Cons}_{\text{para}}(x) = \frac{2}{(T+1)T} \sum_{0 \le t_1 < t_2 \le T} \mathbf{1}[s_{t_1}^* = s_{t_2}^*], \tag{6}$$

and perturbation sensitivity is $\mathrm{Sens}(x) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}[s_t^* \neq s_0^*]$. We also report the dataset-level *sensitive fraction*, the proportion of inputs whose majority structure changes relative to $x^{(0)}$.

3.4 Compile-Style Generation with Structured Intermediate Representations

Beyond analyzing structural reliability, we include a compile-style generation case study. Instead of directly emitting SQL, the model first generates a structured intermediate representation in the AST space, which is then deterministically compiled into executable SQL. This makes program structure an explicit artifact rather than only an implicit token-level outcome.

In our experiments, we compare this full pipeline with conventional direct generation and evaluate its impact on execution accuracy and structural stability. The experiment is not intended to isolate the separate effects of JSON constraints, deterministic compilation, and model reasoning; rather, it tests whether making program structure explicit can serve as a practical mitigation direction.

4 Experiments and Evaluation

Our experiments investigate the structural reliability of SQL programs generated by LLMs and evaluate SQLStructEval. While most existing Text-to-SQL evaluations focus primarily on functional correctness (e.g., execution accuracy), we study an additional dimension: whether generated programs remain structurally stable across generations and under input perturbations.

To this end, we conduct four experiments on Spider. The first two experiments analyze structural variance and its relationship to execution-based evaluation across multiple LLM families. The third experiment studies a compile-style pipeline as a case study rather than an attribution ablation. The fourth experiment evaluates the robustness of generated program structures under semantically equivalent input perturbations such as paraphrases and schema variations.

4.1 Experimental Setup

Dataset We mainly conduct experiments on the **Spider** benchmark (Yu et al., 2018), a widely

used cross-domain Text-to-SQL dataset containing natural language questions paired with SQL queries over multiple relational databases. We use the Spider development set, which contains 1,034 questions across 138 databases. Each generated query can be executed against the corresponding SQLite database to verify functional correctness. We choose Spider because it provides reference SQL, executable databases, and a stable protocol for comparing execution behavior with structural behavior on the same examples. We discuss scope and possible data exposure risks in the Limitations section.

Models and Generation We evaluate LLMs from multiple families, including GPT-4.1-mini and GPT-5-mini (OpenAI), Claude-4.5-Sonnet and Claude-4.5-Opus (Anthropic), Gemini-3-Pro and Gemini-2.5-Flash (Google), and DeepSeek-V3.1.

For each input question, the model is provided with the database schema (table names, column names, and foreign-key relations) and asked to generate an SQL query in the SQLite dialect. To analyze structural variance, we sample multiple SQL queries for the same input using stochastic decoding.

Structural Analysis We parse generated SQL queries into ASTs using the `sqlglot` parser and canonicalize them to remove superficial differences such as alias names and formatting. These representations allow us to compare program structures independent of surface forms and analyze both functional correctness and structural properties. Evaluation measures are defined in Section 3.3.

4.2 Experiment 1: Structural Variance of Generated SQL

We begin by examining whether LLMs generate structurally consistent SQL queries when producing multiple solutions for the same input. While Text-to-SQL systems are typically evaluated using execution accuracy or exact match (Yu et al., 2018; Zhong et al., 2020; Deng et al., 2022), these metrics verify result correctness, but do not capture the stability of program structures across generations.

To analyze structural variance, we sample ten SQL queries for each Spider development example from each model using stochastic decoding, resulting in over 10,000 queries per model. We parse all queries into ASTs using `sqlglot` and canonicalize the resulting structures to remove superficial differences such as alias naming and formatting.

Model	Distinct↓	Majority↑	Entropy↓	Ref Align↑
GPT-5-mini	1.913	0.650	0.413	0.197
GPT-4.1-mini	1.620	0.793	0.301	0.253
Claude-4.5-Opus	0.779	0.687	0.049	0.391
Claude-4.5-Sonnet	1.241	0.775	0.195	0.356
DeepSeek-V3.1	1.023	0.735	0.136	0.321
Gemini-3-Pro	0.845	0.595	0.133	0.369
Gemini-2.5-Flash	0.623	0.556	0.044	0.332

Table 1: Structural statistics of SQL queries generated by different LLMs on the Spider development set. For each question, we sample 10 SQL queries and parse them into canonical ASTs. **Distinct** denotes the average number of distinct AST structures per question; **Majority** is the average proportion of sampled generations that share the dominant AST structure for each question, **Entropy** measures the average entropy of the structure distribution, and **Ref Align** is the annotated-reference structural alignment, i.e., the fraction of parsed generations whose canonical AST matches that of the annotated reference SQL query. Because count-based distinct statistics include all-parse-failure inputs as 0, lower Distinct values should be interpreted together with the validity and success statistics in Table 2, not as stability evidence by themselves.

Table 1 shows substantial structural variance across all models. For example, the main model (GPT-5-mini) produces on average 1.91 distinct structures per question, while the majority structure appears only 65% of the time. Although stronger models such as Claude-4.5-Opus exhibit more concentrated distributions, multiple structural variants still frequently occur for the same input. In extreme cases, up to 5–10 distinct structures can appear within ten generations.

These results indicate that structural variance is common in our Spider setting. Even with identical inputs and schemas, models often produce alternative query compositions, such as different join orders or nested query formulations, motivating a closer look at execution-correct outputs.

4.3 Experiment 2: Execution Accuracy vs. Structural Reliability

Experiment 1 showed that repeated sampling often produces multiple structural variants for the same input. This raises a key question: *Does execution accuracy adequately reflect the reliability of generated programs?*

Execution accuracy measures whether the generated query yields the same result as the annotated reference query and is the dominant evaluation metric in Text-to-SQL benchmarks. However, it only captures functional correctness and does not reveal whether program structures are consistent. This ex-

periment reuses the SQL generations from Experiment 1. For each example, we execute all generated queries on the corresponding SQLite database and compare the results with the annotated reference query. Queries that return identical results are labeled as *execution-correct*. Structural statistics are then computed over both all generations and the subset of execution-correct queries using the canonical AST representations described in Section 3.

Table 2 shows that structural variance remains substantial even among execution-correct queries. For instance, GPT-5-mini achieves an execution accuracy of 0.74, yet execution-correct queries still contain on average 1.38 distinct AST structures per question and exhibit relatively low structural agreement (AST similarity 0.55). Similar patterns follow across model families. Moreover, a significant fraction of questions fall into the category of *execution-correct but structurally different*, where multiple queries produce the correct answer, but correspond to different program structures. Depending on the model, this occurs in roughly 20%–39% of questions. These findings reveal a mismatch between execution accuracy and structural reliability: execution-based metrics capture result correctness on the evaluated database instance, but not structural consistency.

Scope validation beyond Spider. To test whether this pattern is specific to the Spider development set, we additionally apply the same Experiment 2 analysis to 100-example subsets of BIRD Mini-Dev (Li et al., 2023), Spider-Syn (Gan et al., 2021), and Dr.Spider (Chang et al., 2023), sampling 10 SQL queries per example for GPT-5-mini, Claude-4.5-Sonnet, and DeepSeek-V3.1 using the same execution and canonicalization pipeline.

These subsets are not intended for model ranking; the lower execution accuracies reflect dataset shift, adversarial or synthetic perturbations, and the absence of benchmark-specific prompting or tuning. The relevant question is whether execution-correct generations still disagree structurally. Table 3 shows that the answer is yes across all three subsets. *Exec-Corr Struct-Diff* remains non-zero throughout (e.g., 0.060–0.170 on BIRD Mini-Dev, up to 0.220 on Spider-Syn for DeepSeek-V3.1, and 0.030–0.110 on Dr.Spider). *High-Acc Low-Struct* also appears in several settings, including GPT-5-mini on Spider-Syn (0.150) and BIRD Mini-Dev (0.110), indicating that relatively high sample-level execution success does not always imply a single dominant canonical structure. We therefore do not use these subsets to claim broad benchmark-level generalization; rather, they provide a stress check showing that execution-correct structural disagreement persists beyond the original Spider development set.

4.4 Experiment 3: A Full-Pipeline Mitigation Case Study

Experiments 1 and 2 reveal substantial structural variance in LLM-generated SQL queries, even when execution accuracy is relatively high. This raises the question of whether a generation pipeline that makes structure explicit can improve program reliability.

We compare three generation paradigms on the Spider development set (1,034 questions). In the **Direct SQL** setting, the model directly generates SQL queries from the natural language question and database schema. As an additional baseline, we include **DIN-SQL** (Pourreza and Rafiei, 2023), which uses decomposed in-context learning to guide SQL generation through multi-step prompting. In the **compile-style** setting, the model first generates a structured JSON representation corresponding to a SQL AST, which is then deterministically compiled into executable SQL. For all pipelines, we sample 10 outputs per question using the same model and identical decoding configurations. This experiment evaluates the complete pipeline and does not by itself attribute gains to the JSON schema, the compiler, or improved model reasoning separately.

Table 4 summarizes the results. The compile-style pipeline achieves the highest execution accuracy (0.785) and improves AST similarity among execution-correct queries from 0.552 to 0.632. It also produces more distinct structures across all generations. This is not inconsistent: *Distinct Structures (all)* is computed over all generations and partly reflects invalid or incorrect branches and broader intermediate-representation exploration, whereas *AST Similarity (correct)* is computed only over execution-correct generations. Nearly all outputs are valid JSON, and 95.9% complete the full pipeline. These results support compile-style generation as a useful full-pipeline intervention, but not as an ablation isolating the source of the gains.

Model	Exec Acc↑	Success Rate↑	Distinct (all)↓	Distinct (corr)↓	AST Sim (corr)↑	High-Acc Low-Struct↓	Exec-Corr Struct-Diff↓
GPT-5-mini	0.741	0.741	1.913	1.378	0.552	0.351	0.297
GPT-4.1-mini	0.754	0.757	1.620	1.129	0.656	0.250	0.163
Claude-4.5-Sonnet	0.757	0.804	1.241	0.714	0.624	0.054	0.201
Claude-4.5-Opus	0.816	0.816	0.779	0.650	0.600	0.038	0.224
DeepSeek-V3.1	0.771	0.771	1.023	0.774	0.612	0.101	0.191
Gemini-2.5-Flash	0.638	0.743	0.623	0.509	0.484	0.016	0.191
Gemini-3-Pro	**0.842**	**0.843**	0.843	0.721	0.537	0.139	**0.337**

Table 2: Execution accuracy and structural reliability statistics of LLM-generated SQL on the Spider development set. For each question, 10 queries are sampled and executed. We report execution-based metrics (Exec Acc, Success Rate) and structure-based metrics (Distinct, AST Sim), along with two inconsistency indicators: High-Acc Low-Struct and Exec-Corr Struct-Diff. **AST Sim (corr)** is the majority-structure ratio among execution-correct generations. **Exec-Corr Struct-Diff** is the fraction of questions with at least two execution-correct generations and AST Sim (corr) < 0.7. **High-Acc Low-Struct** is the fraction of questions with sample-level execution accuracy ≥ 0.8 and more than one execution-correct canonical structure.

Subset	Model	Parse Rate	Exec Acc	Distinct (all)	Distinct (corr)	Majority (corr)	Entropy (corr)	AST Sim (corr)	Exec-Corr Struct-Diff	High-Acc Low-Struct
BIRD Mini-Dev	GPT-5-mini	0.657	0.145	3.380	0.560	0.091	0.200	0.091	0.090	0.110
BIRD Mini-Dev	Claude-4.5-Sonnet	0.496	0.214	1.060	0.290	0.210	0.041	0.210	0.060	0.040
BIRD Mini-Dev	DeepSeek-V3.1	0.399	0.246	1.600	0.500	0.212	0.152	0.212	0.170	0.090
Spider-Syn	GPT-5-mini	0.713	0.229	2.630	0.540	0.186	0.177	0.186	0.070	0.150
Spider-Syn	Claude-4.5-Sonnet	0.686	0.332	1.200	0.410	0.410	0.000	0.410	0.010	0.000
Spider-Syn	DeepSeek-V3.1	0.453	0.355	1.260	0.530	0.300	0.126	0.300	0.220	0.100
Dr.Spider	GPT-5-mini	0.668	0.122	2.610	0.300	0.121	0.092	0.121	0.030	0.070
Dr.Spider	Claude-4.5-Sonnet	0.720	0.244	1.460	0.290	0.276	0.010	0.276	0.030	0.010
Dr.Spider	DeepSeek-V3.1	0.470	0.266	1.490	0.420	0.287	0.072	0.287	0.110	0.060

Table 3: Scope validation on additional Text-to-SQL subsets. Each row summarizes 100 examples and 10 generations per example. **AST Sim (corr)** and **Majority (corr)** denote the majority-structure ratio among execution-correct generations. **Exec-Corr Struct-Diff** follows the Experiment 2 implementation: at least two execution-correct generations and AST Sim (corr) < 0.7. **High-Acc Low-Struct** denotes sample-level execution accuracy ≥ 0.8 with more than one execution-correct canonical structure.

Metric	Direct SQL	DIN-SQL	Compile-style
Execution Accuracy	0.742	0.736	**0.785**
AST Similarity (correct)	0.552	0.579	**0.632**
Distinct Structures (all)	1.908	1.553	2.527
JSON Valid Rate	–	–	0.998
Compilable Rate	–	–	0.971
SQL Parse Rate	–	–	0.902
End-to-End Success	–	–	0.959

Table 4: Performance comparison of three Text-to-SQL generation paradigms on the Spider development set. We report execution accuracy, majority-structure agreement among execution-correct generations (AST Similarity), and diversity across all generated structures (Distinct Structures).

Model	AST Sim (para)↑	Distinct↓	Sensitivity↓	Sensitive Frac↓
GPT-5-mini	0.328	13.220	0.622	0.900
GPT-4.1-mini	0.493	6.640	0.462	0.770
DeepSeek-V3.1	0.655	3.330	0.280	0.515
Claude-4.5-Sonnet	0.755	2.190	0.179	0.365
Claude-4.5-Opus	0.755	2.190	0.179	0.365
Gemini-2.5-Flash	0.787	2.870	0.158	0.395
Gemini-3-Pro	**0.894**	**1.530**	**0.079**	**0.195**

Table 5: Structural robustness of SQL generation under paraphrase perturbations on the Spider development set. We measure cross-paraphrase structural consistency (AST Sim), diversity (Distinct), and sensitivity to input variations (Sensitivity, Sensitive Frac).

4.5 Experiment 4: Structural Robustness under Input Perturbations

We further ask whether the structural variance observed in the previous experiments is driven by genuine semantic ambiguity or by sensitivity to superficial input changes. To study this, we evaluate structural robustness under two types of semantically equivalent perturbations: natural language paraphrases and schema presentation changes.

We randomly select 200 Spider development questions and construct semantically equivalent variants. Paraphrases are generated with an LLM prompt and manually checked to preserve attributes, filters, aggregation, comparisons, ordering, limits, and schema semantics; schema variants reorder tables or columns while preserving names and foreign keys. We then apply the same repeated-sampling protocol and report cross-variant AST similarity, distinct structures, and perturbation sensitivity.

Paraphrase Perturbations For each selected question, we construct several paraphrases that preserve the original database semantics while varying wording and syntax. Table 5 reports the results.

Paraphrase perturbations frequently change the generated SQL structure. For the main model, cross-paraphrase AST similarity is only 0.33, the average number of distinct structures exceeds 13,

Model	AST Sim (schema)↑	Distinct↓	Sensitivity↓	Sensitive Frac↓
GPT-5-mini	0.490	7.330	0.500	0.640
GPT-4.1-mini	0.727	3.600	0.263	0.360
DeepSeek-V3.1	0.808	1.870	0.170	0.260
Gemini-2.5-Flash	0.895	1.790	0.098	0.145
Claude-4.5-Sonnet	0.955	1.130	0.043	0.065
Claude-4.5-Opus	0.955	1.130	0.043	0.065
Gemini-3-Pro	**0.957**	**1.060**	**0.043**	**0.065**

Table 6: Structural robustness of SQL generation under schema presentation perturbations on the Spider development set. We evaluate cross-variant structural consistency (AST Sim), diversity (Distinct), and sensitivity to schema changes (Sensitivity, Sensitive Frac).

and nearly 90% of questions are structurally sensitive to paraphrasing. Stronger models are more robust, but even the best model (Gemini-3-Pro) still changes structure on about 20% of questions.

This suggests that the model's program construction is not invariant to semantically equivalent inputs, and may depend on superficial linguistic cues. In particular, different phrasings of the same question may trigger different decomposition strategies or join patterns, indicating instability in the underlying reasoning process.

Schema Presentation Perturbations We next keep the natural language question fixed and perturb only the schema presentation, for example by reordering tables or columns while preserving the same database semantics. Results are shown in Table 6.

Compared to paraphrase perturbations, schema perturbations lead to higher structural stability, suggesting that models are more sensitive to linguistic variation than to schema ordering. Models are generally more robust to schema perturbations than to paraphrases, but the effect is still substantial. For GPT-5-mini, cross-schema AST similarity is 0.49 and about 64% of questions remain structurally sensitive. Even strong models such as Claude and Gemini still exhibit non-negligible sensitivity despite much higher robustness.

Overall, Experiment 4 shows that structural instability is strongly influenced by surface-form sensitivity. Even when semantics remain unchanged, perturbing either the question wording or the schema presentation can trigger different SQL structures. These results suggest that structural instability is not merely a consequence of stochastic decoding, but is systematically influenced by surface-form variations. This reinforces the motivation for structural evaluation and for generation strategies that operate more explicitly in the structural space.

5 Conclusion

We introduced SQLSTRUCTEVAL, a structure-aware evaluation framework for analyzing the reliability of LLM-generated SQL programs beyond execution accuracy. SQLSTRUCTEVAL canonicalizes generated SQL queries into AST-based representations and measures complementary aspects of structural behavior, including consistency across repeated generations, diversity of canonical structures, reference-structure alignment, execution-correct structural mismatch, and robustness under semantically invariant perturbations.

Our experiments show that execution accuracy can hide substantial structural instability. Across multiple LLM families on Spider, models often produce different canonical query structures for the same question, including cases where multiple generated queries are execution-correct. Robustness experiments further show that paraphrases and schema-presentation changes can shift the dominant generated structure, indicating that structural behavior is sensitive to surface-form variation rather than only to semantic intent. Additional scope validation on BIRD Mini-Dev, Spider-Syn, and Dr.Spider further shows that execution-correct structural disagreement is not confined to the original Spider development set.

These results clarify the role of structural reliability in Text-to-SQL evaluation. Structural variance is not always an error, since different SQL programs may be execution-equivalent on a given database instance. However, such variance matters when generated SQL is inspected, audited, cached, optimized, reused, or evaluated under database and schema changes. Thus, SQLSTRUCTEVAL is intended to expose hidden structural variation that execution-based metrics cannot observe, rather than to mark every structural difference as incorrect.

Finally, our compile-style case study suggests that making program structure explicit through structured intermediate representations and deterministic compilation can improve both execution accuracy and structural consistency in this setting. Since this experiment evaluates a complete pipeline, future work should isolate the effects of JSON constraints, deterministic compilation, and model planning, and extend structural evaluation to broader Text-to-SQL and program generation benchmarks.

Limitations

The full-scale experiments in this paper are conducted on the Spider benchmark (Yu et al., 2018), which provides reference SQL queries, executable databases, and a stable protocol for jointly analyzing execution correctness and structural behavior. To reduce the risk that our observations are specific to the original Spider development set, we additionally conduct scope validation on BIRD Mini-Dev (Li et al., 2023), Spider-Syn (Gan et al., 2021), and Dr.Spider (Chang et al., 2023). These additional experiments show that execution-correct structural disagreement also appears under dataset shift and adversarial or synthetic variants. However, they are based on 100-example subsets and are intended as scope validation rather than full benchmark-level comparisons. Broader evaluation on larger subsets and additional real-world Text-to-SQL benchmarks remains important future work.

Since Spider has been publicly available for several years, some evaluated models may have been exposed to Spider-style data during pretraining. Our analysis therefore does not use Spider results as leaderboard evidence of absolute model capability; it focuses on relative structural behavior under repeated sampling, execution-correct subsets, and controlled perturbations.

In addition, our analysis relies on canonical AST representations, which capture structural differences between SQL queries but do not fully account for semantic equivalence between different query formulations. The annotated reference SQL is used only as a diagnostic reference structure, not as the unique correct structure. Finally, the compile-style experiment evaluates a complete pipeline and does not isolate the contributions of the JSON schema, deterministic compiler, and model reasoning. Exploring richer semantic metrics, broader benchmark coverage, and targeted ablations remains important future work.

Ethical Considerations

This work studies the structural reliability of LLM-generated SQL queries using publicly available benchmarks. We conduct all experiments on the Spider dataset, which contains no personally identifiable or sensitive user information. Therefore, this work does not involve human subjects or private data.

A potential risk of LLM-based program generation is the production of incorrect or misleading queries, which could lead to erroneous downstream decisions if deployed in real-world systems. Our work aims to mitigate such risks by proposing structure-aware evaluation measures that better reveal instability and hidden failure modes beyond execution accuracy.

All experiments are conducted in an offline evaluation setting, and we do not deploy generated queries to real-world databases. We also acknowledge that different LLM providers may exhibit variability in outputs due to model updates or API changes, which may affect reproducibility. We encourage future work to consider standardized evaluation protocols and open-source implementations to improve transparency and reproducibility.

Data License Our experiments use publicly available Text-to-SQL benchmarks, including Spider, BIRD Mini-Dev, Spider-Syn, and Dr.Spider. We use these datasets in accordance with their original licensing terms and only for offline research evaluation.

The paraphrased inputs used in our robustness experiments are automatically generated transformations of the original benchmark questions and do not introduce new proprietary content. Schema-presentation variants are constructed by reordering schema elements while preserving the original database semantics. All derived data, including canonicalized AST representations, generated SQL outputs, and perturbation variants, are produced or processed during evaluation.

We will release our code and evaluation scripts to facilitate reproducibility. Any released artifacts will comply with the licensing terms of the original datasets and the usage policies of the evaluated model providers.

References

Mohammad Albinhassan, Pranava Madhyastha, Mark Law, and Alessandra Russo. 2025. Learning and enforcing context-sensitive control for LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 834–842, Vienna, Austria. Association for Computational Linguistics.

Jacob Austin, Augustus Odena, Maxwell I. Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie J. Cai, Michael Terry, Quoc V. Le, and Charles Sutton. 2021. Program synthesis with large language models. *CoRR*, abs/2108.07732.

Shuaichen Chang, Jun Wang, Mingwen Dong, Lin Pan, Henghui Zhu, Alexander Hanbo Li, Wuwei

Lan, Sheng Zhang, Jiarong Jiang, Joseph Lilien, Steve Ash, William Yang Wang, Zhiguo Wang, Vittorio Castelli, Patrick Ng, and Bing Xiang. 2023. Dr.spider: A diagnostic evaluation benchmark towards text-to-SQL robustness. In *The Eleventh International Conference on Learning Representations*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

Naihao Deng, Yulong Chen, and Yue Zhang. 2022. Recent advances in Text-to-SQL: A survey of what we have and what we expect. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2166–2187, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yujian Gan, Xinyun Chen, Qiuping Huang, Matthew Purver, John R. Woodward, Jinxia Xie, and Pengsheng Huang. 2021. Towards robustness of text-to-SQL models against synonym substitution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2505–2515, Online. Association for Computational Linguistics.

Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. 2023. Grammar-constrained decoding for structured NLP tasks without finetuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952, Singapore. Association for Computational Linguistics.

Jiaqi Guo, Zecheng Zhan, Yan Gao, Yan Xiao, Jian-Guang Lou, Ting Liu, and Dongmei Zhang. 2019. Towards complex text-to-SQL in cross-domain database with intermediate representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4524–4535, Florence, Italy. Association for Computational Linguistics.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large Language Models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Heegyu Kim, Jeon Taeyang, SeungHwan Choi, Seungtaek Choi, and Hyunsouk Cho. 2025. FLEX: Expert-level false-less EXecution metric for Text-to-SQL benchmark. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4448–4475, Albuquerque, New Mexico. Association for Computational Linguistics.

Sumith Kulal, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, and Percy Liang. 2019. SPoC: Search-based pseudocode to code. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11883–11894.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Wang, Bowen Qin, Ruiying Geng, Nan Huo, Xuanhe Zhou, Chenhao Ma, Guoliang Li, Kevin C.C. Chang, Fei Huang, Reynold Cheng, and Yongbin Li. 2023. Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Yiyang Li, Yonghuang Wu, Ying Luo, Liangtai Sun, Zishu Qin, Lin Qiu, Xuezhi Cao, and Xunliang Cai. 2025. Instance-level randomization: Toward more stable LLM evaluations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3411–3425, Suzhou, China. Association for Computational Linguistics.

Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. Is your code generated by ChatGPT really correct? rigorous evaluation of large language models for code generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Franklin Ma and Alan J. Hu. 2025. Logically constrained decoding. In *Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025)*, pages 150–167, Suzhou, China. Association for Computational Linguistics.

Ansong Ni, Srini Iyer, Dragomir Radev, Ves Stoyanov, Wen-tau Yih, Sida I. Wang, and Xi Victoria Lin. 2023. Lever: learning to verify language-to-code generation with execution. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Gabriel Poesia, Alex Polozov, Vu Le, Ashish Tiwari, Gustavo Soares, Christopher Meek, and Sumit Gulwani. 2022. Synchromesh: Reliable code generation from pre-trained language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Mohammadreza Pourreza and Davood Rafiei. 2023. Din-sql: Decomposed in-context learning of Text-to-SQL with self-correction. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Nitarshan Rajkumar, Raymond Li, and Dzmitry Bahdanau. 2022. Evaluating the Text-to-SQL capabilities of large language models. *CoRR*, abs/2204.00498.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Prateek Rajput, Abdoul Aziz Bonkoungou, Yewei Song, Abdoul Kader Kaboré, Iyiola E. Olatunji, Jacques Klein, and Tegawendé F. Bissyandé. 2025. Dynamic stability of LLM-Generated code. *CoRR*, abs/2511.07463.

Federico Raspanti, Tanir Ozcelebi, and Mike Holenderski. 2025. Grammar-constrained decoding makes large language models better logical parsers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 485–499, Vienna, Austria. Association for Computational Linguistics.

Baptiste Roziere, Marie-Anne Lachaux, Lowik Chanussot, and Guillaume Lample. 2020. Unsupervised translation of programming languages. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Torsten Scholak, Nathan Schucher, and Dzmitry Bahdanau. 2021. PICARD: Parsing incrementally for constrained auto-regressive decoding from language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9895–9901, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yewei Song, Tiezhu Sun, Xunzhu Tang, Prateek Kumar Rajput, Tegawendé F. Bissyandé, and Jacques Klein. 2025. Measuring LLM code generation stability via structural entropy. In *2025 40th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 3922–3926.

Kaishen Wang, Hengrui Gu, Meijun Gao, and Kaixiong Zhou. 2025. DAMO: decoding by accumulating activations momentum for mitigating hallucinations in vision-language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, Zilin Zhang, and Dragomir Radev. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and Text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3911–3921, Brussels, Belgium. Association for Computational Linguistics.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024. How language model hallucinations can snowball. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Qinggang Zhang, Hao Chen, Junnan Dong, Shengyuan Chen, Feiran Huang, and Xiao Huang. 2025. Structure-guided large language models for text-to-SQL generation. In *Forty-second International Conference on Machine Learning*.

Ruiqi Zhong, Tao Yu, and Dan Klein. 2020. Semantic evaluation for Text-to-SQL with distilled test suites. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 396–411, Online. Association for Computational Linguistics.

Jiace Zhu, Yingtao Shen, Jie Zhao, and An Zou. 2024. Path-Consistency: Prefix enhancement for efficient inference in LLM. *CoRR*, abs/2409.01281.

A Structural Metric Details

This appendix provides the full definitions deferred from Section 3.3.

Parse-failure conventions. If $M = 0$, we set structure-count measures such as $\mathrm{Div}(x)$ to 0 and exclude the instance from ratio-based measures such as majority ratio, entropy, and AST similarity. Dataset-level averages are computed over all inputs for count measures and over defined inputs for ratio measures. This convention explains why an average distinct-structure count can be below 1 when many generations fail to parse.

Majority structure under perturbations. For each perturbed input $x^{(t)}$, the majority AST structure is

$$s_t^* = \arg\max_{s \in \mathcal{A}} n_{t,s}, \tag{7}$$

where $n_{t,s}$ denotes the count of structure s among generations for input $x^{(t)}$.

Subset-level structural agreement. For selected subsets $B_x \subseteq \mathcal{I}$ with $m = |B_x|$, we report subset-level structural agreement

$$\mathrm{Sim}(B_x) = \max_{s \in \mathcal{A}} \frac{|\{i \in B_x : a_i = s\}|}{m}, \tag{8}$$

for $m \geq 1$; empty subsets are excluded. In Tables 2–4, *AST Sim (corr)* applies this majority-structure ratio to execution-correct generations ($B_x = C_x$). In Tables 5–6, *AST Sim* reports cross-variant structural agreement ($\mathrm{Cons}_{\text{para}}$) computed from the majority structures $\{s_t^*\}$ defined below.

Execution-structure mismatch indicators. For Experiment 2, we report two dataset-level indicators that connect execution correctness with structural behavior. Let C_x be the set of parsed generations for x that are execution-correct. *Exec-Corr Struct-Diff* is the fraction of inputs for which at least two generations are execution-correct and their execution-correct subset has low structural agreement:

$$\mathbf{1}\left[|C_x| \geq 2 \wedge \mathrm{Sim}(C_x) < 0.7\right]. \tag{9}$$

This captures execution-correct outputs that do not concentrate around a dominant structure. *High-Acc Low-Struct* is the fraction of inputs whose sample-level execution accuracy is at least 0.8 and whose execution-correct generations contain more than one distinct canonical structure. These thresholded indicators are used only as diagnostics for Table 2.

Table-column mapping. Table columns map directly to these definitions: *Distinct* to Div, *Majority* to Cons, *Entropy* to H, *Ref Align* to $\mathrm{RefAlign}$, *AST Sim (corr)* to $\mathrm{Sim}(C_x)$, and *Sensitivity/Sensitive Frac* to perturbation sensitivity.

B Additional Error Analysis

To better understand the behavioral differences between traditional direct SQL generation and the compile-style generation paradigm, we conduct an additional qualitative error analysis on representative examples from the Spider development set.

B.1 Error Taxonomy

Based on manual inspection of generated SQL queries, we categorize errors into six major types:

- **A. Schema Linking Errors.** Incorrect or missing references to tables, columns, or foreign-key relationships between tables.
- **B. Logical Form Errors.** Errors in high-level query logic such as aggregation, grouping, nested queries, or set operations.
- **C. Join Path Errors.** Incorrect or incomplete join paths between tables, including missing joins or incorrect join conditions.
- **D. Predicate / Constraint Errors.** Incorrect or missing filtering conditions in `WHERE` or `HAVING` clauses, such as incorrect comparison operators or incorrect attribute references.
- **E. Structural Variance but Executable.** Queries that produce correct execution results but differ substantially in structure from the annotated reference query or other correct solutions.
- **F. Intermediate Representation / Compilation Errors** Errors introduced by the structured intermediate representation or the compilation process in the compile-style generation pipeline.

B.2 Error Distribution

Table 7 shows the distribution of error types observed in baseline and compile-style generation. Each ratio represents the proportion of inspected examples assigned to the corresponding category. We manually inspected 100 Spider development examples for each setting (100 direct-SQL generations and 100 compile-style generations), sampled

Error Type	Baseline	Compile-style	Δ
A. Schema linking	0.20	0.13	-0.07
B. Logical form	0.18	0.15	-0.03
C. Join path	0.14	0.09	-0.05
D. Predicate / constraint	0.16	0.10	-0.06
E. Structural variance (exec OK)	0.28	0.40	+0.12
F. Intermediate representation / compile failure	0.04	0.13	+0.09

Table 7: Error type statistics for baseline and compile-style generation on the Spider development set. Each value represents the fraction of the 100 inspected examples in that setting assigned to the corresponding primary category. **Baseline** refers to direct SQL generation, while **Compile-style** refers to generation using structured intermediate representations followed by deterministic compilation into SQL. Δ denotes the difference between the two settings (compile-style minus baseline). Positive values indicate that the error type becomes more frequent under compile-style generation.

from cases with execution errors or structural disagreement. One author performed the initial annotation and a second author checked the labels. Categories are assigned as mutually exclusive primary error types; when multiple issues appear in one example, we label the earliest error that determines the final failure or structural divergence. Thus, the ratios in each column sum to 1 over the 100 inspected examples for that setting.

As shown in Table 7, compile-style generation reduces several common semantic errors, including schema linking, logical form mistakes, incorrect join paths, and predicate errors. However, two categories increase: structurally diverse but executable queries (Type E) and intermediate representation or compilation-related failures (Type F). This pattern is consistent with the quantitative observations reported in Experiment 3.

B.3 Representative Case Studies

We further illustrate typical failure modes using representative examples from the Spider development set.

Case Group A: Baseline Execution Failures. In many cases, baseline generation produces incorrect SQL queries due to schema linking or join path errors. Compile-style generation often recovers the correct structure by explicitly modeling table relations in the intermediate representation.

For example, for the question:

> "Which model of the car has the minimum horsepower?"

the annotated reference query joins `CAR_NAMES` and `CARS_DATA` to retrieve the corresponding model. Baseline generation frequently produces queries that omit this join and directly access the `CARS_DATA` table, resulting in incorrect results. In contrast, compile-style generation constructs the join structure explicitly in the intermediate representation, producing SQL queries that are structurally equivalent to the annotated reference query.

Case Group B: Execution-Correct but Structurally Divergent Queries. We also observe cases where multiple generated queries produce identical execution results but differ substantially in their syntactic structure. For example, for the question:

> "Show all countries and the number of singers in each country."

baseline generation often produces multiple structurally distinct SQL queries, including variants with redundant subqueries or unnecessary `DISTINCT` operators. Although these queries produce identical results, they correspond to different AST structures and therefore increase structural entropy. Compile-style generation tends to produce a more consistent structural form, typically converging to a simple `GROUP BY` aggregation.

Case Group C: Paraphrase Sensitivity. Semantically equivalent questions can trigger different query structures in baseline generation. For example, the two questions

> "Show all countries and the number of singers in each country"
> "How many singers are from each country?"

share the same semantic intent and annotated reference SQL query. However, baseline generation often produces different structural patterns for the two inputs, including variations involving subqueries or derived tables. Compile-style generation shows improved structural consistency under paraphrasing, although some degree of variance remains.

Case Group D: Compile-style Failure Cases. Despite the overall improvements, compile-style generation introduces new failure modes related to the intermediate representation and compilation process. For instance, when answering the question

> "How many car models are produced in the USA?"

the correct query requires joining MODEL_LIST, CAR_MAKERS, and COUNTRIES. In some compile-style generations, the intermediate representation omits the final join to COUNTRIES and instead applies the filtering condition directly to the CAR_MAKERS table. This leads to execution errors despite the structured generation pipeline.

Overall, these case studies highlight complementary strengths and limitations of the two paradigms. Compile-style generation reduces many semantic reasoning errors but may introduce new failure modes related to the intermediate representation, suggesting opportunities for future improvements in intermediate representation design and compilation strategies.

C Implementation Details

This section describes the implementation details of SQLStructEval, including SQL parsing and canonicalization, canonical AST representation, execution evaluation, and generation settings.

C.1 SQL Parsing and Canonicalization

We parse all generated SQL queries using the `sqlglot` parser, which provides robust SQL parsing and AST construction across multiple SQL dialects. In our experiments we use the SQLite dialect to ensure compatibility with the Spider benchmark databases.

To remove surface-level differences that are irrelevant to program structure, we apply both text-level normalization and AST-level canonicalization.

Text-level normalization. After rendering the parsed AST back into SQL text, we perform lightweight normalization including: (1) removing trailing semicolons and leading/trailing whitespace, (2) replacing line breaks with spaces and collapsing redundant whitespace, (3) removing simple column aliases (e.g., `SELECT col AS x`), and (4) converting the SQL string to lowercase.

AST-level canonicalization. To ensure that structurally equivalent queries share the same representation, we further canonicalize the AST structure:

- **Alias normalization.** Table aliases are renamed according to their first appearance order (e.g., `t1, t2, ...`). All column references are updated accordingly so that structurally identical joins remain equivalent under different alias choices.
- **Logical operator normalization.** Nested AND predicates are flattened into condition lists, sorted by their SQL string representation, and reconstructed into canonical AST nodes. This ensures that logically equivalent expressions such as `A AND B` and `B AND A` share the same canonical form.

These normalization steps are shared across all experiments (Exp1–Exp4) and applied to both baseline and compile-style generated queries.

C.2 Canonical AST Representation

For each SQL query we obtain a canonical representation through the following pipeline:

1. Parse the SQL query into an AST using `sqlglot`;
2. Apply AST-level canonicalization (alias normalization and logical operator normalization);
3. Render the canonical AST back into SQL;
4. Apply text-level normalization.

The resulting canonical SQL string serves as our canonical representation of the AST structure and is used as the basis for all structural measures in our analysis.

For each question with k generated SQL queries, we compute the frequency distribution over canonical structures. Structural statistics are then derived from this distribution:

- **Distinct structures:** the number of unique canonical SQL representations.
- **Majority structure ratio:** the relative frequency of the most common structure among the k generations.
- **Structural entropy:** the Shannon entropy of the structure distribution, measuring structural diversity.

To compare generated structures with the annotated reference query, the reference SQL is canonicalized using the same pipeline, and structural matches are determined by canonical representation equality. This comparison is used as a reference-alignment diagnostic rather than as a claim that the annotated SQL is the only valid structure.

C.3 Execution Environment

Execution-based evaluation follows the standard Spider evaluation protocol. All experiments use the official Spider SQLite database files.

For each generated SQL query, we execute the query against the corresponding SQLite database and obtain the result set. A generated query is considered execution-correct if its execution result exactly matches the result produced by the annotated reference SQL query. Queries that produce execution errors or mismatched results are considered incorrect.

C.4 Generation Settings

For each question we generate multiple candidate SQL queries from each model to analyze structural variation.

Sampling configuration. We sample $k = 10$ SQL queries for every $(question, schema)$ pair. For APIs that cannot return multiple samples in a single request, we perform multiple independent calls and aggregate the generated outputs.

Model names. We use the following models: GPT-5-mini, GPT-4.1-mini, Claude-4.5-Opus, Claude-4.5-Sonnet, Gemini-3-Pro, Gemini-2.5-Flash, and DeepSeek-V3.1. API outputs can change over time because hosted models may be updated by providers; the experiments were run between February and April 2026. We record model identifiers, decoding settings, and generation timestamps in the released metadata.

Decoding parameters. Unless restricted by the model API, we use stochastic decoding with temperature $= 1.0$ and default nucleus sampling parameters. Beam search or greedy decoding is not used, as our goal is to analyze structural variation under realistic sampling conditions.

Compile-style generation. In Experiment 3, the compile-style setting requires models to generate a structured intermediate representation instead of directly producing SQL text. The structured representation is expressed as JSON and contains fields such as `select`, `from`, `joins`, `where`, `group_by`, `having`, `order_by`, and `limit`. A deterministic compiler converts the JSON representation into executable SQL. The compiled SQL is then evaluated using the same canonicalization and execution evaluation pipeline as the baseline generation.

D Prompt and Generation Settings

This section provides the prompt templates used in our experiments.

D.1 Prompt for Direct SQL Generation

In the baseline setting, models are prompted to directly generate a SQL query given a natural language question and the corresponding database schema.

```
You are an expert SQL generator.
Use SQLite dialect.
Only output ONE SQL query, no explanation.

Database ID: <db_id>

Database schema (JSON):
<schema_json>

Question:
<natural_language_question>

SQL:
```

D.2 Prompt for Compile-style JSON Generation

In the compile-style setting, the model is instructed to generate a structured JSON representation of the SQL query rather than directly producing SQL text. The prompt used for this setting is shown below.

```
You are an expert Text-to-SQL system for the
    Spider benchmark.
Your task is to write a structured JSON
    representation of a SQL query
for the given question and database schema.

Requirements:

- Use ONLY tables and columns that exist in the
     provided schema.
- Assume the database uses the SQLite dialect.
- You MUST output a single JSON object, and
    nothing else (no explanations).
- The JSON must describe the logical structure
    of the SQL query with the following fields:
  - type: "query"
  - query: {
    select: [ ... ],
    from: { ... },
    joins: [ ... ],
    where: [ ... ],
    group_by: [ ... ],
    having: [ ... ],
    order_by: [ ... ],
    limit: ...,
    distinct: ...
  }
- Do NOT include any natural language text in
    the JSON.

Database ID: {db_id}

Database schema (JSON):
```

```
{schema_json}

Question:
{question}

Now output ONLY the JSON object for the query
    structure:
```

D.3 Prompt for Paraphrase Generation and Checking

For the paraphrase perturbation experiment, we generate candidate paraphrases with an LLM prompt and then manually check them before use. The prompt asks the model to preserve the database semantics of the original question:

```
Rewrite the following Text-to-SQL question into
     semantically equivalent
paraphrases. Preserve the requested attributes,
     aggregation, filters,
comparison operators, ordering, limits, and all
     schema-specific meanings.
Do not add or remove constraints. Do not make
    the question more ambiguous.

Original question:
<question>

Output <n> paraphrases, one per line.
```

We exclude paraphrases that change selected columns, aggregation targets, predicates, comparison relations, ordering, limits, or the intended schema entities. We also exclude variants that add new conditions, omit necessary constraints, or introduce ambiguity. This filtering is intended to keep the perturbation semantic-preserving while still varying surface wording.

E Additional Structural Examples

To complement the quantitative structural metrics reported in the main experiments, we present several representative examples illustrating common forms of structural variation in LLM-generated SQL queries.

Example 1: Execution-correct but structurally different queries **Question.** Which model of the car has the minimum horsepower? (db_id = car_1)

Reference SQL.

```
SELECT T1.Model
FROM CAR_NAMES AS T1
JOIN CARS_DATA AS T2 ON T1.MakeId = T2.Id
ORDER BY T2.Horsepower ASC
LIMIT 1;
```

Generated SQL (baseline).

```
SELECT DISTINCT cn.Model
FROM cars_data cd
JOIN car_names cn ON cd.Id = cn.MakeId
WHERE cd.Horsepower = (
  SELECT MIN(Horsepower) FROM cars_data
);
```

Although both queries return identical execution results, their canonical AST structures differ substantially, as illustrated in Figure 2. The reference query computes the minimum value using an ORDER BY + LIMIT structure, while the generated query uses an aggregate subquery with MIN(Horsepower) and an equality predicate. This introduces a nested subquery node in the AST, resulting in a significantly different structural representation.

Example 2: Structural sensitivity to paraphrase Consider the following pair of paraphrased questions from the Spider dataset (db_id = concert_singer):

Question A. Show all countries and the number of singers in each country.

Question B. How many singers are from each country?

Both questions correspond to the same annotated reference SQL query:

```
SELECT country, COUNT(*)
FROM singer
GROUP BY country;
```

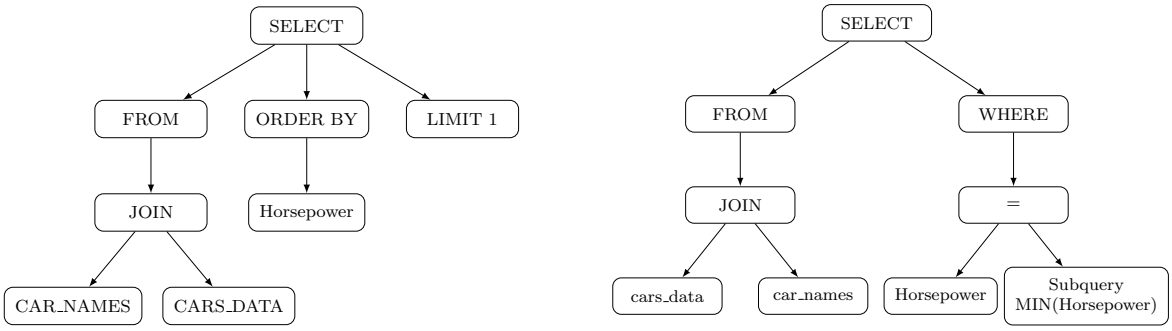
However, the baseline model produces slightly different structural variants:

```
-- Variant A
SELECT Country, COUNT(Singer_ID)
FROM singer
GROUP BY Country;

-- Variant B
SELECT Country, COUNT(*)
FROM singer
GROUP BY Country;
```

Although both queries are execution-equivalent, their AST structures differ because the aggregation argument differs (COUNT(*) vs. COUNT(Singer_ID)). Across multiple generations, additional variations such as derived tables or DISTINCT operators may also appear. This example illustrates how small linguistic changes in the question can lead to different structural query plans.

Example 3: Compile-style correcting structural errors **Question.** What are the names and ids of all countries with at least one car maker? (db_id = car_1)



(a) AST of reference SQL query

(b) AST of generated SQL query

Figure 2: AST structures of two execution-equivalent SQL queries in Example 1 of Appendix E.

Reference SQL.

```
SELECT T1.CountryName, T1.CountryId
FROM COUNTRIES AS T1
JOIN CAR_MAKERS AS T2 ON T1.CountryId = T2.
    Country
GROUP BY T1.CountryId
HAVING COUNT(*) >= 1;
```

Baseline SQL (typical failure).

```
SELECT CountryName, CountryId
FROM COUNTRIES
JOIN CAR_MAKERS
ON COUNTRIES.CountryId = CAR_MAKERS.Country;
```

The baseline model frequently omits the GROUP BY and HAVING clauses, producing incorrect query results.

Compile-style SQL.

```
SELECT COUNTRIES.CountryName, COUNTRIES.
    CountryId
FROM COUNTRIES
JOIN CAR_MAKERS
ON COUNTRIES.CountryId = CAR_MAKERS.Country
GROUP BY COUNTRIES.CountryId
HAVING COUNT(*) >= 1;
```

In the compile-style generation setting, the structured intermediate representation explicitly models aggregation components such as group_by and having. As a result, the compiled SQL consistently contains the correct aggregation structure and closely matches the annotated reference AST.

F Example of the Compile-style Generation Pipeline

To illustrate the compile-style generation process used in Experiment 3, we present a concrete example showing the full pipeline from natural language to structured intermediate representation, internal AST, and final SQL query.

F.1 Input Natural Language Query

Question.

Show the name and capacity of the stadium with the largest average attendance.

Database schema (simplified).

```
{
  "tables": [
    {"table_name": "stadium",
     "columns": ["Stadium_ID", "Name", "
    Location", "Capacity", "Average"]},
    {"table_name": "concert",
     "columns": ["concert_ID", "Stadium_ID", "
    Year"]}
  ],
  "foreign_keys": [
    {"source_table": "concert", "source_column":
     "Stadium_ID",
     "target_table": "stadium", "target_column":
     "Stadium_ID"}
  ]
}
```

F.2 LLM Output: Structured JSON Intermediate Representation

In compile-style generation, the model does not directly produce SQL text. Instead, it generates a structured JSON representation that encodes the program structure.

```
{
  "type": "query",
  "query": {
    "select": [
      {"expr": {"col": ["stadium", "Name"]}, "
    alias": null},
      {"expr": {"col": ["stadium", "Capacity"]},
     "alias": null}
    ],
    "from": {"table": "stadium", "alias": null},

    "joins": [],
    "where": [],
    "group_by": [],
    "having": [],
```

```
    "order_by": [
      {"expr": {"col": ["stadium", "Average"]},
      "direction": "desc"}
    ],
    "limit": 1,
    "distinct": false
  }
}
```

This JSON structure explicitly represents the logical components of the query, including `select`, `from`, `order_by`, and `limit`.

F.3 Intermediate AST Representation

The compiler then converts the JSON structure into an internal AST:

```
Select(
  columns = [
    Column(table="stadium", name="Name"),
    Column(table="stadium", name="Capacity")
  ],
  from = Table(name="stadium"),
  joins = [],
  where = [],
  group_by = [],
  having = [],
  order_by = [
    OrderBy(
      expr = Column(table="stadium", name="
    Average"),
      direction = DESC
    )
  ],
  limit = 1
)
```

This AST provides a canonical structural representation of the query, which can be analyzed and compared across different generations.

F.4 Compiled SQL Query

Finally, the AST is deterministically compiled into an executable SQL query:

```
SELECT stadium.Name, stadium.Capacity
FROM stadium
ORDER BY stadium.Average DESC
LIMIT 1;
```

This compiled SQL query is then used for execution evaluation and structural analysis in our experiments.