Measuring Cross-Cohort Generalization Under Distribution Shift in SQL Representations

Pablo Rivas and Donald R. Schwartz

Department of Computing Technology, School of Computer Science and Mathematics, Marist University, 3399 North Road, Poughkeepsie, NY 12601, USA
{Pablo.Rivas,Donald.Schwartz}@Marist.edu

Abstract. Predictive models built from program text can appear reliable while relying on features that change when data are collected in a new setting. This issue is central for SQL submissions, where table names, schemas, and assignment batches may identify the source of an example as strongly as its query structure indicates its outcome. Here we show that representations of 6,980 SQL submissions from three cohorts contain substantial source-associated signal, and that masking identifiers and literals changes transfer behavior. Across three frozen representation families, raw SQL made schema identity highly recoverable, whereas masked SQL reduced this signal but did not remove cohort or source information. For the optimized multitask CNN, masking improved mean performance when one cohort was held out for correctness, remark-label, and numeric-grade prediction, although raw SQL remained stronger within cohort. Greater representation distance was also associated with larger loss in numeric-grade transfer. These findings show that evaluation confined to one data source can overstate transfer performance and that cross-cohort tests are needed before SQL outcome predictions are used with newly collected submissions.

Keywords: SQL embeddings · distribution shift · representation analysis · cross-cohort transfer · code representations

1 Introduction

SQL outcome prediction is useful only to the extent that predictions remain credible when a new collection of submissions differs from the data used for fitting. In practice, SQL corpora can change when database schemas, assignment variants, query patterns, or source batches change. A model may then learn cues that are predictive within one collection but less reliable under transfer. This issue matters for the prediction of correctness, remark labels, and numeric grade because those outputs can be interpreted too confidently when the evaluation is limited to random partitions of one source pool.

Our earlier SQL work established a sequence of learned prediction models: CNN prediction based on attention, CNN prediction based on multi-output, transformer-based modeling, and analysis oriented to explanation [10,14,13,15,12].

That line culminates in the previously published optimized multitask CNN of Rivas and Schwartz (2025) [11]. We ask how schema or source-batch change is expressed in the representation space and related to held-out outcome-prediction performance. We address this question conservatively as *schema/source-associated* shift. The observed cohorts vary jointly in schema, source batch, query structure, and label support, so the study tests are transferred under real corpus additions rather than isolating a sole causal factor.

We analyze 6,980 SQL submissions across three cohorts, seven schemas, and nine source groups. For each submission, we compare cleaned SQL with identifiers preserved against an aligned identifier-and-literal-masked form. We examine TF–IDF/SVD, CodeBERT, and MiniLM representations, linear source-recovery and outcome probes, and the optimized multitask CNN trained within valid fixed partitions. The primary evaluation is leave-one-cohort-out (LOO) transfer, complemented by a bounded same-schema source-batch control over the repeated `tv` schema (the fixed-`tv` control) and a representation-distance analysis.

The evidence supports four contributions:

1. We show that raw SQL representations make schema identity nearly recoverable across all three frozen representations, while masking sharply reduces this direct signal but leaves measurable cohort and source-group information.
2. We show that, for the optimized multitask CNN, masking improves mean LOO prediction for correctness, remark labels, and numeric grade, although raw SQL remains stronger under within-cohort evaluation.
3. We show that residual transfer loss persists across source batches that share the `tv` schema, which limits any reading based on schema vocabulary alone.
4. We show that numeric-grade transfer loss is positively associated with representation distance across TF–IDF/SVD, CodeBERT, and MiniLM distance spaces in the studied cross-domain comparisons.

Taken together, these results argue for source-aware evaluation before outcome predictions are interpreted on newly collected SQL data. They do not support replacing local validation or human judgment with model outputs.

The remainder of this paper is organized as follows. Section 2 reviews previous work; Sections 3–5 define the study and tests; Section 6 reports results; and Sections 7 and 8 address limits and implications.

2 Related Work and Prior SQL Prediction Studies

SQL assessment and outcome prediction. SQL submissions admit multiple semantically valid forms, while errors can arise from predicates, joins, grouping, nesting, or schema references. Earlier SQL assessment systems addressed correctness checking, feedback, and partial-credit analysis through test generation or structural comparison [6,2,1]. More recently, bounded SQL-equivalence checking has been presented as a complementary route for validating complex queries under integrity constraints [5]. These methods establish the importance of SQL evaluation, but do not answer whether learned outcome-prediction models remain stable when their source distribution changes.

Prior project line. This study extends a six-work project line. The first study used an attention-based CNN to predict SQL correctness [10]. A later multi-output CNN expanded the prediction setting to correctness, remark labels, and numeric grade [14]. A BERT-based model then examined transformer representations for SQL outcomes [13], while an attribution study and a later multi-task CNN study addressed explanation of predictions [15,12]. A sixth published conference paper reports the optimized multitask CNN evaluated here [11]. The present paper changes the central question from prediction and explanation within accumulated data to transfer under schema/source-associated cohort change.

Representations and cross-schema SQL. CodeBERT learns code-oriented representations from programming-language and natural-language/code data [3]; therefore, we use it as the primary frozen code representation, with TF–IDF/SVD and MiniLM as comparisons. Cross-schema generalization is also central in text-to-SQL work: Spider separates databases across train and test so that models must handle unseen schemas and unseen queries [17]. This neighboring task motivates attention to schema variation, but text-to-SQL generation and outcome prediction from submitted SQL are different problems.

Shift-aware evaluation. Distribution-shift research treats performance on source distributions not seen during fitting as a separate evaluation problem, and emphasizes that model-selection policy affects how such results should be interpreted [4,7]. Our leave-one-cohort-out design and fixed-`tv` check follow that reasoning: they measure transfer across observed source changes without claiming that schema names alone cause the effect.

3 Study Setting and Data

This section describes the corpus, targets, SQL controls, and interpretation scope.

3.1 Corpus, cohorts, and source structure

We study 6,980 SQL submissions comprising 5,897 unique normalized SQL strings. The corpus was assembled through three data additions, which we treat as *early*, *middle*, and *late* cohorts. These groups are not random partitions of a single assignment pool: they encode changes in database schema, source batch, and query form. The early cohort contains 225 submissions from the `movie_person` schema; the middle cohort contains 4,270 submissions from four schemas; and the late cohort contains 2,485 submissions from three schemas. Overall, the data contain seven schemas and nine source groups.

Table 1 gives composition statistics. The early cohort is small and structurally distinct: its queries use more tokens and substantially more joins on average than the later groups. The `tv` schema is the one repeated schema across data additions: it occurs in two middle-cohort source batches and one late-cohort source batch. We use this overlap later for a bounded fixed-schema source-batch check.

Table 1. Dataset composition and query characteristics by cohort. Correct is the percentage labeled correct; tokens and joins are per-submission means. Remark labels include two very rare classes: *Non-Interpretable* ($n = 31$) and *Cheating* ($n = 2$).

Cohort	Rows	Schemas	Correct (%)	Grade	Tokens	Joins
Early	225	1	58.2	79.0	48.8	0.476
Middle	4,270	4	58.1	85.8	43.1	0.031
Late	2,485	3	59.0	87.5	37.9	0.033
Total	6,980	7	58.4	86.2	41.4	0.046

3.2 Prediction targets and SQL input controls

Each submission has three prediction targets: binary correctness, a categorical remark label, and a numeric grade. Correctness prevalence is similar across the broad cohorts (58.2% in early, 58.1% in middle, and 59.0% in late), but remark support is not balanced. The corpus contains 4,076 *Correct* and 2,871 *Partially Correct* remarks, compared with only 31 *Non-Interpretable* and 2 *Cheating* remarks. We retain the all-class remark endpoint as defined in the study, while later reporting a two-supported-class diagnostic to prevent the rare labels from carrying an overly broad interpretation.

For each submission, we compare two aligned SQL inputs. The `raw_clean` variant preserves cleaned SQL with schema names intact. The `schema_anon` variant masks schema identifiers and literals. This control removes much of the direct lexical source cue while preserving query-level structure required for outcome-prediction analysis.

3.3 Scope of the shift

The paper evaluates *schema/source-associated* shift rather than a sole-cause effect of schema naming. Cohort, source batch, schema, and query structure vary together in the observed corpus. E.g., the mean number of joins is 0.476 in early submissions but approximately 0.03 in both middle and late submissions. Thus, held-out transfer tests measure whether prediction models generalize across real corpus additions; the fixed-`tv` check asks more narrowly whether batch-associated shift remains measurable when the schema label is held constant.

4 Methods

Next, we discuss representations, prediction models, and distance-to-loss tests.

4.1 SQL inputs and representation families

Let x_i denote the cleaned SQL submission for instance i, and let $a(\cdot)$ replace schema identifiers and literals with anonymous symbols. We evaluate the inputs

$$x_i^{\text{raw}} = x_i, \qquad x_i^{\text{anon}} = a(x_i). \tag{1}$$

This paired design measures sensitivity to removing explicit lexical source cues while preserving the query structure that remains after masking; it does not isolate a causal effect of schema naming.

We use three representation families. The lexical baseline concatenates word TF–IDF features with 1–3-grams and character TF–IDF features with 3–5-grams, followed by truncated SVD to 384 dimensions. CodeBERT is the primary pretrained code representation because it was designed to learn general-purpose representations from programming and natural-language/code data [3]. For its final hidden states $\mathbf{h}_{it}$ and attention mask m_{it}, we use masked mean pooling:

$$\mathbf{z}_i = \frac{\sum_{t=1}^{T_i} m_{it}\mathbf{h}_{it}}{\sum_{t=1}^{T_i} m_{it}}. \tag{2}$$

CodeBERT outputs 768-dimensional vectors from sequences truncated to 256 tokens. Because CodeBERT is code-oriented rather than SQL-specific, its usefulness here is evaluated by the present probes and transfer tests rather than assumed from pretraining. As a general sentence-representation comparison, we use `all-MiniLM-L6-v2`, producing 384-dimensional embeddings with a maximum input length of 256 tokens. This comparison combines a compact MiniLM encoder [16] with sentence-embedding use motivated by cosine-based comparison and clustering [9]. Frozen embeddings are computed once per input variant; no prediction-target label is used during embedding extraction.

4.2 Source recoverability and local representation structure

A representation may encode source information even when a two-dimensional projection visually overlaps. We therefore make formal source-signal statements from high-dimensional tests rather than from UMAP alone. For group label g_i and the set $\mathcal{N}_k(i)$ of k nearest neighbors of sample i, local same-group purity is

$$\text{Purity}_k = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k}\sum_{j\in\mathcal{N}_k(i)}\mathbb{I}[g_j = g_i], \qquad k \in \{5, 10, 25\}. \tag{3}$$

We compute purity, silhouette scores, and centroid distances for cohort, schema, and source-group labels. We also fit linear source-recovery models to predict each of these labels from frozen embeddings under stratified five-fold cross-validation. Logistic regression is the primary reported recovery model; a linear SVM is retained as a support check. Macro F1 is primary because group sizes are uneven.

4.3 Frozen outcome probes

Frozen outcome probes measure how readily each representation supports the three prediction targets without end-to-end encoder fitting. For correctness, a weighted logistic probe estimates

$$p\left(y_i^{(c)} = 1 \mid \mathbf{z}_i\right) = \sigma\left(\mathbf{w}^\top \mathbf{z}_i + b\right). \tag{4}$$

We use a weighted multiclass linear probe for remarks and ridge regression for numeric grade. Each probe is fitted separately within every fixed split and text variant. Thus, transfer differences reflect the available frozen representation and the train/test source relation, not adaptation of the pretrained encoder to held-out labels.

4.4 Optimized multitask CNN

We evaluate the published optimized multitask CNN of Rivas and Schwartz (2025) [11], which extends prior attention/convolution-based SQL outcome prediction [10,14,12]. Its configuration uses a 16,384-token vocabulary, maximum length capped at 256, a 1,024-dimensional learned token embedding, 512 convolution filters for each width in $\{3, 4, 5, 7\}$, global max pooling, a 512-unit ReLU bottleneck, and dropout 0.0562. The profile uses Adam with learning rate 1.93×10^{-4}, batch size 128, early stopping with patience 15 and best-weight restoration, and learning-rate reduction with patience 5.

The CNN jointly estimates correctness, one-hot remark outputs, and numeric grade. Its reported objective is

$$\mathcal{L} = 0.142\,\mathcal{L}^{(c)}_{\mathrm{BCE}} + 0.740\,\mathcal{L}^{(r)}_{\mathrm{BCE}} + 0.118\,\mathcal{L}^{(g)}_{\mathrm{MSE}}. \tag{5}$$

The remark output is evaluated as a single-label categorical endpoint using macro F1, even though the published configuration uses a binary-cross-entropy-coded remark loss. This design choice is stated directly because rare remark classes make its interpretation consequential.

This previously published configuration is evaluated over the required within-cohort and LOO runs. Because its prior design included reported split conditions, these results are not framed as independent confirmation.

4.5 Representation distance and transfer loss

For a cohort or source group A, let $\boldsymbol{\mu}_A$ be the mean frozen representation. We quantify shift using cosine distance between group centroids:

$$\boldsymbol{\mu}_A = \frac{1}{|A|}\sum_{i \in A} \mathbf{z}_i, \qquad \delta(A, B) = 1 - \frac{\boldsymbol{\mu}_A^{\top}\boldsymbol{\mu}_B}{\|\boldsymbol{\mu}_A\|_2 \|\boldsymbol{\mu}_B\|_2}. \tag{6}$$

Transfer loss is defined relative to the within-target baseline:

$$\begin{aligned} \Delta_{\mathrm{BA}} &= \mathrm{BA}_{\mathrm{within}} - \mathrm{BA}_{\mathrm{cross}}, \\ \Delta_{\mathrm{F1}} &= \mathrm{F1}_{\mathrm{within}} - \mathrm{F1}_{\mathrm{cross}}, \\ \Delta_{\mathrm{MAE}} &= \mathrm{MAE}_{\mathrm{cross}} - \mathrm{MAE}_{\mathrm{within}}. \end{aligned} \tag{7}$$

Positive values therefore indicate degraded transfer for all three endpoints. Our main association statistic is Spearman rank correlation,

$$\rho_S = \mathrm{corr}(\mathrm{rank}(\delta), \mathrm{rank}(\Delta)), \tag{8}$$

which tests monotone distance-to-loss association without assuming a linear scale for representation distance.

5 Evaluation Protocol

This section sets the splits, endpoints, bounded control, and reporting rules.

5.1 Fixed splits and bounded control

All models and SQL variants use split files generated once with seed 42. Within-cohort splits estimate in-source performance using approximately 80%/10%/10% train/validation/test partitions. Leave-one-cohort-out (LOO) evaluation is the primary transfer test: two cohorts form the fitting pool and the complete third cohort is held out for testing (225 early, 4,270 middle, or 2,485 late submissions). Pairwise cohort transfer is used for frozen-probe and distance-to-loss support analyses, not for the primary CNN comparison. Stratification uses remark labels where support permits; as the late test set contains only the two well-supported remark classes, all-class remark comparisons are interpreted with the rare-label limit stated explicitly.

The split audit found zero sample-identifier overlap and zero normalized-SQL-hash overlap across train, validation, and test roles in all twelve fixed split settings. The bounded fixed-`tv` control additionally removes matching normalized hashes before fitting each directed source-batch comparison. It tests whether shift remains under one repeated schema label; it is not a causal isolation test.

5.2 Endpoints and reporting policy

Frozen probes use the three frozen representation families under fixed partitions. The optimized multitask CNN [11] is trained only within the required within-cohort and LOO partitions. Because its prior design included reported split conditions, these results are not presented as independent confirmation.

We report balanced accuracy (BA) for correctness, macro F1 for remark labels, and mean absolute error (MAE) for numeric grade. The all-class remark endpoint is retained, with the prespecified supported-class diagnostic reported alongside its interpretation:

$$\mathrm{BA} = \frac{\mathrm{TPR} + \mathrm{TNR}}{2}, \qquad \mathrm{F1}_{\mathrm{macro}} = \frac{1}{C}\sum_{c=1}^{C} \mathrm{F1}_c, \qquad \mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|. \quad (9)$$

Distance-to-loss results compute Pearson and Spearman correlations, treating Spearman correlation as primary. The main frozen-probe analysis contains nine cross-domain comparisons per variant; CNN distance correlations contain only three LOO targets per variant and are treated as exploratory.

The recorded environment includes scikit-learn 1.8.0, transformers 5.9.0, torch 2.11.0 with CUDA 12.8 support, sentence-transformers 5.5.1, TensorFlow 2.21.0, and Keras 3.14.1. The reported analyses concern aggregate performance and source-associated shift.

Table 2. Five-fold logistic-regression macro F1 for recovering source-associated labels from frozen SQL representations. Anon. masks identifiers and literals. Values are macro F1; larger values indicate stronger label recoverability.

	TF–IDF/SVD		CodeBERT		MiniLM	
Target	Raw	Anon.	Raw	Anon.	Raw	Anon.
Schema	0.999	0.492	0.995	0.511	1.000	0.499
Cohort	0.986	0.647	0.980	0.666	0.987	0.652
Source group	0.902	0.464	0.892	0.499	0.903	0.475

6 Results

This section describes source signal, transfer, and distance-to-loss association.

6.1 Representation and Source Signal

Table 2 and Fig. 1 test whether frozen SQL representations encode corpus membership before any outcome target is predicted. Under raw SQL, schema identity is nearly recoverable from every representation: logistic-regression macro F1 is 0.999 for TF–IDF/SVD, 0.995 for CodeBERT, and 1.000 for MiniLM. Cohort recovery is similarly high (0.980–0.987), while recovery of the finer nine-way source group remains high but lower (0.892–0.903). Thus, raw query strings expose a strong signal linked to the schema and collection source.

Masking identifiers and literals sharply reduces this easy source signal. Schema-recovery macro F1 falls to 0.492–0.511, a decrease of at least 0.485 across representations. Source-group recovery falls to 0.464–0.499, while cohort recovery remains between 0.647 and 0.666. Masking therefore removes much of the directly recoverable schema vocabulary, but it does not erase the structure associated with cohort or source batch. This residual signal is consistent with the data setting: query form and assignment/source properties vary across corpus additions in addition to schema naming.

The CodeBERT results also separate recoverability from a claim of globally compact grouping. For schema labels, its cosine silhouette score is only 0.036 with raw SQL and -0.025 after masking, while its linear schema-recovery macro F1 is 0.995 and 0.511, respectively. At $k = 10$, CodeBERT schema-neighbor purity decreases from 0.729 to 0.395 against a 0.158 class-frequency baseline; source-group purity decreases from 0.649 to 0.322 against a 0.138 baseline. Hence, the source-associated signal is available to local neighborhoods and linear probes even where a global cluster claim is not supported. Fig. 2 shows the corresponding cohort-colored CodeBERT UMAP projections for raw and masked SQL. These projections support visual comparison, while the result statement is based on the high-dimensional tests and recovery probes.

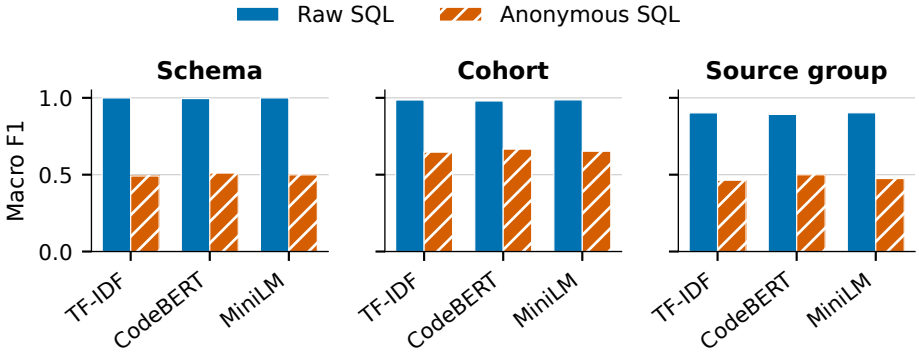

Fig. 1. Source-label recoverability after masking identifiers and literals. Bars report five-fold logistic-regression macro F1 from frozen representations; hatching identifies anonymous SQL for grayscale reading. Exact values appear in Table 2. Quantitative evidence is carried by the recovery metrics, not by a two-dimensional projection.

Table 3. Primary leave-one-cohort-out outcome prediction. Values are means over the three held-out cohorts. Bold marks the better input within each model. The optimized multitask CNN is a published configuration. Remark F1 is rare-label-sensitive; under anonymous SQL, supported-class F1 is 0.635 for CodeBERT and 0.626 for the CNN.

Model	Input	Correct. BA ↑	Remark F1 ↑	Grade MAE ↓
CodeBERT probe	Raw	0.599	0.372	21.994
CodeBERT probe	Anon.	**0.662**	**0.375**	**19.484**
Optimized multitask CNN	Raw	0.564	0.385	21.316
Optimized multitask CNN	Anon.	**0.660**	**0.481**	**19.966**

6.2 Cross-Cohort Outcome-Prediction Transfer

Table 3 presents the primary leave-one-cohort-out (LOO) outcome-prediction results. For the optimized multitask CNN, anonymous SQL improves all three mean LOO endpoints relative to raw SQL: correctness balanced accuracy increases from 0.564 to 0.660, all-class remark macro F1 increases from 0.385 to 0.481, and numeric-grade MAE decreases from 21.316 to 19.966. These gains are specific to held-out transfer rather than a general advantage of masking. Under within-cohort evaluation, the same CNN performs better with raw SQL than with anonymous SQL for correctness (0.741 versus 0.686), remarks (0.608 versus 0.596), and grade prediction (MAE 15.507 versus 20.566). Explicit schema-bearing tokens can therefore aid in-source prediction while harming transfer across cohort additions. Fig. 3 summarizes this contrast for the optimized multitask CNN, showing the raw-versus-anonymous change under within-cohort and LOO evaluation for all three endpoints.

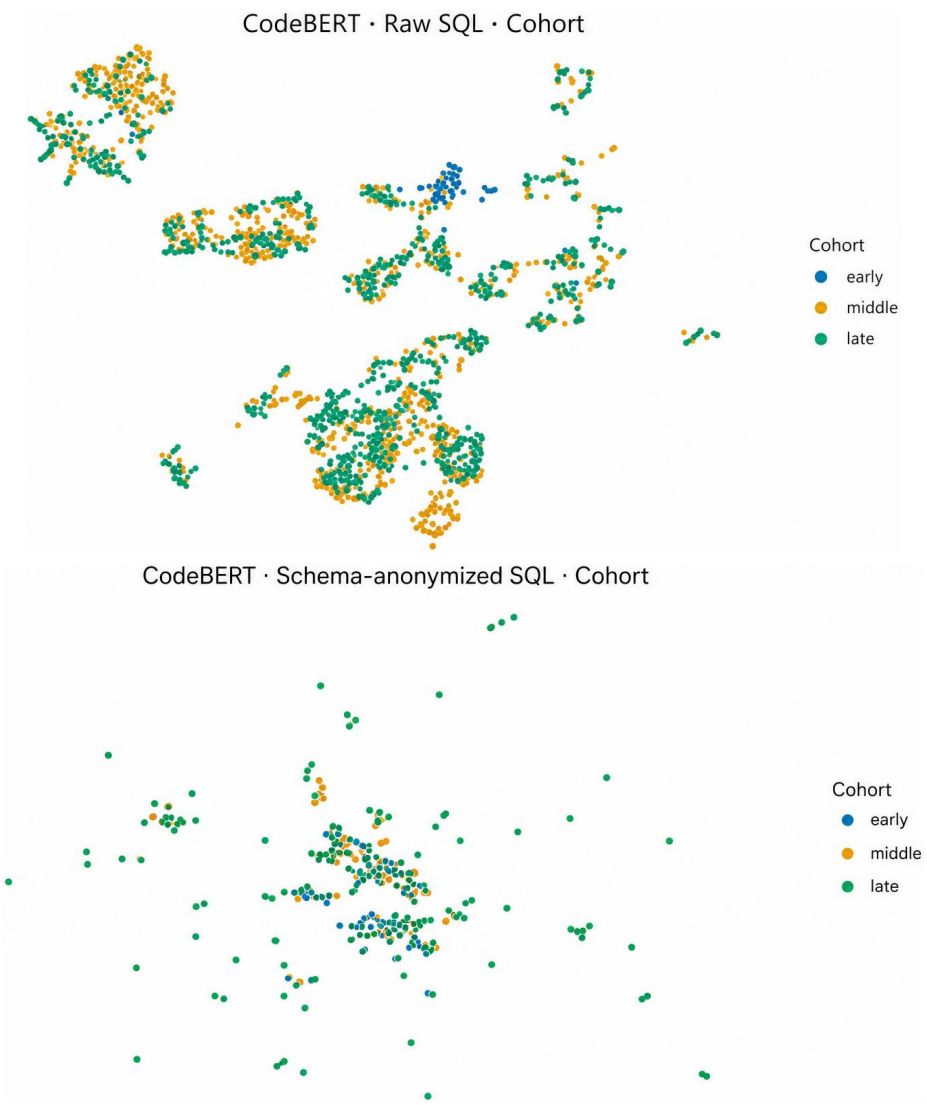

Fig. 2. CodeBERT UMAP projections of SQL submissions colored by cohort for raw SQL (top) and identifier-and-literal-masked SQL (bottom). Each point represents one submission; colors identify the early, middle, and late cohorts. The panels provide a visual comparison of cohort-associated geometry before and after masking.

Under anonymous LOO evaluation, CodeBERT and the optimized multitask CNN are effectively tied on correctness prediction (0.662 versus 0.660 balanced accuracy) and close on numeric-grade prediction (19.484 versus 19.966 MAE). The CNN has higher reported all-class remark macro F1 (0.481 versus 0.375). That comparison requires qualification: the remark labels are severely imbalanced, and

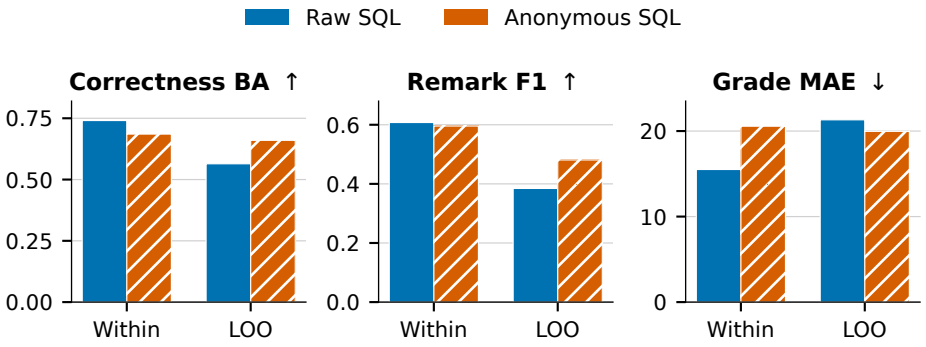

Fig. 3. Performance of the optimized multitask CNN within cohort and under leave-one-cohort-out (LOO) transfer. Anonymous SQL improves all three mean LOO outputs, while raw SQL is stronger within cohort. Lower values are better only for numeric-grade MAE. The model is a published configuration rather than an independent confirmation result.

the supported-class diagnostic over *Correct* and *Partially Correct* is nearly tied, with CodeBERT at 0.635 and the CNN at 0.626. Further, the CNN's anonymous all-class remark F1 varies from 0.202 on the held-out early cohort to 0.773 on the held-out late cohort. The result supports a higher reported mean remark score for the optimized multitask CNN, not a uniform advantage across cohorts or remark classes.

A bounded repeated-`tv` check using the CodeBERT probe provides a same-schema source-batch control. Although every comparison uses the same schema label, anonymous SQL retains source-batch recoverability (macro F1 0.637) and positive mean transfer losses remain across six directed source-batch comparisons: 0.169 in correctness balanced accuracy and 10.584 in grade MAE. In this fixed-schema control, masking does not reduce mean transfer loss relative to raw SQL. The primary cohort-level LOO improvement should therefore be read as an empirical transfer benefit in that setting, not as evidence that masking removes all source-associated shift.

6.3 Representation Distance and Transfer Loss

Representation distance is most informative for numeric-grade prediction. Table 4 reports the association between distance in each frozen representation space and transfer loss of the CodeBERT frozen outcome probe. Using CodeBERT distances, numeric-grade MAE loss increases with cross-domain distance under both raw SQL ($\rho_S = 0.740$, $p = 0.0228$, $n = 9$) and anonymous SQL ($\rho_S = 0.783$, $p = 0.0125$, $n = 9$). The same positive association holds when distance is measured in TF–IDF/SVD or MiniLM space; all six grade-distance correlations in Table 4 are positive with $p < 0.05$.

Table 4. Spearman association between representation distance and numeric-grade MAE transfer loss for the CodeBERT frozen outcome probe. Each cell uses nine cross-domain comparisons.

	Raw SQL		Anonymous SQL	
Distance space	ρ_S	p	ρ_S	p
TF-IDF/SVD	0.767	0.0159	0.899	0.0010
CodeBERT	0.740	0.0228	0.783	0.0125
MiniLM	0.833	0.0053	0.773	0.0145

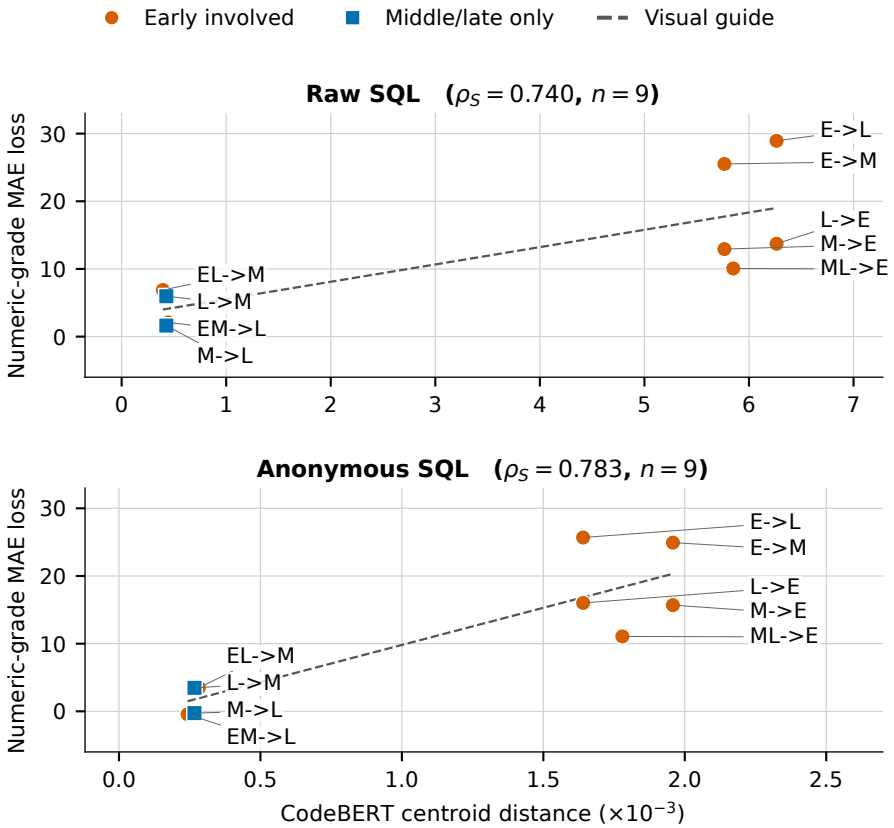

Fig. 4. Numeric-grade MAE transfer loss against CodeBERT centroid distance for the CodeBERT frozen outcome probe. Each panel contains nine cross-domain comparisons; point labels identify directed train-to-test comparisons. Comparisons involving the early cohort account for much of the far-distance range. Lines are visual guides; inference is based on the rank statistics and Table 4.

The association is endpoint-specific rather than a general claim about every outcome. For CodeBERT distance, correctness transfer loss has weaker and

uncertain rank associations (raw $\rho_S = 0.168$, $p = 0.6656$; anonymous $\rho_S = 0.467$, $p = 0.2054$), while remark-label loss does not show a consistent positive relation (raw $\rho_S = 0.227$, $p = 0.5571$; anonymous $\rho_S = -0.517$, $p = 0.1544$). This difference matters: representation distance signals risk for numeric-grade transfer in these data, but does not function as a single proxy for every prediction endpoint.

Fig. 4 also makes the range limitation visible. The comparisons involving the early cohort occupy the high-distance region and carry the largest numeric-grade losses. Since the early cohort is smaller and structurally different from the later groups, the association should be read as a useful warning signal within the studied source changes, not as a universal law of SQL outcome prediction. Distance correlations for the optimized multitask CNN use only three LOO targets per input variant and are therefore treated as exploratory rather than as primary evidence.

7 Discussion, Limits, and Responsible Use

The results establish a focused claim: SQL representations carry source-associated information, and this information matters for held-out prediction of numeric grade and, in the primary LOO setting, for the optimized multitask CNN's three reported outcomes. Raw SQL makes schema and source identity highly recoverable; masking removes much of that direct lexical signal. Yet the fixed-`tv` control shows that batch-associated transfer loss remains when the schema label is held constant. The findings therefore concern *schema/source-associated* shift, not schema names as a sole causal mechanism.

Limits. First, the observed cohorts link schema, source batch, query form, and label distribution. The early cohort is especially important: it contains 225 submissions, one schema, and substantially more joins than the later cohorts, and it sets much of the high-distance range in Fig. 4. Second, the remark endpoint has only 31 *Non-Interpretable* and 2 *Cheating* cases. We retain the specified all-class endpoint but qualify it with the supported-class diagnostic.

Responsible interpretation. Cross-source evaluation is necessary when a predictive model may face distributions different from those used for fitting, and model-selection choices must be made visible in such comparisons [4]. In an educational setting, responsible interpretation requires clarity about the records used and the meaning and limits of the model predictions [8]. Consequently, the present results support transfer testing before predictions are interpreted on newly collected schemas or source batches. They do not support using numeric-grade predictions as final instructional decisions without local validation and human oversight.

8 Conclusion

This study examines cross-cohort generalization for SQL outcome prediction under schema/source-associated distribution shift. Across frozen representations, source-

associated information remains recoverable from SQL text; masking identifiers and literals reduces direct schema and source recoverability without eliminating cohort-associated structure. In leave-one-cohort-out testing, the optimized multitask CNN benefits from masking across its three reported outcome endpoints, while the distance analysis identifies a positive association between representation distance and numeric-grade transfer loss across the tested representation spaces. Together, these findings show that held-out transfer behavior can differ from within-cohort performance.

The study does not attribute transfer degradation to schema names alone: in this corpus, cohort, schema, source batch, query form, and label distribution are linked. The evidence instead supports cross-cohort evaluation as a necessary test for SQL outcome-prediction models intended for newly collected data. Future work should separate these linked factors through confirmatory protocols and additional source groups; practical use of numeric-grade predictions should remain subject to source-aware validation and local human oversight.

Acknowledgments. Part of this work was funded by the National Science Foundation under grant CNS-2136961. The authors thank Dr. Gissella Bejarano Nicho for facilitating access to high-performance computing equipment. The authors thank the Rivas.AI Lab (`https://lab.rivas.ai`) for the support and helpful feedback throughout this project.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Chandra, B., Banerjee, A., Hazra, U., et al.: Automated grading of SQL queries. pp. 1630–1633. IEEE (2019). `https://doi.org/10.1109/ICDE.2019.00159`
2. Chandra, B., Joseph, M., Radhakrishnan, B., et al.: Partial marking for automated grading of SQL queries. Proceedings of the VLDB Endowment **9**(13), 1541–1544 (2016). `https://doi.org/10.14778/3007263.3007304`
3. Feng, Z., Guo, D., Tang, D., et al.: CodeBERT: A pre-trained model for programming and natural languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 1536–1547. Association for Computational Linguistics (2020). `https://doi.org/10.18653/v1/2020.findings-emnlp.139`
4. Gulrajani, I., Lopez-Paz, D.: In search of lost domain generalization (2020). `https://doi.org/10.48550/arXiv.2007.01434`
5. He, Y., Zhao, P., Wang, X., et al.: VeriEQL: Bounded equivalence verification for complex SQL queries with integrity constraints. Proceedings of the ACM on Programming Languages **8**(OOPSLA1), 1071–1099 (2024). `https://doi.org/10.1145/3649849`
6. Kleiner, C., Tebbe, C., Heine, F.: Automated grading and tutoring of SQL statements to improve student learning. In: Koli Calling '13: Proceedings of the 13th Koli Calling International Conference on Computing Education Research. pp. 161–168. Association for Computing Machinery (2013). `https://doi.org/10.1145/2526968.2526986`
7. Koh, P.W., Sagawa, S., Marklund, H., et al.: WILDS: A benchmark of in-the-wild distribution shifts (2020). `https://doi.org/10.48550/arXiv.2012.07421`

8. Nguyen, A., Ngo, H.N., Hong, Y., et al.: Ethical principles for artificial intelligence in education. Education and Information Technologies **28**(4), 4221–4241 (2023). https://doi.org/10.1007/s10639-022-11316-w
9. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. pp. 3980–3990. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1410
10. Rivas, P., Schwartz, D.R.: Modeling SQL statement correctness with attention-based convolutional neural networks. In: 2021 International Conference on Computational Science and Computational Intelligence (CSCI). pp. 64–71. IEEE (2021). https://doi.org/10.1109/CSCI54926.2021.00086
11. Rivas, P., Schwartz, D.R.: Designing multi-objective cnn architectures for sql query modeling with evolution strategies. In: Proc. of the 27th International Conference on Artificial Intelligence (ICAI'25), CSCE'25. Las Vegas, NV, USA (Jul 2025), https://rivas.ai/pdfs/rivas2025designing.pdf, july 21–24, 2025
12. Rivas, P., Schwartz, D.R.: Explainable AI for SQL grading: A practical approach with multi-task CNNs. pp. 57–73. Springer Nature Switzerland (2025). https://doi.org/10.1007/978-3-031-86623-4_5
13. Rivas, P., Schwartz, D.R., Quevedo, E.: BERT goes to SQL school: Improving automatic grading of SQL statements. pp. 83–90. IEEE (2023). https://doi.org/10.1109/CSCE60160.2023.00019
14. Schwartz, D.R., Rivas, P.: An automated SQL query grading system using an attention-based convolutional neural network. arXiv preprint (2024). https://doi.org/10.48550/arXiv.2406.15936
15. Sooksatra, K., Khanal, B., Rivas, P., Schwartz, D.R.: Attribution scores of BERT-based SQL-query automatic grading for explainability. pp. 213–220. IEEE (2023). https://doi.org/10.1109/CSCI62032.2023.00039
16. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers (2020). https://doi.org/10.48550/arXiv.2002.10957
17. Yu, T., Zhang, R., Yang, K., et al.: 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. pp. 3911–3921. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/D18-1425