

JOB-COMPLEX: A Challenging Benchmark for Traditional & Learned Query Optimization (Extended Version)

Johannes Wehrstein
Technical University of Darmstadt

Roman Heinrich
Technical University of Darmstadt
DFKI

Timo Eckmann
Technical University of Darmstadt

Carsten Binnig
Technical University of Darmstadt
DFKI

ABSTRACT

Query optimization is a fundamental task in database systems that is crucial to providing high performance. To evaluate learned and traditional optimizer’s performance, several benchmarks, such as the widely used JOB benchmark, are used. However, in this paper, we argue that existing benchmarks are inherently limited, as they do not reflect many real-world properties of query optimization, thus overstating the performance of both traditional and learned optimizers. In fact, simple but realistic properties, such as joins over string columns or complex filter predicates, can drastically reduce the performance of existing query optimizers. Thus, we introduce JOB-COMPLEX, a new benchmark designed to challenge traditional and learned query optimizers by reflecting real-world complexity. Overall, JOB-COMPLEX contains 30 SQL queries and comes together with a plan-selection benchmark containing nearly 6000 execution plans, making it a valuable resource to evaluate the performance of query optimizers and cost models in real-world scenarios. In our evaluation, we show that traditional and learned cost models struggle to achieve high performance on JOB-COMPLEX, providing a runtime of up to 11x slower compared to the optimal plans.

VLDB Workshop Reference Format:

Johannes Wehrstein, Timo Eckmann, Roman Heinrich, and Carsten Binnig. JOB-COMPLEX: A Challenging Benchmark for Traditional & Learned Query Optimization (Extended Version). VLDB 2025 Workshop: Applied AI for Database Systems and Applications (AIDB 2025).

VLDB Workshop Artifact Availability:

The source code, data, and/or other artifacts have been made available at <https://github.com/DataManagementLab/JOB-Complex>.

1 INTRODUCTION

The State of Query Optimization. Efficient query optimization is the backbone of providing high-performance database management systems. For a given SQL statement, the query optimizer has to choose optimally among many logical execution plans. This is highly important, as a bad plan may take orders of magnitude longer than the fastest one [6, 11]. Thus, query optimizers aim to

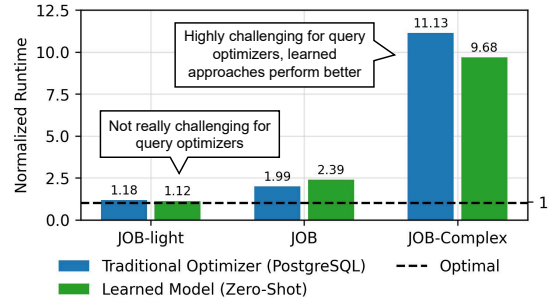


Figure 1: Query optimization performance for PostgreSQL and a selected learned cost model Zero-Shot [7] over JOB-Light, JOB and JOB-COMPLEX (ours). Here, we report the sum of the selected plan runtimes, normalized by the optimal runtime (optimization gap). In contrast to existing benchmarks, JOB-COMPLEX poses significant challenges for query optimization by showing an high optimization gap of up to 11.13.

minimize the query runtime, e.g., by enumerating different join orders or selecting between physical join operators. However, query optimization is a highly challenging task, as the number of plan candidates grows quickly with the query complexity. For that reason, query optimizers have been under constant improvement for decades [9]. While commercial database systems today typically apply static rule-based optimizations and hand-crafted cost models, *machine learning* approaches have been proposed in recent years to further improve query optimization[18, 25], covering Learned Cost Models (LCMs) [7, 13, 19], learned cardinalities [8, 24, 27], learned query hinting [14], and even end-to-end learned query optimizers[15, 26, 29].

Is Query Optimization A Solved Problem (in times of AI)?

To exercise the performance of query optimizers, many benchmarks were proposed, such as JOB-light [10], JOB [11] and TPC-H [22]. Particularly, JOB and JOB-light specifically evaluate the performance of query optimizers and cost models in choosing good join orders (hence the name Join-Order-Benchmark). However, in this paper, we examined classical optimizers like PostgreSQL and learned optimizers such as ZeroShot [7] and our findings, and find, that these optimizers provide already almost optimal performance on these benchmarks, as shown in Figure 1. For instance, on JOB-light, PostgreSQL achieves to select plans only 1.18x slower than the optimal plan which we refer to as an *optimization gap* of 1.18. While JOB is significantly more complex than JOB-light, the overall performance is still remarkably high. Here, PostgreSQL chooses for

50% of the queries a plan that is less than 29% slower than the optimal plan. On the whole benchmark, the optimization gap amounts to only 1.99, as shown in Figure 1. Similarly, LCMs like Zero-Shot [7] achieve a similar quality of plan selection with optimization gaps of 1.12 for JOB-light and 2.39 for JOB. Hence, we want to raise the question in this paper: *Do we need to further invest in query optimization and learned models, or is query optimization an already solved problem in times of AI?*

The Need For A New Benchmark. While these standard benchmarks have been a valuable tool for evaluating query optimizers, our experience with them revealed that they lack many complex real-world challenges. As a result, these benchmarks (while valuable) do not effectively stress query optimizers or expose their limitations. For instance, TPC-H uses synthetic data lacking real-world correlations and complex data distributions, often leading to overly accurate cardinality estimates, as recently discussed [23]. JOB & JOB-light were proposed to address this using the IMDB movie dataset, a real-world dataset with complex correlations. While these benchmarks capture varying levels of query complexity, they, however, still largely operate under simplifying assumptions, such that data exists in a normalized schema or that primary/foreign key (PK/FK) constraints are rigidly enforced. However, the scenarios faced by modern database systems often extend far beyond these assumptions [1, 2, 12, 17, 23]. For instance, users frequently join on non-key columns [1], use string-based join conditions[21], or employ complex filter predicates involving LIKE or IN clauses with many values [1, 2, 12]. Moreover, data might not always be perfectly normalized, and schemas can evolve over time [23]. However, current standard benchmarks fail to adequately capture these properties, limiting their effectiveness in evaluating query optimizers and cost models and often resulting in overly optimistic performance assessments.

Introducing JOB-COMPLEX. To address this gap of realistic benchmarks, we introduce JOB-COMPLEX, a novel benchmark to evaluate query optimization and cost models. JOB-COMPLEX consists of 30 hand-crafted SQL queries that reflect real-world conditions, such as joins on string columns, joins on non-primary keys, and complex filter predicates. Moreover, different from other benchmarks, JOB-COMPLEX includes multiple execution plans for each query, encompassing physical plans along with their estimated and actual costs and cardinalities. This enables a reproducible and comparable evaluation of cost models and query optimizers, making JOB-COMPLEX a valuable resource for benchmarking. To create JOB-COMPLEX, we adapted query patterns from the JOB benchmark, demonstrating how relatively simple modifications can result in a significantly more challenging and realistic query optimization benchmark. As shown in Figure 1, JOB-COMPLEX presents a much greater challenge for both traditional and learned approaches compared to JOB and JOB-Light, with optimization gaps of up to 11.13 for PostgreSQL and 9.68 for the learned ZeroShot model.

Contributions. Overall, we present the following contributions in this paper:

- (1) We introduce JOB-COMPLEX¹, a novel benchmark for query optimization and cost estimation that incorporates real-world

challenges such as joins on non-primary/foreign key columns, joins on string columns, and complex filter predicates.

- (2) We additionally provide a plan selection benchmark that covers a broad range of pre-executed plans for each SQL query of JOB-COMPLEX, as well as JOB and JOB-light. This enables better training of learned approaches and fine-grained evaluation of plan selection.
- (3) In our experimental section, we show that both traditional and learned query optimizers and cost models struggle with the increased complexity of JOB-COMPLEX, indicating considerable room for improvement. Through a detailed evaluation of PostgreSQL and several recent learned cost models, we highlight their respective strengths and weaknesses when applied to these more challenging queries.
- (4) We make JOB-COMPLEX publicly available to foster future research and development in query optimization, encouraging the community to address these identified challenges.

Outline. In the remainder of this paper, we first present the design ideas and specific modifications made to create JOB-COMPLEX in Section 2. We then present a comprehensive evaluation of traditional and learned query optimization techniques on this new benchmark in Section 3. Finally, we conclude our work and outline future research directions in Section 4.

2 JOB-COMPLEX

In the following, we first present JOB-COMPLEX, which is the core contribution of this work in Section 2.1, before we discuss the necessity of plan enumeration datasets in Section 2.2.

2.1 Benchmark Design

In this section, we present the core ideas of JOB-COMPLEX and discuss how it differs from previous benchmarks by representing real-world query optimization challenges, as summarized in Table 1. Overall, our benchmark comes with 30 queries that have between 5-14 joins. Moreover, it contains various real-world properties, that we discuss in the following. By that, it shows an optimization gap of 11.13, posing way harder challenges for query optimization than existing benchmarks.

Real World Properties. To design a novel benchmark that presents a significantly greater challenge for query optimizers than existing ones, we build on the established JOB benchmark [11], addressing its simplifying assumptions through targeted query modifications. This approach enables a meaningful comparison between JOB-COMPLEX and JOB by preserving similar query structures, thereby avoiding superficial changes such as only increasing the number of joins. Different from JOB, however, JOB-COMPLEX incorporates real-world query conditions, as detailed in the following:

- (1) **Joins on Non-Primary/Foreign Key Columns:** Unlike JOB and JOB-Light, which primarily use joins on columns with PK/FK relationships, JOB-COMPLEX includes queries where joins are performed on columns that do not have such constraints. This is a common pattern in real-world analytical queries or when integrating denormalized data. In fact, the absence of PK/FK relationships makes cardinality estimation for these joins significantly harder, as the estimator cannot rely on uniqueness or referential integrity assumptions.

¹<https://github.com/DataManagementLab/JOB-Extended>

Benchmark	Number of Queries	Number of Joins	String Filters	Join on Non-PK/FK columns	Join on String Columns	Runtime of Optimal Plans (s)	Runtime of PG selected Plans (s)	Optimization Potential
JOB-light [10]	70	1-3	—	—	—	2359.72	2795.53	1.18
JOB [11]	113	3-14	✓	—	—	156.79	312.23	1.99
JOB-COMPLEX (ours)	30	5-14	✓	✓	✓	53.34	593.50	11.13

Table 1: Key characteristics of existing benchmarks. JOB-COMPLEX comes with real-world query properties that pose a significant challenge for query optimizers, as reflected in a very high runtime of selected plans by PostgreSQL, that is 11.13x slower than the optimal plans. The runtimes are aggregated over all queries in each benchmark.

- (2) **Joins on String Columns:** Existing benchmarks predominantly use integer columns for join conditions. In contrast, JOB-COMPLEX introduces joins on string columns. String comparisons are computationally more expensive, and the distribution of string values can be more complex and harder to summarize than numerical data, posing additional challenges for selectivity estimation and cost estimation.
- (3) **Comparisons Between Columns of Same Table:** In JOB-COMPLEX, we intentionally added comparisons between columns within the same table, e.g., `cn.name_pcode_nf = cn.name_pcode_sf`. Such appear in data validation checks or complex business logic e.g., start-time before end-time. This is particularly challenging for query optimizers because they typically lack statistics on intra-row correlations and assume column independence, leading to inaccurate selectivity and cardinality estimates. Additionally, such comparisons limit the use of common optimization techniques like index usage or predicate push-down.
- (4) **Complex Filter Predicates:** Similar to JOB, we incorporated more complex filter attributes, such as LIKE predicates for pattern matching on strings and IN clauses with multiple values. Such filters are often simplified or underrepresented in benchmarks. Estimating the selectivity of such predicates accurately is a highly challenging problem.
- (5) **Preservation of Read-Sets and Correlations:** We intentionally kept the set of tables accessed by a query (read-sets) from the original JOB queries. The IMDB dataset underlying JOB is known for its real-world data distributions and high cross-table correlations. That way, JOB-COMPLEX inherits this inherent data complexity, ensuring that the new challenges are layered on top of an already non-trivial data landscape.

Example Query. In the following, we present an example query (Nr. 12) from JOB-COMPLEX, that covers most of those key ideas from JOB-COMPLEX that were previously discussed.

```

SELECT MIN(chn.name), ...
FROM complete_cast cc, comp_cast_type cct1,
     comp_cast_type cct2, ... -- (11 other tables)
WHERE cct1.kind = 'cast'
AND cct2.kind LIKE '%complete%'
AND chn.name IS NOT NULL
AND (chn.name LIKE '%man%'
     OR chn.name LIKE '%Man%') -- Complex Predicates
AND k.keyword IN (...)
AND ... -- (other filters)
AND chn.id = ci.person_role_id -- w/o PK
AND ak.name_pcode_cf=n.name_pcode_cf -- on strings
AND ak.name_pcode_nf=chn.name_pcode_nf -- on strings

```

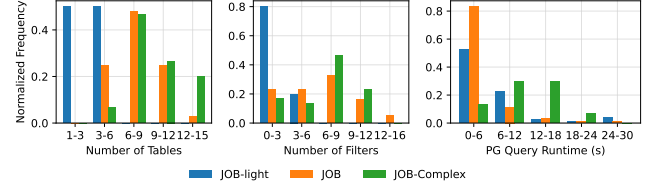


Figure 2: Normalized distribution of number of tables and filter predicates per query across benchmarks. JOB-COMPLEX is similar in the number of tables and filters to existing benchmarks. However, JOB-light & JOB both show a skewed distribution of PostgreSQL runtimes of the queries, while JOB-COMPLEX shows a more symmetric runtime distribution covering both short and long-running queries

AND ... -- (other join conditions)

Benchmark Design. To construct JOB-COMPLEX, we selected 30 representative queries from the JOB benchmark and systematically applied the modifications described above. For instance, an original JOB query that joins on `A.id = B.fk_id` (where `id` is a primary key and `fk_id` is a foreign key) is transformed in JOB-COMPLEX to join on columns such as `A.name = B.title` (string columns without key constraints) or `A.value1 = B.value2` (non-key numerical columns). This preserves the overall SQL complexity by maintaining similar numbers of joins, filters, and aggregations while introducing more realistic and challenging join and filter conditions. By focusing on these targeted modifications, JOB-COMPLEX enables a direct evaluation of how query optimizers cope with real-world complexities rather than simply increasing query size or structural complexity. As illustrated in Figure 2, the distribution of tables and filter predicates per query in JOB-COMPLEX closely mirrors that of JOB, which importantly enables better comparisons. This alignment ensures that any observed differences in optimizer performance are due to the added real-world complexities in JOB-COMPLEX rather than differences in query structure or size.

Additionally, unlike JOB-light, all queries in JOB-COMPLEX are guaranteed to produce non-empty results, preventing optimizers from exploiting shortcuts based on empty intermediate results.

2.2 The Need for Query Plans

Evaluating LCMs and query optimization techniques *end-to-end* requires far more than a single plan per query. In reality, in query optimization scenarios, cost-models will be applied on a large set of query plan candidates, using their cost estimates to select the presumably fastest plan. Hence, to evaluate if a cost model’s cost

estimation leads to the correct ranking and selection of plans, a **diverse set of alternative execution plans** along with their actual execution costs is needed. Therefore, along the 30 SQL queries of JOB-COMPLEX we also provide a meaningful selection of (on average) around 200 enumerated plans per SQL-query as the second key contribution of this work. Providing enumerated and pre-executed plans allows developers of learned cost models and query optimizers to directly apply their cost models, thereby enabling the evaluation of query optimization and plan selection performance on JOB-COMPLEX without the need to manually explore the search space for possible plans or construct a diverse set of plan candidates.

Challenges in Plan Enumeration. However, generating a diverse and representative set of plans is challenging for multiple reasons:

- (1) The exponential growth of the search space with the number of joins. This makes the enumeration of all plans computationally infeasible for queries with more than a few joins.
- (2) The tendency of naive enumeration to produce a skewed runtime distribution, where most plans are either extremely slow or very similar in performance.
- (3) The lack of existing strategies for generating plans with 15+ joins while maintaining a uniform runtime distribution as these queries amplify both previous issues.

Proposed Solution. To address these challenges, we developed a steered plan generation procedure that ensures a diverse and balanced set of plans:

(1) *Oracle-Guided Plan Generation:* We pre-calculated all intermediate cardinalities for all possible join orders (where feasible within a 30-second timeout per sub-plan/join-pair). These actual cardinalities were then injected into the PostgreSQL optimizer by modifying its estimates using [5]. This process helps to identify very good plans, as it mitigates the uncertainty introduced by inaccurate cardinality estimations that are provided by the default optimizer. That way, this strategy aims to find plans close to the true optimal.

(2) *Diversified Random Enumeration:* Since the oracle-guided approach primarily finds good plans, it alone is not sufficient to explore suboptimal but plausible alternatives. Thus, we also randomly enumerated join orders and enforced them in the PostgreSQL optimizer. Though simply enumerating all possible join orders is computationally prohibitive due to exponential growth with the number of joins. In addition, a purely random enumeration often leads to a vast majority of plans being extremely slow and only a few achieving moderate or fast runtimes. To address this problem, we propose a strategy where we only enforce a *global join order* using [5] (i.e., the top-level sequence of joins, e.g., $A \bowtie B \bowtie C$) but allow the PostgreSQL optimizer to choose the best join *nesting* (e.g., $((A \bowtie B) \bowtie C)$ vs. $(A \bowtie (B \bowtie C))$) if the global order allows for it) and physical operators for that enforced global order. Further, we considered only tables with more than ten thousand rows in the global join order. Smaller tables are omitted since they overall do not influence plan runtime significantly but drastically increase the number of possible join orders. This approach ensures a diverse distribution of overall plan structures while leveraging PostgreSQL’s capabilities to optimize local decisions. This approach avoids an overabundance of extremely poor plans and yields a more informative, diverse set of alternatives with a broad runtime spread. By combining oracle-guided and randomized enumeration, we provide, for each query, a representative set of plans (on average 200 per

SQL query), enabling realistic and reproducible evaluation of cost models and plan selection strategies.

3 INITIAL RESULTS USING JOB-COMPLEX

In this section, we use JOB-COMPLEX to stress-test different cost models by comparing the performance of the PostgreSQL cost model for plan selection and several state-of-the-art learned cost models on JOB-COMPLEX against established benchmarks like JOB and JOB-Light. In particular, we analyze (1) the overall query optimization performance, (2) the performance of traditional, (3) and learned plan selection and (4) cardinality estimation accuracy.

Experimental Setup. All queries are executed on PostgreSQL v16 on instances of the academic cloud provider *CloudLab* [3]. For the cost models, in addition to the cost model of PostgreSQL(16) we included a large variety of learned approaches including: DACE [13], ZeroShot [7], QPPNet [16], T3 [19], FlatVector [4], E2E [20], QueryFormer [28], the cost-model of NEO [15] and MSCN [10]. For the training procedures, we follow those described in [6], to ensure a fair comparison by training the models all on the same datasets as much as possible. Database-specific models are trained on the IMDB dataset that is underlying JOB / JOB-light and JOB-COMPLEX, while database-agnostic models were pre-trained on a diverse set of 19 different databases, similar to [7].

3.1 Query Optimization Performance

Our initial analysis reveals that JOB-COMPLEX poses a significantly greater challenge to traditional query optimization than established benchmarks like JOB and JOB-Light. This increased difficulty is visible in the substantially larger *optimization gap* observed with JOB-COMPLEX. Figure 3 illustrates this gap by comparing the runtime of the plan selected by PostgreSQL against the runtime of the optimal plan (from our enumerated set) for each query, sorted by the optimal runtime. A higher optimization gap denotes a greater difficulty for the optimizer in finding the best plan.

The reasons for this larger gap on JOB-COMPLEX are twofold: First, both the absolute and relative differences between the optimal plan runtimes and the PostgreSQL-selected plan runtimes are considerably larger for JOB-COMPLEX. This holds true across the entire spectrum of query execution times. For example, even for queries where the optimal plan executes in under 1 second, PostgreSQL frequently chooses plans on JOB-COMPLEX that are 10 seconds or slower. Second, JOB-COMPLEX features a more uniform runtime distribution, unlike the long-tailed distributions of JOB and JOB-Light, where a few long running queries dominate. This reduces bias and ensures optimizers are tested consistently across both short and long-running queries, making JOB-COMPLEX a more robust challenge.

These findings underscore that query optimization is far from a solved problem, particularly when faced with the real-world complexities captured by JOB-COMPLEX. This naturally leads to the question: Why does PostgreSQL struggle to find optimal, or even near-optimal, plans on JOB-COMPLEX? Is this primarily due to inaccuracies in its cardinality estimations or perhaps limitations within its cost model when dealing with these complex queries? And how do LCMs perform in more challenging scenarios? To examine these

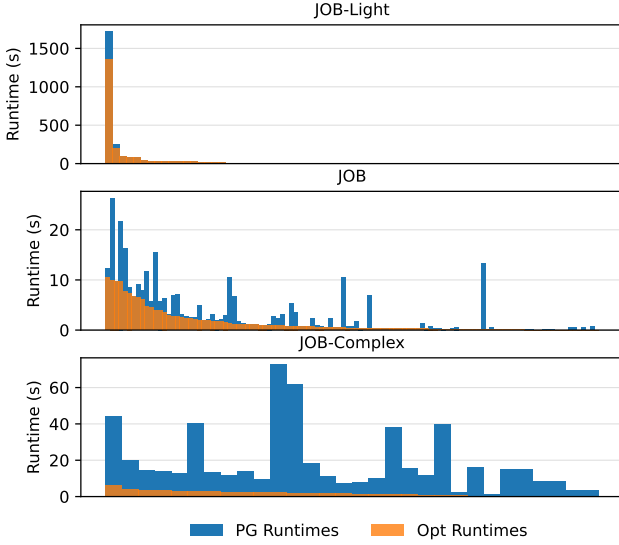


Figure 3: Optimal runtime vs. PostgreSQL selected plan runtime per query of JOB-Light, JOB and JOB-COMPLEX, sorted in descending runtime order. Overall, queries provided in JOB-COMPLEX come with a high optimization potential.

questions, the following sections will analyze the accuracy of plan selection, predicted costs and cardinality estimates.

3.2 How do Traditional Approaches Perform?

In the following, we take a deeper look to understand why JOB-COMPLEX stresses PostgreSQL more. For this, we first analyze the accuracy of cost estimates as understanding the reliability of these estimates on complex queries is vital. Figure 4 shows PostgreSQL’s estimated cost against the actual runtime for enumerated plans from JOB-COMPLEX. For JOB and JOB-Light, there is a tight correlation where estimated costs tend to increase with actual runtimes. However, on JOB-COMPLEX, this correlation is much weaker. While JOB-COMPLEX has a more narrow runtime distribution (a desirable property, as discussed), the cost estimates from PostgreSQL vary by orders of magnitude and show little correlation with actual runtimes. This underscores the difficulty traditional cost models face in capturing the complexities introduced by JOB-COMPLEX. Their struggle is also reflected in the median Q-error of the cost estimation by PostgreSQL as it is substantially higher on JOB-COMPLEX in comparison to JOB-light and JOB as can be seen in Table 2 (1015.67 on JOB, 3.87 on JOB-light and 2669.22 on JOB-COMPLEX).

However, the question about the overall query optimization performance is not answered simply by cost estimation accuracy, as shown by [6]. Therefore, we analyze the optimization potential (selected plan runtime divided by optimal plan runtime) for PostgreSQL as well. On JOB, PostgreSQL achieves an impressively small optimization gap of roughly 2x, as previously shown in Figure 1 and Figure 3. In stark contrast, the same experiment on JOB-COMPLEX reveals that traditional approaches do not consistently find near-optimal plans as the optimization gap is 11x. Since AI models have

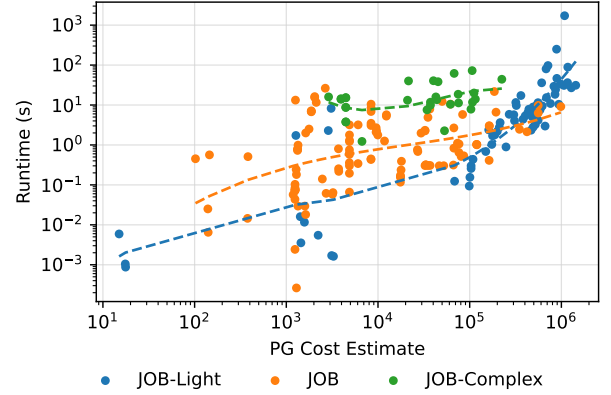


Figure 4: PostgreSQL’s estimated cost vs. actual runtime for queries in JOB-COMPLEX. Each point represents a query plan. The dispersion indicates challenges in accurately modeling costs for the complex queries of JOB-COMPLEX. Ideally, points should lie close to a line with a positive slope, indicating a linear relationship.

shown promising results for query optimization in recent years, we will in the following analyze if LCMs show similar struggles.

3.3 How do AI Approaches Perform?

To evaluate whether AI models show similar or different challenges, we selected nine representative LCMs, encompassing various architectures (GNNs, Transformers, set-based models, etc.) and approaches (database-specific and database-agnostic/zero-shot models like T3, ZeroShot and DACE). Following the same steps to analyze the performance of traditional approaches, we first analyze the median Q-error of cost estimations before discussing the overall optimization gap of LCMs.

When examining the median Q-error of cost estimations for all models computed across all enumerated plans, including optimal and sub-optimal plans, as shown in Table 2, an interesting pattern emerges: Despite JOB-COMPLEX proving significantly harder for plan selection (i.e., a larger optimization gap as shown in Section 3.1), the median Q-errors for cost estimation by LCMs are often comparable or even slightly better on JOB-COMPLEX than on JOB. While accurate cost estimates are desirable, the ability to rank plans correctly for selection is essential.

What about plan selection? Next, we examine how challenging JOB-COMPLEX is for LCMs when tasked with selecting the best execution plan from the enumerated set based on their cost estimates for each plan. The results in Figure 5 show the previously discussed optimization gap for PostgreSQL against the selected LCMs across the three benchmarks. On JOB, most advanced LCMs perform on par with PostgreSQL, showing an optimization potential of around 2 to 2.5. However, consistent with findings in [6], most models, such as E2E, QueryFormer, NEO, and MSCN, show performance challenges. On JOB-Light, PostgreSQL and several LCMs (e.g., ZeroShot, QPPNet, FlatVector) achieve near-optimal performance, with a significantly lower overall optimization potential, while other models fail to provide good plan selections. In

Median Q-Error	JOB	JOB-light	JOB-COMPLEX
Postgres (v16)	1015.67	3.87	2669.22
DACE	1.91	2.34	1.81
E2E	13.21	5.03	23.39
FlatVector	2.15	1.72	1.52
MSCN	13.85	3.38	26.56
NEO	13.20	2.22	4.09
QPPNet	17.50	1.45	67.20
QueryFormer	8.30	3.49	9.88
T3	4.07	1.72	2.30
ZeroShot	1.58	1.42	1.60
Median of all Models	10.75	2.28	6.99

Table 2: Cost Estimation Median Q-Error for Postgres and LCMs on all enumerated plan selection candidates. The increased difficulty in JOB-COMPLEX significantly impacts cost estimation accuracy of PostgreSQL. However, the learned cost models show to be more robust against more complex queries and exhibit a roughly similar Q-error range than JOB.

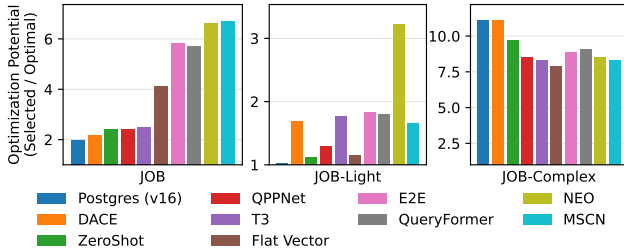


Figure 5: Optimization Potential (selected plan runtime / optimal plan runtime) for PostgreSQL (PG) and various LCMs across JOB-Light, JOB, and JOB-COMPLEX. Higher values indicate poorer cost model performance and greater room for improvement. Overall, JOB-COMPLEX clearly shows a higher optimization potential across all models.

stark contrast to this optimization potential of roughly 2 to 2.5 on JOB, the optimization potential of the best-performing LCMs, such as T3 or FlatVector is roughly 8x on JOB-COMPLEX. Clearly, this gap of 8x compared to the optimal is far from acceptable and highlights that query optimization is not solved, and there is a pressing need for improved cost models to tackle real-world query complexities. However, LCMs seem to perform better than traditional approaches on the more challenging JOB-COMPLEX benchmark (gap of 8x for T3 and FlatVector vs. 11x for Postgres). Motivated by [11] showcasing the huge impact of cardinality estimates on query optimization performance, we investigate in the following whether the increased complexity of JOB-COMPLEX can also be attributed to cardinality estimation challenges.

3.4 How Accurate are Cardinalities?

Finally, we investigate the accuracy of cardinality estimates, which are fundamental inputs for cost models [11], including PostgreSQL’s

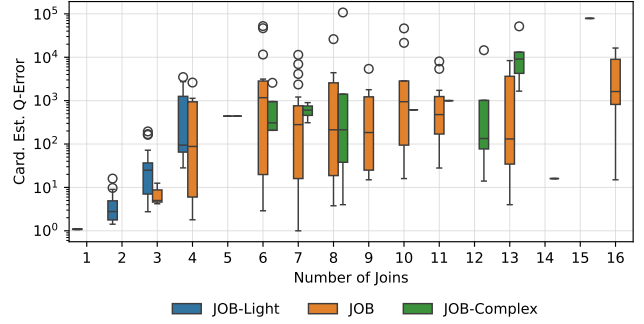


Figure 6: Cardinality estimation error of PostgreSQL (q-error) at the root operator versus the number of tables in the query, compared across JOB-Light, JOB, and JOB-COMPLEX. JOB-COMPLEX consistently exhibits higher q-errors, indicating less accurate cardinality estimates, especially as query complexity increases.

native model and many LCMs (especially zero-shot models). This experiment aims to determine whether the large optimization gap observed on JOB-COMPLEX is primarily due to inaccurate cardinality estimates, or if the cost models themselves are unable to translate even reasonable cardinality estimates into good plan choices. By analyzing the cardinality estimation errors, we can assess whether improving cardinality alone would be sufficient, or if further improvements in cost modeling are necessary. Figure 6 compares the Q-error of PostgreSQL’s cardinality estimates at the root operator, broken down by the number of joins in the query, across the three benchmarks. The results show that while cardinality estimates for JOB-COMPLEX are generally worse (higher Q-error) than for JOB-Light, they are in a somewhat similar range to those for JOB. As expected, JOB-Light, with its simpler queries and fewer complex correlations, yields better cardinality estimates. The fact that the cardinality estimation errors on JOB-COMPLEX are not drastically worse than on JOB, despite the introduction of challenging features like joins on non-key and string columns (notorious for causing estimation difficulties), is significant. It suggests that the large optimization gap observed on JOB-COMPLEX does not solely stem from catastrophically worse cardinality estimates, as often suggested [11]. Instead, it points towards inherent difficulties within the cost models themselves – both traditional and learned – in generalizing to and accurately costing more complex query patterns and operator inter-dependencies, even when underlying cardinality estimates are not excessively erroneous. This indicates that improvements are needed not just rooted in cardinality estimation but also in how these estimates (and other plan features) are used by cost functions to predict final plan costs and guide plan selection.

3.5 How Well do Cost Models Rank?

Beyond the accuracy of cost estimates, effective plan selection critically depends on the model’s ability to rank candidate plans correctly. Even if absolute cost predictions are largely accurate, a model that does not preserve the correct ordering of plans (especially for the fast running plans) cannot guide the optimizer towards

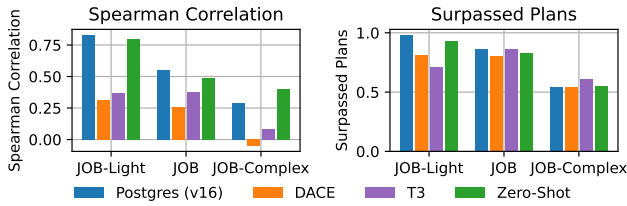


Figure 7: Correlation and Surpassed-Plans across JOB-Light, JOB, and JOB-COMPLEX. Correlation describes the spearman correlation of the estimated runtime and the actual runtime. Surpassed plans [6] measures the relative rank of a chosen plan by showing what fraction of alternative plans it outperforms (e.g., 0.5 means half are worse and half are better). JOB-COMPLEX exhibits a drastically decreasing correlation compared to JOB-light and JOB. Similarly surpassed-plans drops significantly to 0.55-0.6 which is only slightly better than random plan selection (0.5).

near-optimal choices. In other words, cost models must not only be accurate regressors, but also robust sorting functions. Achieving success on JOB-COMPLEX therefore requires models to consistently assign lower costs to plans with lower runtimes, ensuring that the best-performing plan is selected.

Figure 7 illustrates this challenge by reporting both Spearman correlation and surpassed-plans[6] scores. On JOB-Light, cost models achieve high correlations (often above 0.8), indicating that the relative ordering of plans is well preserved. On JOB, correlations already decline substantially, and on JOB-COMPLEX all approaches deteriorate severely, reaching at most a Spearman correlation of 0.4. This collapse in rank correlation directly reflects the difficulty of ordering plans correctly in more complex workloads. The surpassed-plans metric (relative rank of picked plan) confirms this picture: while models perform reasonably well on JOB-Light and JOB, on JOB-COMPLEX they achieve only 0.55–0.6, barely better than random selection (0.5).

Taken together, these results highlight that the central bottleneck exposed by JOB-COMPLEX is not merely inaccurate cost and cardinality estimation, but the inability of both traditional and learned cost models to produce reliable plan rankings in complex query spaces. Thus, the challenge is not only to reduce overall cost prediction errors, but to design models that are fundamentally better at ranking plan candidates in the presence of intricate operator interactions and non-trivial query structures.

4 CONCLUSION AND FUTURE WORK

In this work, we demonstrated that query optimization, particularly for complex Select-Project-Aggregate-Join (SPAJ) queries covering real-world properties, is far from a solved problem. While existing benchmarks like JOB and JOB-Light suggest that systems like PostgreSQL or query optimization built on learned cost models achieve acceptable or even near-optimal query optimization performance, our findings indicate that this is an incomplete picture. Thus, we introduced JOB-COMPLEX, a novel benchmark derived from JOB, specifically designed to incorporate these real-world challenges.

By introducing realistic complexities, such as joins on non-key columns, string-based joins, and intricate filter predicates, we have shown that the optimization gap can quickly exceed 10x. On JOB-COMPLEX, PostgreSQL exhibits an 11x average optimization gap, and even state-of-the-art LCMs show a significant gap of around 8x. Thus, our results highlight substantial room for improvement in both traditional, cost-model-based query optimizers and emerging learned approaches. Interestingly, while LCMs struggle with the increased complexity of JOB-COMPLEX as well, they tend to outperform traditional cost models. This suggests that LCMs might possess a competitive advantage in handling more intricate query patterns, which, however requires further investigation. Nevertheless, both paradigms are far from achieving optimal performance on these more challenging queries, signaling a clear need for new research into more robust and accurate cost estimation and plan selection techniques.

Future Work. Our work represents a first step towards developing more realistic and challenging query optimization benchmarks. Thus, it opens up several promising directions for future research.

- (1) **Increase Complexity of JOB-COMPLEX:** The first direction is to extend JOB-COMPLEX to cover an even broader range of real-world complexities, such as queries involving materialized views, semi-structured or nested data (e.g., JSON, XML), user-defined functions, and a wider spectrum of analytical query patterns (e.g., window functions, common table expressions).
- (2) **Analyze Failures of Cost Models:** Second, we aim to conduct a deeper root cause analysis for the suboptimal plans selected by traditional and learned approaches on JOB-COMPLEX to better understand *why* they selected those plans. However, this is particularly challenging for learned approaches, as they leverage deep neural networks and thus are not yet *explainable*.
- (3) **Advance Learned Approaches:** Next, we see a large potential in learned approaches, as they have proven to provide highly accurate estimates in general. Thus, we suggest to further investigate the development of more sophisticated learned models, covering learned query plan representation, learned plan searching, and hybrid approaches that strategically combine the strengths of traditional optimizers with learned components. By providing various plan enumerations for the same queries in JOB-COMPLEX, we importantly enable the development of such approaches.
- (4) **Evaluating Query Optimization Across Systems:** Moreover, we propose to evaluate JOB-COMPLEX and its future extensions across a wider range of database systems, including open-source and commercial DBMSs, to validate the generality of our findings and identify system-specific strengths, weaknesses, and optimization strategies.
- (5) **Generate Hard Queries Automatically:** Finally, the creation of challenging benchmarks is a non-trivial task that requires expert knowledge and testing. Thus, we propose to explore automated techniques for identifying and generating "hard" queries that maximally stress-test query optimizers beyond current benchmarks, which could be addressed with the help of LLMs, adversarial generation, or evolutionary algorithms.

Ultimately, we aim to contribute to the development of more robust, effective, and truly general query optimizers capable of handling the full spectrum and complexity of real-world workloads.

ACKNOWLEDGMENTS

This work has been supported by the LOEWE program (Reference III 5-519/05.00.003-(0005)), hessian.AI at TU Darmstadt, as well as DFKI Darmstadt.

REFERENCES

- [1] Bailu Ding, Surajit Chaudhuri, Johannes Gehrke, and Vivek Narasayya. 2021. DSB: A decision support benchmark for workload-driven and traditional database systems. *Proceedings of the VLDB Endowment* 14, 13 (2021), 3376–3388.
- [2] Markus Dreseler, Martin Boissier, Tilmann Rabl, and Matthias Uflacker. 2020. Quantifying TPC-H choke points and their optimizations. *Proceedings of the VLDB Endowment* 13, 8 (2020), 1206–1220.
- [3] Dmitry Duplyakin, Robert Ricci, Aleksander Maricq, Gary Wong, Jonathon Duerig, Eric Eide, Leigh Stoller, Mike Hibler, David Johnson, Kirk Webb, Aditya Akella, Kuang-Ching Wang, Glenn Ricart, Larry Landweber, Chip Elliott, Michael Zink, Emmanuel Cecchet, Snigdhaswin Kar, and Prabodh Mishra. 2019. The Design and Operation of CloudLab. In *Proceedings of the 2019 USENIX Annual Technical Conference, USENIX ATC 2019, Renton, WA, USA, July 10-12, 2019*, Dahlia Malkhi and Dan Tsafir (Eds.). USENIX Association, 1–14. <https://www.usenix.org/conference/atc19/presentation/duplyakin>
- [4] Archana Ganapathi, Harumi A. Kuno, Umeshwar Dayal, Janet L. Wiener, Armando Fox, Michael I. Jordan, and David A. Patterson. 2009. Predicting Multiple Metrics for Queries: Better Decisions Enabled by Machine Learning. In *Proceedings of the 25th International Conference on Data Engineering, ICDE 2009, March 29 2009 - April 2 2009, Shanghai, China*, Yannis E. Ioannidis, Dik Lun Lee, and Raymond T. Ng (Eds.). IEEE Computer Society, 592–603. <https://doi.org/10.1109/ICDE.2009.130>
- [5] Naruhiko Hayashi and contributors. 2025. `pg_hint_plan`: PostgreSQL hinting engine. https://github.com/oss-c-db/pg_hint_plan. Accessed: 2025-06-10.
- [6] Roman Heinrich, Manisha Luthra, Johannes Wehrstein, Harald Kornmayer, and Carsten Binnig. 2025. How Good are Learned Cost Models, Really? Insights from Query Optimization Tasks. *Proc. ACM Manag. Data* 3, 3, Article 172 (June 2025), 27 pages. <https://doi.org/10.1145/3725309>
- [7] Benjamin Hilprecht and Carsten Binnig. 2022. Zero-Shot Cost Models for Out-of-the-box Learned Cost Prediction. *Proc. VLDB Endow.* 15, 11 (2022), 2361–2374. <https://doi.org/10.14778/3551793.3551799>
- [8] Benjamin Hilprecht, Andreas Schmidt, Moritz Kulesa, Alejandro Molina, Kristian Kersting, and Carsten Binnig. 2020. DeepDB: Learn from Data, not from Queries! *Proc. VLDB Endow.* 13, 7 (2020), 992–1005. <https://doi.org/10.14778/3384345.3384349>
- [9] Matthias Jarke and Jürgen Koch. 1984. Query Optimization in Database Systems. *ACM Comput. Surv.* 16, 2 (1984), 111–152. <https://doi.org/10.1145/356924.356928>
- [10] Andreas Kipf, Thomas Kipf, Bernhard Radke, Viktor Leis, Peter A. Boncz, and Alfons Kemper. 2019. Learned Cardinalities: Estimating Correlated Joins with Deep Learning. In *9th Biennial Conference on Innovative Data Systems Research, CIDR 2019, Asilomar, CA, USA, January 13-16, 2019, Online Proceedings*. www.cidrdb.org, <http://cidrdb.org/cidr2019/papers/p101-kipf-cidr19.pdf>
- [11] Viktor Leis, Andrey Gubichev, Atanas Mirchev, Peter A. Boncz, Alfons Kemper, and Thomas Neumann. 2015. How Good Are Query Optimizers, Really? *Proc. VLDB Endow.* 9, 3 (2015), 204–215. <https://doi.org/10.14778/2850583.2850594>
- [12] Viktor Leis, Bernhard Radke, Andrey Gubichev, Atanas Mirchev, Peter Boncz, Alfons Kemper, and Thomas Neumann. 2018. Query optimization through the looking glass, and what we found running the join order benchmark. *The VLDB Journal* 27 (2018), 643–668.
- [13] Zibo Liang, Xu Chen, Yuyang Xia, Runfan Ye, Haitian Chen, Jiandong Xie, and Kai Zheng. 2024. DACE: A Database-Agnostic Cost Estimator. In *40th IEEE International Conference on Data Engineering, ICDE 2024, Utrecht, The Netherlands, May 13-16, 2024*. IEEE, 4925–4937. <https://doi.org/10.1109/ICDE60146.2024.00374>
- [14] Ryan Marcus, Parimarjan Negi, Hongzi Mao, Nesime Tatbul, Mohammad Alizadeh, and Tim Kraska. 2020. Bao: Learning to Steer Query Optimizers. *CoRR abs/2004.03814* (2020). arXiv:2004.03814 <https://arxiv.org/abs/2004.03814>
- [15] Ryan Marcus, Parimarjan Negi, Hongzi Mao, Chi Zhang, Mohammad Alizadeh, Tim Kraska, Olga Papaemmanouil, and Nesime Tatbul. 2019. Neo: A Learned Query Optimizer. *Proc. VLDB Endow.* 12, 11 (2019), 1705–1718. <https://doi.org/10.14778/3342263.3342644>
- [16] Ryan Marcus and Olga Papaemmanouil. 2019. Plan-Structured Deep Neural Network Models for Query Performance Prediction. *Proc. VLDB Endow.* 12, 11 (2019), 1733–1746. <https://doi.org/10.14778/3342263.3342646>
- [17] Parimarjan Negi, Laurent Bindshaedler, Mohammad Alizadeh, Tim Kraska, Jyoti Leeka, Anja Gruenheid, and Matteo Interlandi. 2023. Unshackling database benchmarking from synthetic workloads. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. IEEE, 3659–3662.
- [18] Parimarjan Negi, Matteo Interlandi, Ryan Marcus, Mohammad Alizadeh, Tim Kraska, Marc T. Friedman, and Alekh Jindal. 2021. Steering Query Optimizers: A Practical Take on Big Data Workloads. In *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, Guoliang Li, Zhanhui Li, Stratos Idreos, and Divesh Srivastava (Eds.). ACM, 2557–2569. <https://doi.org/10.1145/3448016.3457568>
- [19] Maximilian Rieger and Thomas Neumann. 2025. T3: Accurate and Fast Performance Prediction for Relational Database Systems With Compiled Decision Trees. *Proceedings of the ACM on Management of Data* 3, 3 (2025), 1–27.
- [20] Ji Sun and Guoliang Li. 2019. An End-to-End Learning-based Cost Estimator. *Proc. VLDB Endow.* 13, 3 (2019), 307–319. <https://doi.org/10.14778/3368289.3368296>
- [21] Jan Vincent Szlang, Sebastian Bress, Sebastian Cattes, Jonathan Dees, Florian Funke, Max Heime, Michel Oleynik, Ismail Oukid, and Tobias Maltenberger. 2025. Workload Insights From The Snowflake Data Cloud: What Do Production Analytic Queries Really Look Like. *Proc. VLDB Endow.* 18, 11 (2025).
- [22] Transaction Processing Performance Council. 2010. TPC Benchmark H (Decision Support). <http://www.tpc.org/tpch/>. Standard Specification, Revision 2.17.1.
- [23] Alexander van Renen, Dominik Horn, Pascal Pfeil, Kapil Vaidya, Wenjian Dong, Murali Narayanaswamy, Zhengchun Liu, Gaurav Saxena, Andreas Kipf, and Tim Kraska. 2024. Why TPC Is Not Enough: An Analysis of the Amazon Redshift Fleet. *Proc. VLDB Endow.* 17, 11 (2024), 3694–3706. <https://doi.org/10.14778/3681954.3682031>
- [24] Johannes Wehrstein, Carsten Binnig, Fatma Özcan, Shobha Vasudevan, Yu Gan, and Yawen Wang. [n.d.]. Towards Foundation Database Models. ([n.d.]).
- [25] Ziniu Wu, Ryan Marcus, Zhengchun Liu, Parimarjan Negi, Vikram Nathan, Pascal Pfeil, Gaurav Saxena, Mohammad Rahman, Balakrishnan Narayanaswamy, and Tim Kraska. 2024. Stage: Query Execution Time Prediction in Amazon Redshift. In *Companion of the 2024 International Conference on Management of Data, SIGMOD/PODS 2024, Santiago AA, Chile, June 9-15, 2024*, Pablo Barceló, Nayat Sánchez-Pi, Alexandra Meliou, and S. Sudarshan (Eds.). ACM, 280–294. <https://doi.org/10.1145/3626246.3653391>
- [26] Zongheng Yang, Wei-Lin Chiang, Sifei Luan, Gautam Mittal, Michael Luo, and Ion Stoica. 2022. Balsa: Learning a Query Optimizer Without Expert Demonstrations. In *SIGMOD '22: International Conference on Management of Data, Philadelphia, PA, USA, June 12 - 17, 2022*, Zachary G. Ives, Angela Bonifati, and Amr El Abbadi (Eds.). ACM, 931–944. <https://doi.org/10.1145/3514221.3517885>
- [27] Zongheng Yang, Amog Kamsetty, Sifei Luan, Eric Liang, Yan Duan, Xi Chen, and Ion Stoica. 2020. NeuroCard: One Cardinality Estimator for All Tables. *Proc. VLDB Endow.* 14, 1 (2020), 61–73. <https://doi.org/10.14778/3421424.3421432>
- [28] Yue Zhao, Gao Cong, Jiachen Shi, and Chunyan Miao. 2022. QueryFormer: A Tree Transformer Model for Query Plan Representation. *Proc. VLDB Endow.* 15, 8 (2022), 1658–1670. <https://doi.org/10.14778/3529337.3529349>
- [29] Rong Zhu, Wei Chen, Bolin Ding, Xingguang Chen, Andreas Pfadler, Ziniu Wu, and Jingren Zhou. 2023. Lero: A Learning-to-Rank Query Optimizer. *Proc. VLDB Endow.* 16, 6 (2023), 1466–1479. <https://doi.org/10.14778/3583140.3583160>