

CCQ: A Multi-State Child Care Quality Dataset to Support AI for Children’s Health Research

VICTOR LI, Emory University, USA
 YUZHANG XIE, Emory University, USA
 ZIWEI DONG, Emory University, USA
 QINGYANG ZHU, Emory University, USA
 WENJING MA, Emory University, USA
 CARL YANG, Emory University, USA
 JINBING BAI, Emory University, USA
 HUIWEN XU, Emory University, USA
 JIAYING LU*, Emory University, USA

High-quality child care in early life is a critical determinant of children’s growth and development. Research on child care quality has been constrained by fragmented, non-research-friendly, and privacy-bound datasets. We present *CCQ* (Child Care Quality), a large-scale, de-identified dataset for applied data science research at the intersection of AI and early childhood health. *CCQ* integrates 59,372 child care provider records across 12 U.S. states, covering diverse provider types as well as data schemas. To ensure research utility while protecting privacy, we implement an automated, LLM-based curation pipeline that anonymizes, cleans, and standardizes raw state records into two complementary releases: a cleaned textual release and a fully preprocessed tabular release. We also benchmark traditional machine learning models, tabular foundation models, and language models on quality rating prediction and important features analytics. Within a state, tabular classifiers on the preprocessed tables perform best. Across states, zero-shot transfer is near chance, but modest target-state supervision recovers most of the within-state performance, and pretraining on other states benefits finetuned language models. We release both datasets with all code to accelerate AI-driven research on child care quality and ultimately improve children’s health and development.

CCS Concepts: • **Applied computing** → **Health informatics**; • **Computing methodologies** → *Machine learning*; • **Information systems** → *Data cleaning*.

Additional Key Words and Phrases: AI for children’s health, child care quality rating prediction, benchmark and dataset, LLM-based data curation

1 Introduction

Early-life environmental exposures, such as physical and social environments, affect children’s health and development [10]. Among these, non-parental child care environments constitute a majority of sustained exposure. More than eight hours in care covers most of a young child’s waking hours [3], and about 59% of U.S. children age five and younger have at least one weekly non-parental care arrangement [18]. Furthermore, approximately 15.7 million children in the U.S. under the age of five have routine child care arrangement [76], where children spend an average of 36 hours a week in these settings [86]. These early and prolonged experiences are linked to development well into adolescence: more hours of non-parental care before age four and a half predict more risk

*Corresponding author: Jiaying Lu, jiaying.lu@emory.edu

Authors’ Contact Information: Victor Li, Emory University, Atlanta, GA, USA, victor.li@emory.edu; Yuzhang Xie, Emory University, Atlanta, GA, USA, yuzhang.xie@emory.edu; Ziwei Dong, Emory University, Atlanta, GA, USA, ziwei.dong@alumni.emory.edu; Qingyang Zhu, Emory University, Atlanta, GA, USA, qingyang.zhu@emory.edu; Wenjing Ma, Emory University, Atlanta, GA, USA, wenjing.ma@emory.edu; Carl Yang, Emory University, Atlanta, GA, USA, j.carlyang@emory.edu; Jinbing Bai, Emory University, Atlanta, GA, USA, jinbing.bai@emory.edu; Huiwen Xu, Emory University, Atlanta, GA, USA, huiwen.xu@emory.edu; Jiaying Lu, Emory University, Atlanta, GA, USA, jiaying.lu@emory.edu.

taking and impulsivity at age 15 [77]. Higher-quality care is also associated with stronger cognitive and language skills by age three [57] and better cognitive and social development through second grade [64]. It also correlates with higher academic achievement and less externalizing behavior at age 15 [77, 78]. As a common part of family life, understanding the quality of child care is crucial for promoting children’s health and development.

In order for families to act on these conclusions, accessible and relevant data on child care quality is needed. Provider-level quality ratings published through state-sponsored quality rating and improvement systems (QRIS) are designed to close this gap and inform parents’ decisions [35, 49]. Yet, parents do not use all the information available to them when judging quality [52]. Specifically, in a study of low-income families, parents’ evaluation of their child’s program bore little correlation to these external measures [9]. However, since QRIS participation is voluntary in most states and coverage is often low, the providers a family is considering may simply have no rating [9, 35]. This coverage gap can be large: among the QRIS that reported participation in 2017, the median system rated only 60% of licensed centers and 25% of family child care programs, and New Jersey rated only 1% of its centers [56]. Beyond supporting parents, states use QRIS to help providers improve the quality of their care, and many pay higher subsidy rates to programs with higher ratings [35, 44]. More complete coverage would make these efforts more effective as well. It would also advance research on how care quality relates to children’s health and developmental outcomes, such as the NICHD Study of Early Child Care and Youth Development, which followed 1,364 children from birth and directly observed the child care settings they attended [78], as well as research on the early-life environmental exposures examined by programs such as NIH ECHO [10].

Machine learning offers a way to close the coverage gap by estimating quality for providers that have not been rated. In early care and education, it has been implemented to assess preschool teaching quality [43] and to predict the accreditation levels of early childhood institutions [59], but each study used a small sample from a single rating system. A model fit to one rating system risks overfitting to that system’s features and rating rules, and may not generalize to providers outside it, which is exactly the case where parents have no rating to rely on. Additionally, to support models that generalize across rating systems, the data must capture the many interconnected factors that shape quality. These include the physical environment (e.g., equipment, toys, facilities), food service (e.g., meals offered), policies and procedures (e.g., operational rules, parental access), program activities (e.g., curriculum, planned indoor/outdoor activities), and staffing (e.g., staff-to-child ratios) [3].

To our knowledge, no publicly available dataset features these factors with provider-level quality ratings across multiple rating systems. We attribute this gap to the challenges of collecting and preprocessing such heterogeneous data. Child care quality ratings (QR) are distributed by state across unstandardized QRIS, ranging from government-owned licensing databases and third-party websites to self-reported program descriptions [21, 74]. Each system stores and publishes its findings in its own schema [74], which also contain missing values, errors, and outdated information [51]. In addition to these data limitations, preparing real-world children’s health data often relies on costly expert annotation [36] and ad hoc, domain-specific preprocessing [46] with teams of domain experts and data scientists [34]. Existing approaches for automated data curation pipelines also rely heavily on domain-specific knowledge [55], with hard-coded rules that make pipelines impractical to maintain, as well as, failing to generalize across diverse datasets [55].

To fill this gap, we construct *CCQ* (Child Care Quality), a curated dataset that combines provider-level quality ratings with provider features across twelve U.S. states. Because hand-written rules and expert annotation are time-consuming and labor-intensive at this scale, we leverage large language models (LLMs) to automatically collect, anonymize, and curate the raw state-specific records into two *AI-ready* releases. They cover two families of models: a cleaned text release (*raw*) that preserves the textual features of source records for language-based deep learning models, and a fully preprocessed tabular release (*preprocessed*) that is strictly numerical and supports out-of-the-box analysis. Both releases are verified by hand for private information and data leakage, with the final values spot-checked against the source portals.

Using *CCQ*, we benchmark models on two tasks of QR prediction, mirroring how state agencies would actually deploy them. In within-state prediction, a state already has many rated providers and wants to assign ratings to its remaining unrated providers under its own rating system. In cross-state prediction, a target state has few or no rated providers and must borrow from models trained on other states, whose schemas and rating systems differ. This harder setting tests the generalization across rating systems that single-system studies could not examine. We benchmark a representative set of models, spanning classical machine learning baselines, tabular deep learning and foundation models, and pretrained language models. First, we evaluate their predictive accuracy using ordinal agreement and balanced accuracy. Second, we apply feature attribution over the preprocessed features to identify which provider attributes drive predictions. Third, we measure cross-state transferability under different amount of labeled data, from zero-shot to 80% of the target state’s providers. Lastly, we examine whether full transfer learning in cross-state prediction improves prediction accuracy over the within-state setting. Together, these results establish reference baselines for the two QR prediction tasks and show where current models break down under realistic, heterogeneous conditions.

Our contributions are summarized as follows:

- We construct *CCQ*, a comprehensive provider-level dataset that combines child care quality ratings with provider features across twelve U.S. states. To our knowledge, it is the first publicly available, multi-state, AI-ready dataset of real-world child care quality ratings, supporting future AI research on children’s health.
- We build an automated curation pipeline that leverages LLMs to efficiently collect, anonymize, and clean raw state-specific records into AI-ready releases. The pipeline reduces the time and labor that manual curation of multi-dimensional child care data typically requires. We verify every output by hand for private information and data leakage, and spot-check its values against the source portals.
- We benchmark 10 representative models, ranging from classical machine learning classifiers and tabular foundation models to finetuned language models, on within-state and cross-state QR prediction. We then use feature attribution to identify the provider attributes that drive predictions, and measure how cross-state transferability improves with more of target state labeled.

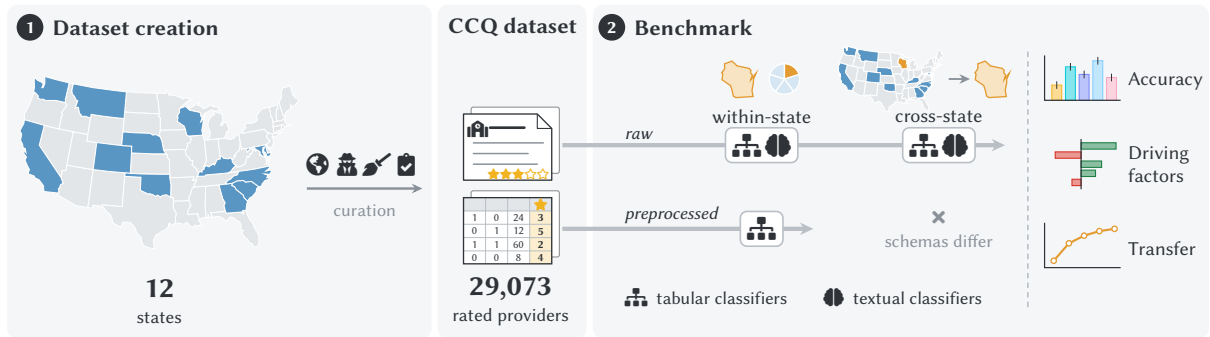


Fig. 1. Overview of the *CCQ* study. **Stage 1** curates the QRIS records of twelve states in four steps: collection, anonymization, cleaning, and verification (Figure 2). The *raw* release retains textual feature while the *preprocessed* release is strictly numerical. **Stage 2** benchmarks tabular classifiers and language models on two tasks. Within-state prediction uses five-fold cross-validation and cross-state prediction holds out one state (orange) and trains on the others (blue). Because the *preprocessed* data schemas differ across states, it is only used for within-state prediction. The trained models are analyzed for prediction accuracy, feature attribution, and cross-state transferability.

2 Related Work

2.1 Child Care Quality Datasets

Though several child care quality datasets [6, 8, 26, 75] are publicly available and could support quality prediction modeling, they are limited in their coverage and in the features they record compared to *CCQ*. The Australian Children’s Education & Care Quality Authority publishes the ratings of more than 17,000 early childhood services, all assessed under the National Quality Standard [6, 7]. Indonesia’s national accreditation agency publishes the accreditation scores for more than 174,000 early childhood education institutions [8, 66]. Beyond the ratings themselves, however, both record little about the factors that shape quality. The Australian dataset records each service’s provider, location, care types, number of approved places, and operating hours [6]. The Indonesian dataset generally records only each institution’s accreditation score [8]. On the other hand, the Canadian province of Alberta releases quarterly records of 2,720 licensed child care providers with their location, type, capacity, and inspection results, but assigns them no quality ratings [26]. The U.S. Head Start Family and Child Experiences Survey [75] also records observed quality scores (CLASS and ECERS-R) for 669 classrooms in 347 Head Start centers [2], but assigns no quality rating. Lastly, the Quality Compendium [74] catalogs QRIS programs across U.S. states, but it is curated at the system level rather than the provider level. To our knowledge, no unified, multi-state, provider-level QRIS dataset is publicly available. *CCQ* fills this gap by assembling provider-level records and quality ratings across twelve U.S. states into a single, consistently curated dataset with features that span the factors shaping care quality [3], although each state’s schema differs.

2.2 Child Care Quality Prediction

Predictive modeling of child care and early childhood education quality has not kept up with modern machine learning. Existing studies rely mostly on regression models and classical classifiers [2, 14, 19, 25, 29, 43, 59, 71]. To our knowledge, no study evaluates tree ensembles, tabular foundation models, or language models. Char et al. [14] and Harrison et al. [29] analyzed the Australian national register of quality ratings [6] using multinomial logistic regression. Siraj et al. [71] found correlations between these ratings and observed classroom quality using hierarchical multiple regression. Aikens et al. [2] analyzed Head Start classroom quality in the FACES dataset [75] using linear and logistic regression. Noripansyah et al. [59] compared support vector classifiers (SVC), K-Nearest Neighbors (KNN), and neural network approaches for predicting the accreditation scores of 225 early childhood education providers in South Kalimantan, Indonesia [8, 66]. Other studies rely on samples that the researchers collected themselves. Goelman et al. [25] developed a path-analytic prediction model of center-based child care quality across 239 Canadian centers, while Doherty et al. [19] used hierarchical regression to predict the quality of 231 Canadian family child care providers. These studies all share two limitations that our benchmark addresses. First, each study fits and evaluates models using data from a single program with a limited number of labeled providers, so the reported performance does not reflect how models behave on large, heterogeneous datasets. Second, no study examines whether a model trained on one rating system transfers to another. In contrast, we evaluate a more comprehensive set of conventional machine learning methods, tabular foundation models, and deep learning models on much larger datasets. We also explicitly study the heterogeneous cross-state transfer setting, in which models are trained on several states’ data despite differing schemas, which reflects the fragmentation of real-world child care rating programs.

3 Dataset Creation

3.1 Data Sources

The *CCQ* (Child Care Quality) dataset integrates publicly available child care data across 12 U.S. states: California (CA), Colorado (CO), Georgia (GA), Kentucky (KY), Maryland (MD), Montana (MT), North Carolina (NC), Nebraska

(NE), Oklahoma (OK), South Carolina (SC), Washington (WA), and Wisconsin (WI). In total, we assemble 59,372 records, originally released by state agencies or their designated partners. Each state agency operates its own QRIS, which differ in design but share the purpose of helping families make informed decisions [35, 49]. Table 1 lists the source agency, QRIS, QR scale, and provider counts for each state. With the exception of Georgia, which uses a 1–3 scale, all states publish a 1–5 QR scale. Overall, only 29,073 providers (49.0%) in the dataset possess an assigned QR score, highlighting the need for predictive models that can estimate ratings for the remaining facilities.

Table 1. Data sources overview. #Feature includes the provider ID and QR score.

State	Data Source	#Provider	#Rated	QRIS	QR Scale	#Feature
CA	CA Child Care Resource & Referral Network [54]	10,715	847	Quality Counts California	1–5	142
CO	CO Dept of Early Childhood [17]	4,508	3,423	Colorado Shines Rating	1–5	64
GA	GA Dept of Early Care and Learning [24]	8,192	2,897	Quality Rated	1–3	203
KY	KY Cabinet for Health and Family Services [37]	1,929	1,896	Kentucky All STARS	1–5	61
MD	MD State Dept of Education [50]	7,106	5,005	Maryland EXCELS	1–5	32
MT	MT Dept of Public Health & Human Services [53]	213	181	Best Beginnings STARS to Quality	1–5	12
NC	NC Div. of Child Development & Early Education [60]	3,264	3,050	NC Star Rated License	1–5	30
NE	NE Dept of Education [58]	3,274	1,079	Step Up to Quality	1–5	38
OK	OK Dept of Human Services [61]	2,557	2,554	Reaching for the Stars	1–5	179
SC	SC Dept of Social Services [72]	2,400	1,223	ABC Quality	1–5	10
WA	WA Dept of Children, Youth, & Families [82]	10,399	3,168	Early Achievers	1–5	169
WI	WI Dept of Children and Families [84]	4,815	3,750	YoungStar	1–5	164
Total	–	59,372	29,073	–	–	–

3.2 Construction Pipeline

Since each state administers and publishes its own QRIS [74], there is no federal registry, common schema, or common format for the data. The practical consequence is that data from each state require substantial state-specific collection and preprocessing. We believe this lack of standardization is a key reason why no national-scale dataset of provider-level quality ratings currently exists. Our construction pipeline uses Georgia as a human-implemented reference that the remaining states are adapted from. It contains four main steps: collection, anonymization, cleaning, and verification. We chose Georgia as the reference because it has the most published fields of the 12 states (Table 1), which all publish its data on easy accessible portals.

3.2.1 Collection. For every state, we collect a single table where rows represent providers and columns represent provider attributes. The first two columns are always the provider identifier (ID) and the QR score. In the hand-written collection pipeline for Georgia, providers are enumerated through the state agency’s search portal, fetching all available information. We then prompted a coding agent (Claude [4]) to adapt that reference to each remaining state, following prior work that uses LLMs to generate web scrapers [33] and LM agents to write code [87]. Each state’s collected columns were then verified by hand, using the column names and their unique values, confirming that every informative attribute available had been recorded.

3.2.2 Cleaning. To support different machine learning models, every state is cleaned to produce two separate releases of the data: *raw* and *preprocessed*. The *raw* release applies minimal preprocessing and preserves the tabular text of the source records, so that each provider record can be directly fed as natural language into text-based models. On the other hand, the *preprocessed* release applies a complete preprocessing pipeline for out-of-the-box analysis, creating a strictly numerical dataset. Both releases remove sensitive private columns as well as data-leakage columns. The private columns are those that identify a provider or a person rather than

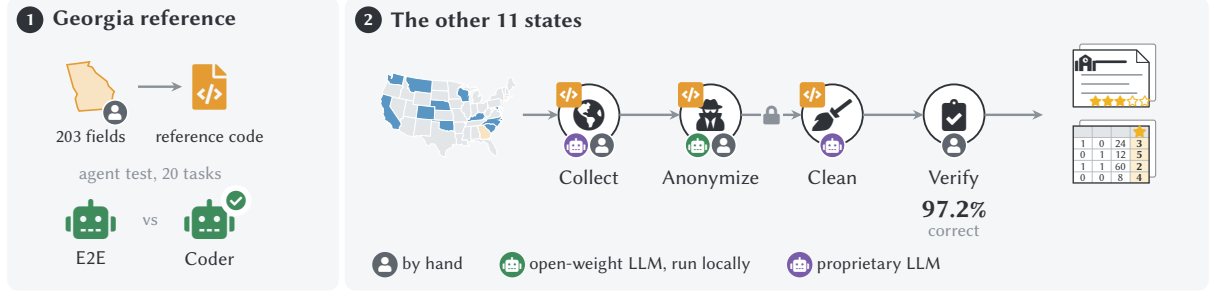


Fig. 2. The CCQ construction pipeline. **(1)** The Georgia reference is implemented by hand. Its human-curated data are the ground truth for testing two agents on twenty curation tasks: a per-sample agent (E2E) and a code-writing agent (Coder), which is adopted. **(2)** The other eleven states are collected, anonymized, and cleaned with the help of LLMs. Human verification at each step is shown with icons indicating who performed each step. The orange code icon indicate when Georgia’s code is used as and in-context example. Lastly, two annotators spot-checked the curated data against the live portals, finding 97.2% of the checked values correct.

describe the program, such as provider names, staff names, addresses, emails, phone numbers, and websites. Removing them is a de-identification step in the sense used by NIST, which covers establishments as well as individuals [23]. This also includes free-text narratives some states publish, such as inspection, complaint, and violation reports, which carry private information about the provider. Coarse geography, county, and region are retained. The data-leakage columns are those that disclose the rating rather than predict it, such as award dates or different formats of rating scores. Missing values are left empty in the *raw* release, but encoded explicitly in the *preprocessed* release. Lastly, unrated providers are also removed.

As with collection, Georgia serves as the reference, and its preprocessing was carried out under strict expert supervision. Against that hand-curated ground truth we then measured two zero-shot agents on twenty curation tasks for anonymization and cleaning: a per-sample agent (*Qwen3-8B* [67], “E2E”), prompted once per record, and a code-writing agent (*Qwen3-Coder-30B* [68], “Coder”), prompted to write one reusable cleaning function per field with the ground-truth code withheld. Table 2 reports the results, grouping the twenty tasks into five types. Each row averages over the tasks of one type, and the #Tasks column gives their number.

Although the coding agent does not achieve perfect performance, we adopt it to conduct anonymization and generate cleaning scripts. Its accuracy is comparable to the per-sample agent, except on rare edge cases within the data, at a fraction of the cost. To remedy these edge cases, we include the Georgia pipeline as an in-context example in every anonymization and cleaning prompt. The in-context example is always code: the reference implementation’s anonymization and cleaning functions are supplied to the agent, and no provider record from any state is ever placed in the prompt. We also use a proprietary LLM (Claude [4]), whose code-generation ability lets us produce an entire state’s pipeline in a single pass rather than one cleaning function at a time. Importantly, no private information was exposed to the proprietary model as anonymization was performed locally with open-source models beforehand, and every column flagged as sensitive was removed by hand before any cleaning script was generated.

The following boxes give one prompt from each agent. The anonymization prompt asks the per-sample agent for a single privacy decision about one column, given the column name and a sample of that column’s values. The cleaning prompt asks the code-writing agent for one reusable function, given the field name and the field’s unique values, with the requirements stated as numbered constraints. The remaining prompts follow these two shapes and vary only in the operation requested and the output columns written.

Table 2. LLM-based curation results on the reference state Georgia. Token counts are per provider record, so the code-writing agent’s one-time cost is amortized over the 8,192 records in the Georgia reference implementation.

Curation Task	Model	#Tasks	Acc%	In Tok.	Out Tok.
<i>Anonymization</i>					
Column removal	E2E	1	97.12	71.0	1.0
Column removal	Coder	1	99.28	103.1	1.0
De-identification	E2E	3	35.2 ± 42.8	107.2 ± 22.6	4.4 ± 2.5
De-identification	Coder	3	100.0 ± 0.0	0.03 ± 0.00	0.05 ± 0.01
<i>Cleaning</i>					
Text → numeric	E2E	5	93.3 ± 13.4	165.3 ± 18.4	2.7 ± 1.2
Text → numeric	Coder	5	99.98 ± 0.02	0.14 ± 0.12	0.03 ± 0.01
Text → multi-label	E2E	8	97.2 ± 3.8	390.1 ± 103.3	22.2 ± 9.0
Text → multi-label	Coder	8	88.9 ± 22.1	0.53 ± 0.96	0.08 ± 0.03
Compound field	E2E	3	95.8 ± 3.0	511.3 ± 345.4	13.9 ± 12.0
Compound field	Coder	3	90.2 ± 11.9	1.92 ± 1.29	0.08 ± 0.03

Prompt of the E2E Agent for Anonymization: Column Removal

You are an AI researcher studying child care quality data. In order to support open science, you need to determine which fields in the original data need to be anonymized to protect privacy. Given the variable name “{PLACEHOLDER_NAME}” and sample associated values “{PLACEHOLDER_VAL1, PLACEHOLDER_VAL2, ...}”, respond “Yes” if it contains protected privacy information, “No” if not.

Prompt of the Coder Agent for Cleaning: Compound Field Split

You are an AI research assistant creating a child care quality dataset. In order to allow future researchers to easily use this data, some of the fields in the raw data require cleaning. Given the field name ‘{PLACEHOLDER_FIELD_NAME}’ and unique values ‘{PLACEHOLDER_UNIQUE_VALUES}’, write a min-max split function with the following requirements.

1. The function should have two inputs: ‘csv_file_path’ and ‘column_name’.
2. The function should first read the ‘csv_file_path’ into a pandas dataframe and then extract the ‘column_name’ column from it.
3. The function should then clean the column by splitting the input into two separate values: min and max. Then, remove any leading non-numerical prefix (symbols/words/spaces) from each value. If the input is nan, missing, unknown (i.e., it is np.nan, ‘nan’, ‘NaN’, ‘None’, ‘null’, ‘unknown’, an empty string, or whitespace-only), set the cleaned value to be None.
4. After splitting the input into min and max, replace the original column with two new columns titled ‘predicted_min’ and ‘predicted_max’ and save the new dataframe to the original ‘csv_file_path’.

Now, output the ONLY the code for the function. Do not output any extra text or explanations.

3.2.3 Verification. Human verification effort was allocated for efficiency under time constraints. First, the Georgia reference implementation code was fully verified for collection, anonymization, and cleaning. For the other states, only the anonymization step was fully verified by hand. Both critical failure modes, privacy and data leakage, were checked to ensure that no text or derived feature contains any private or quality-rating information. However, the generated cleaning scripts were not verified line by line for correctness. Instead, we spot check the correctness of the end-to-end output, from collection to cleaned data, through a human review against the live source portals.

To estimate end-to-end collection and cleaning error, two authors independently re-derived a small set of fields by hand from each sampled provider’s page on the state portal and compared them to the released values. We reviewed 10 of the 12 states. In each, we drew 20 rated providers at random, stratified by quality rating in proportion to the state’s rating distribution. Georgia is excluded because it is the reference implementation and Montana because its quality ratings originate from a standardized database. For each of the 200 providers (0.69% of the 29,073 rated providers), we checked the four most common and critical fields: quality rating, provider type, licensed capacity, and ages served. Since not all of these fields exist in every state, we exclude each state-field pair where the field is absent. This leaves 740 checked values per annotator.

Table 3 reports the results per state, and its last row (ALL) pools them over the ten states. The two annotators judged 96.8% and 97.6% of the released values correct, or 1,438 of the 1,480 checks combined (97.2%). They agreed with each other on 730 of the 740 values. Pooled over states, Cohen’s kappa [16] is 98.1 for quality rating, 99.4 for provider type, 95.9 for licensed capacity, and 99.4 for ages served, all within the range of almost perfect agreement [41]. Agreement is lowest for licensed capacity in Kentucky (68.0). Five of the ten disagreements are Kentucky capacities that one annotator flagged and the other accepted. Furthermore, there are 16 values that both annotators flagged (2.2%). Four are provider types in Colorado, where both annotators found 4 of the 20 values wrong. Five are quality ratings, one each in Kentucky, Nebraska, and Wisconsin and two in Washington. The remaining seven are four licensed capacities and three ages-served values. We attribute these mismatches to recently updated records that changed between time of collection and verification. However, we did not explicitly determine the cause of each mismatch.

Table 3. Manual data verification results per state (20 providers per state). H1 and H2 give the percentage of released values each annotator judged correct against the live portal. κ is Cohen’s kappa ($\times 100$) between their correct/incorrect judgments. The ALL row pools the states that publish each field. “NA” indicates that the state does not publish the field.

State	QR score			Provider type			Licensed capacity			Ages served		
	H1	H2	κ	H1	H2	κ	H1	H2	κ	H1	H2	κ
CA	100.0	100.0	100.0	NA	NA	NA	100.0	100.0	100.0	95.0	95.0	100.0
CO	100.0	100.0	100.0	80.0	80.0	91.3	90.0	95.0	94.6	100.0	100.0	100.0
KY	90.0	95.0	91.6	100.0	100.0	100.0	70.0	95.0	68.0	NA	NA	NA
MD	100.0	100.0	100.0	100.0	100.0	100.0	95.0	95.0	100.0	100.0	100.0	100.0
NC	100.0	95.0	92.4	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
NE	95.0	95.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
OK	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SC	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	NA	NA	NA
WA	90.0	85.0	89.1	100.0	100.0	100.0	95.0	100.0	93.8	100.0	100.0	100.0
WI	95.0	95.0	100.0	100.0	100.0	100.0	95.0	95.0	100.0	90.0	90.0	94.3
ALL	97.0	96.5	98.1	97.8	97.8	99.36	94.5	98.0	95.87	98.1	98.1	99.35

3.3 Dataset Overview

CCQ contains 59,372 child care providers drawn from 12 states: four in the West (CA, CO, MT, WA), two in the Midwest (NE, WI), and six in the South (GA, KY, MD, NC, OK, SC). Records span learning centers, family child care homes, and other specialized provider types. Of the 59,372 providers, 29,073 (49.0%) carry a published quality rating and form the labeled portion of the dataset. The size of each state’s dataset differs greatly (Table 1), from 181 rated providers in Montana to 5,005 in Maryland. Additionally, feature schemas are defined state by state and unique, ranging from 10 attributes in South Carolina to 203 in Georgia, with a median of 62.5 attributes.

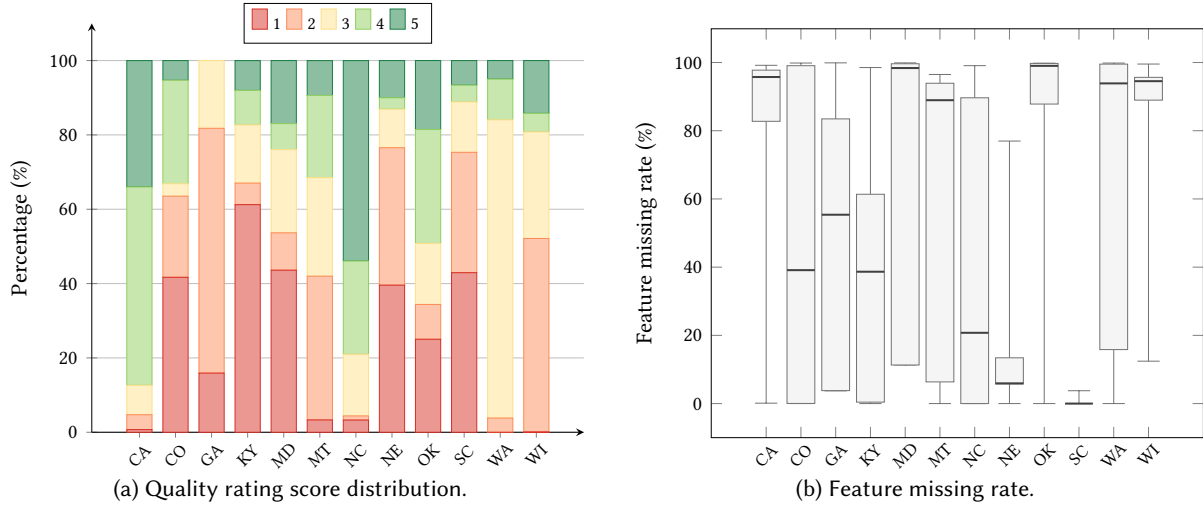


Fig. 3. Overview of the CCQ dataset. (a) Distribution of quality rating scores across rated providers in each state. (b) Feature missing rates for each state’s *raw* release, excluding the provider identifier and rating column. Boxes span the interquartile range and whiskers mark the 5th and 95th percentiles.

Rating coverage and rating distributions also vary widely across states. The rated share of a state’s providers runs from 7.9% in CA to 99.9% in OK (Table 1). Figure 3a shows the share of rated providers at each level of a state’s scale. Every state’s rating distribution except OK’s is skewed, but the direction of the skew is not shared. Five states concentrate at the bottom of their scale: KY and MD place 61.23% and 43.60% of providers at one star, and NE, SC, and CO place 76.55%, 75.31%, and 63.51% of providers in their bottom two levels. Two states concentrate at the top: CA assigns four- and five-star ratings to 87.36% of providers, and NC assigns five stars to 53.93%. Four states concentrate in the middle: WA places 80.24% of providers at exactly three stars, WI places 52.08% at two stars, MT places 65.19% at two or three stars, and GA places 65.86% at two stars within its three-level scale. Only OK resembles a balanced distribution. In addition, some classes are extremely small: WA, WI, MT, and CA report 0, 2, 6, and 6 providers, respectively, at the one-star level.

Furthermore, the feature missing rate differs across states. Figure 3b summarizes, for each state, the distribution across that state’s feature columns of the percentage of providers for which the column carries no value. The figure describes the *raw* release, in which an attribute a provider does not report is left empty rather than encoded (Section 3.2.2). On the lower end, SC is nearly complete: seven of its eight feature columns carry a value for every provider, and only capacity is ever missing (5.81%). NE (5.93%) and NC (20.75%) sit below a third of empty values for the median field, followed by KY (38.66%), CO (39.12%), and GA (55.37%). Lastly, MT (88.95%), WA (93.88%), WI (94.53%), CA (95.75%), MD (98.38%), and OK (99.02%) have nearly empty columns for their median field. MD shows the sharpest split of the twelve: only one of its 30 feature columns (program type) is fully populated, nine more are missing for at most 15% of providers, and 19 of the remaining 20 are empty for more than 86%.

3.4 Dataset Release

To promote reproducibility and facilitate future research, the CCQ dataset is made publicly available exclusively for non-commercial, academic research purposes. The dataset is hosted on the Hugging Face platform (<https://huggingface.co/datasets/GAIN-Lab/CCQ>) as a gated dataset. Access is granted upon request: prospective users

are required to sign a data use agreement, specify the intended purpose of use, and acknowledge the non-commercial restriction. This mechanism ensures that all usage complies with the CC-BY-NC-SA 4.0 license and the ethical guidelines governing the dataset. The code for data curation and for reproducing the benchmarking experiments is available at <https://github.com/Veeeeee7/CCQ>.

4 Benchmark Design

Using *CCQ*, we design a benchmark centered on the QR prediction task: given the curated record of a single child care provider, predict the quality rating that its state would assign. The benchmark contains two tasks that mirror how state agencies would deploy such models. In within-state prediction, a model is trained on a state’s own rated providers and predicts ratings for the rest. In cross-state prediction, a model for one state is trained on other states, whose schemas and rating systems differ.

4.1 Within-State QR Prediction

States use QRIS to evaluate the quality of child care providers and communicate it to families [35]. However, most QRIS do not rate every provider [56]. Participation is voluntary in most states [35], and ratings rely on in-person observations that are costly and time-consuming to collect [49]. Accurate within-state QR prediction can help state agencies close this coverage gap by assigning approximate ratings to unrated providers. Such models could also lower the cost of rating and extend ratings to more providers.

Formally, let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ denote the dataset of a single state, where $\mathbf{x}_i \in \mathcal{X}$ is the feature vector of child care provider i and $y_i \in \mathcal{Y}$ is its quality rating. The label space $\mathcal{Y} = \{1, \dots, K\}$ is ordinal. In the within-state setting, the training set $\mathcal{D}^{\text{train}}$ and the test set $\mathcal{D}^{\text{test}}$ are both drawn *i.i.d.* from the same distribution P over $\mathcal{X} \times \mathcal{Y}$. They therefore share the same feature schema $\mathcal{X}$ and label space $\mathcal{Y}$. The task is to learn a classifier $f : \mathcal{X} \rightarrow \mathcal{Y}$ from $\mathcal{D}^{\text{train}}$ that performs best on the held-out $\mathcal{D}^{\text{test}}$. Since there is no schema mismatch or distribution shift, this is a standard supervised ordinal classification problem.

The evaluation protocol should follow common practices in supervised machine learning. We conduct stratified five-fold cross-validation [39], with the fold assignments fixed once per state and shared by all compared methods. Since quality ratings are ordinal, we use ordinal-aware metrics such as quadratic weighted kappa (QWK) [16]. Balanced accuracy (BA) [13] is also reported as a complementary measure that accounts for label imbalance. This matters for *CCQ*, where some rating levels hold only a handful of providers (Section 3.3).

4.2 Cross-State QR Prediction

In practice, QRIS coverage varies widely across states and depends on local policy. Among the QRIS that reported participation in 2017, the share of licensed centers with a rating ranged from 1% in New Jersey to 100% in states whose QRIS places every licensed program at the first level [56]. High-coverage states could therefore lend their data to states whose rated subset is too small to train a within-state model. However, each QRIS uses its own data schema, variable definitions, and QR scale (e.g., 1–5 vs. 1–3). This creates a need for models that transfer knowledge across states despite limited and incompatible labeled data.

Formally, let $\mathcal{D}_s$ denote the pooled dataset of a set of source states, with feature space $\mathcal{X}_s$ and ordinal label space $\mathcal{Y}_s$. Let $\mathcal{D}_t$ denote the dataset of a target state t , with feature space $\mathcal{X}_t$ and label space $\mathcal{Y}_t$. In general, $\mathcal{X}_s \neq \mathcal{X}_t$ and $\mathcal{Y}_s \neq \mathcal{Y}_t$. The marginal distributions $P_s(\mathbf{x})$ and $P_t(\mathbf{x})$ may also differ, as may the conditional distributions $P_s(y|\mathbf{x})$ and $P_t(y|\mathbf{x})$. The task is to learn a model $f : \mathcal{X}_t \rightarrow \mathcal{Y}_t$ from the labeled source data $\mathcal{D}_s^{\text{train}}$, with some labeled target data, so that it generalizes to the held-out target test set $\mathcal{D}_t^{\text{test}}$. A model must therefore handle feature schema mismatch, label space mismatch, and distribution shift at the same time.

We adopt a *leave-one-state-out* protocol for evaluating cross-state QR prediction. It is the state-level analogue of holding out one domain for testing in domain generalization [28]. Each state t serves as the target once, and

the remaining states $\{\mathcal{D}_{s_1}, \mathcal{D}_{s_2}, \dots\}$ are pooled into a single source training set. To reflect real-world conditions, we evaluate both zero-shot and target-supervised transfer settings. First, a stratified 20% test set is held out from the target state $\mathcal{D}_t$ and used to score every setting. Holding this test set fixed keeps scores comparable across supervision levels. Then, a stratified draw of $p\%$ of the target state’s providers is created as the adaptation set, with $0 \leq p \leq 80$. At $p=0$, the setting is zero-shot, and no target label is used anywhere in training. At $p=80$, the adaptation set is the entire 80% that remains after the test set is held out. For levels in between, the providers outside the draw are left unused. At every level, all source states provider records are used for training. We evaluate cross-state prediction with the same metrics as within-state prediction (QWK and BA).

5 Experiments

5.1 Experimental Setup

5.1.1 Compared Models. We evaluate two groups of models on CCQ for both within- and cross-state QR prediction tasks. The first group, **tabular classifiers**, consumes the *preprocessed* release natively or the *raw* release embedded as a numerical table. The second group, **textual classifiers**, consumes the *raw* release by processing text natively. Each tabular classifier is marked with the representation it consumes: *-num* for *preprocessed* and *-txt* for *raw*. The textual classifiers consume only the *raw* representation and carry no suffix. The compared models are:

- **Tabular Classifiers:** (a.1) a majority-class baseline derived from the training set (*Dummy*); (a.2) logistic regression (*LR*); (a.3) random forest (*RF*) [12]; (a.4) XGBoost (*XGB*) [15]; (a.5) a tabular neural network (*TabNN*) [5]; and (a.6) a tabular prior-fitted network (*TabPFN*) [31].
- **Textual Classifiers:** (b.1) ModernBERT [81] finetuned on textualized records with a classification head (*MBERT*); (b.2) the same model trained with the CORN rank-consistent ordinal loss [70] instead of cross-entropy (*MBERT-CORN*); (b.3) an LLM (Qwen3-4B [67]) with a finetuned classification head over the frozen backbone (*LLM-CLS*); (b.4) the same frozen-backbone configuration with the most representative labeled provider of each class prepended to the input, a simplified form of retrieval augmentation [42] (*LLM-RAG*); and (b.5) the same LLM with LoRA adapters [32] trained alongside the classification head (*LLM-LoRA*).

For cross-state QR prediction (Section 4.2), we evaluate six target supervision levels, $p \in \{0, 16, 32, 48, 64, 80\}$. The models follow one of two training regimes. *Dummy*, *LR*, *RF*, *XGB*, and *TabPFN* are one-shot-fit models, which are fit once on the union of the pooled source states and the target draw. *TabNN*, *MBERT*, *MBERT-CORN*, and the three LLM variants are sequentially trained models, which are first pretrained on the pooled source states and then finetuned on the target draw.

5.1.2 Implementation Details. Table 4 lists the complete configuration of every model. We use scikit-learn [63] for *Dummy*, *LR*, and *RF*, PyTorch [62] for the neural models, and Hugging Face Transformers [85] for the language models. We tune hyperparameters only when required for convergence or for consistency across experiments, and otherwise run every method at its library defaults. Our goal is a uniform set of baselines rather than a per-experiment optimum, and a single configuration keeps comparisons consistent and fair across states. Since class imbalance is consistently present in CCQ, we apply class weights, or subsampling when a model has no weighting mechanism.

The tabular classifiers (a.1–a.6) natively consume the *preprocessed* representation (*-num*), while the textual classifiers (b.1–b.5) consume the *raw* representation. To write each *raw* record as text, we follow the text-based representation of PedCA-FT [46] (Figure 4). Each record becomes natural-language text, with its fields grouped into labeled sections in a fixed order: general provider information, curriculum, facility capabilities, enrollment details per age group, and multi-year compliance history. Boolean fields list only their true values and missing values are omitted. For within-state QR prediction, the tabular classifiers are evaluated on both representations. On

Table 4. Complete hyperparameter configuration. Values are held fixed across all 12 states, both tasks, and all six supervision levels. “Phase” is either (1) source pretraining or (2) target adaptation in sequential training. Each phase derives its own class weights and validation split.

Component	Configuration
<i>Training Protocol</i>	
Seed	42 globally (splits, folds, subsamples, initializations, bootstrap)
Cross-validation	5-fold stratified, shuffled; folds cached per state and shared by all methods
Target test block	stratified 20%, carved once per target, scored at every supervision level
Validation split	stratified 15% of each training phase; best epoch by validation QWK, patience 2 (textual classifiers; TabNN below)
<i>Tabular classifiers</i>	
Dummy	most-frequent class
LR	ℓ_2 penalty, $C=1.0$, lbfgs, max 2,000 iterations, balanced class weights
RF	100 trees, unlimited depth, 1 sample per leaf, $\sqrt{d}$ features per split, balanced subsample weights
XGB	500 rounds, depth 6, learning rate 0.05, row subsample 0.8, column subsample 0.8, $\lambda=1.0$, histogram tree method, per-row balanced sample weights
TabNN	$n_d=n_a=16$, 3 steps, $\gamma=1.5$, $\lambda_{\text{sparse}}=10^{-4}$, entmax masks [65]; Adam [38] at 2×10^{-2} with StepLR(20, 0.9); ≤ 200 epochs, patience 40 on balanced accuracy; batch $\min(256, n_{\text{train}})$ with virtual batch $\min(64, \text{batch})$; inverse-frequency class weights; transfer phase 2 warm-starts from phase 1
TabPFN	TabPFN v2, 8 ensemble members, in-context fitting (no gradient training); training set class-balanced by downsampling to the smallest class, then capped at 10,000 context rows
<i>Textual classifiers</i>	
MBERT	ModernBERT-base, 8,192-token window, classification head; class-weighted cross-entropy, AdamW [45] at 2×10^{-5} , linear decay with 10% warmup, weight decay 0.01, 4 epochs per phase, batch 8, length-grouped batches with dynamic padding
MBERT-CORN	identical backbone and optimization trained with the CORN rank-consistent ordinal loss, class-weighted by the same balanced weights as MBERT; prediction by the canonical cumulative-sigmoid count decode, probabilities by telescoping the cumulative terms
LLM-CLS	Qwen3-4B (dense) with the backbone frozen and only a classification head trained; head learning rate 10^{-3} , 4 epochs per phase, batch 8 with gradient accumulation 2 (effective 16), 8,192-token window, 10% warmup, weight decay 0.01, balanced class weights
LLM-RAG	identical head and optimization; each input is additionally prefixed with one exemplar per rating level, chosen as the MiniLM class medoid. Exemplars are drawn from the training pool only and self-excluded, so no scored provider ever supplies or receives its own exemplar; in the zero-shot setting exemplars come from the source pool, since no target labels exist
LLM-LoRA	same backbone with LoRA adapters ($r=16$, $\alpha=32$, dropout 0.05) on the attention and MLP projections, learning rate 10^{-4} , gradient checkpointing enabled

the *raw* representation, each textualized record is embedded with the sentence encoder *all-MiniLM-L6-v2* [69, 80]. Records longer than the encoder window are split into chunks, and the unit-normalized chunk embeddings are mean-pooled and normalized, so no record is truncated. The textual classifiers are evaluated on the textualized records only. In cross-state QR prediction, all models use the *raw* representation, since the *preprocessed* one follows state-specific schemas that cannot be pooled.

5.1.3 Evaluation Metrics. For both within- and cross-state QR prediction, we report balanced accuracy (BA) [13] and quadratic weighted kappa (QWK) [16] (Section 4). Since quality ratings are ordinal, QWK is the primary metric. It penalizes each error by its squared distance on the rating scale and corrects for the agreement expected

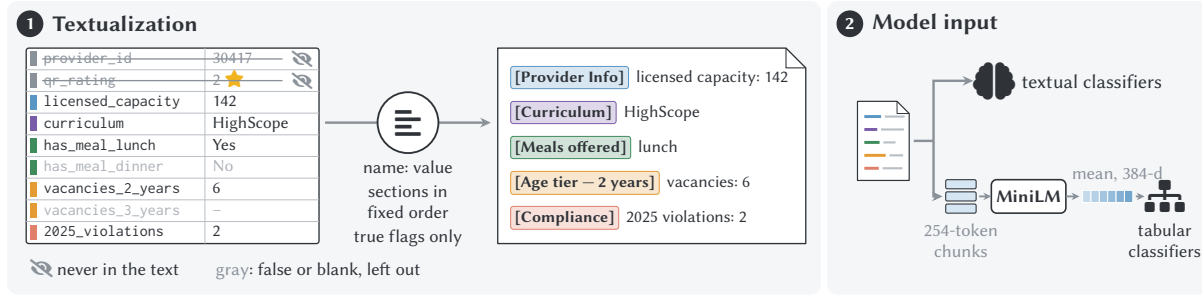


Fig. 4. Text-based representation of a provider record, following PedCA-FT [46]. (1) A simplified record in Georgia’s schema is written as text. False and blank values are left out, and the provider identifier and the rating never appear. (2) The textual classifiers read this text directly. For the tabular classifiers, all-MiniLM-L6-v2 embeds the text in 254-token chunks, and the normalized mean of the chunk embeddings is the 384-dimensional input.

by chance. Within-state results are the mean over the five folds, and cross-state results are a single point estimate on the held-out target test set.

To summarize overall performance across states and metrics, we also report Elo scores [11]. Elo gives each model a single rating derived from pairwise comparisons. Each combination of state and metric is one match between every pair of models, and the model with the higher score wins. Following existing work [11, 20], ratings are updated after each match, with a gap of Δ points giving the higher-rated model an expected win rate of $1/(1 + 10^{-\Delta/400})$. A 400-point gap therefore corresponds to a 91% expected win rate, and equal ratings to 50%. As in TabArena [20], we anchor RF-txt at 1,000 points in both settings and report 95% bootstrap confidence intervals. Each bootstrap round resamples the twelve states with replacement, and the reported score is the mean over 2,000 rounds. In the cross-state setting, the six supervision levels of a state are not independent samples. We therefore compute Elo separately at each level on the same resampled states and average it over the six levels, so that each level carries equal weight. Because the within- and cross-state ratings are computed from different sets of matches, scores can be compared within a setting but not across settings.

5.2 Within-State QR Prediction Experiment

Table 5 reports the within-state QR prediction results. Each state’s models are trained and evaluated only on its own providers, using stratified five-fold cross-validation on that state’s native rating scale. Tabular classifiers are evaluated on both representations, and textual classifiers on the *raw* representation only. Figure 5a summarizes overall performance with Elo scores computed from BA and QWK across all twelve states. Figure 6 shows the trade-off between performance and model complexity as the Pareto frontier of Elo score versus the number of tuned parameters.

5.2.1 Predictive Performance. Tabular classifiers on the *preprocessed* representation achieve the highest Elo scores (Figure 5a). XGB-num ranks first with 1519.1, followed by TabPFN-num (1422.2), LR-num (1375.1), and RF-num (1346.9). Every tabular classifier also scores higher on the *preprocessed* representation than on the *raw* one. Across LR, RF, XGB, TabPFN, and TabNN, the *-num* version exceeds the *-txt* version by 303.5 Elo points on average, an 85.2% expected win rate. The gap is smallest for LR (175.0 points) and largest for XGB (390.7 points). On the *raw* representation, MBERT is the strongest classifier, with an Elo score of 1308.0. It ranks fifth overall, ahead of TabNN-num (1172.0). Its expected win rate is 22.9% against XGB-num and 44.4% against RF-num. LLM-LoRA (1226.1) and LR-txt (1200.1) follow. MBERT-CORN scores 1162.1, 145.9 points below MBERT. Its QWK is higher than MBERT’s in five states (CO, MD, NC, OK, and WI), but its BA is lower in 11 of the 12. Lastly, XGB-txt (1128.4),

Table 5. Within-state QR prediction results on each state’s native QR scale. Each cell is the mean over five-fold cross-validation and reports BA and QWK, both as percentages. The best value per state and metric is in **bold**, and the second best is underlined.

Method	CA		CO		GA		KY		MD		MT		NC		NE		OK		SC		WA		WI	
	BA	QWK	BA	QWK	BA	QWK	BA	QWK	BA	QWK	BA	QWK	BA	QWK	BA	QWK	BA	QWK	BA	QWK	BA	QWK	BA	QWK
Dummy	20.00	0.00	20.00	0.00	33.33	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	25.00	0.00	23.00	0.00
<i>Preprocessed feature representation</i>																								
LR-num	39.45	29.33	41.37	51.69	<u>45.23</u>	<u>21.04</u>	41.51	32.43	44.75	64.31	37.86	27.23	<u>50.29</u>	44.90	<u>42.44</u>	<u>29.03</u>	49.48	64.64	<u>41.68</u>	36.25	<u>43.51</u>	27.09	44.71	50.59
RF-num	34.01	44.82	37.06	<u>56.82</u>	35.47	6.45	35.95	40.38	43.23	65.07	<u>42.67</u>	31.67	42.37	59.19	39.27	26.79	52.83	76.43	38.76	39.65	32.81	20.44	38.52	<u>52.75</u>
XGB-num	36.22	<u>42.59</u>	39.53	57.98	41.15	19.00	<u>41.76</u>	48.10	44.27	65.23	43.85	26.63	46.02	61.73	38.84	29.99	<u>53.07</u>	<u>76.44</u>	39.66	<u>42.18</u>	39.07	33.63	40.80	52.63
TabPFN-num	<u>36.65</u>	21.16	<u>40.67</u>	51.24	48.04	23.33	45.93	<u>40.70</u>	45.45	64.14	40.14	<u>27.49</u>	54.15	51.18	46.22	26.74	56.43	77.34	42.76	38.66	44.32	24.88	30.74	18.17
TabNN-num	24.78	18.58	36.22	46.92	40.18	14.27	38.13	29.53	44.90	63.82	39.85	20.62	48.40	48.01	42.09	18.57	48.76	64.75	40.28	37.72	40.93	21.37	<u>43.23</u>	47.82
<i>Raw feature representation</i>																								
LR-txt	33.66	29.80	35.66	42.34	39.36	10.85	38.90	33.18	44.18	64.38	40.14	26.53	48.27	49.81	38.22	26.16	43.95	62.58	40.32	37.98	35.58	20.39	37.70	37.86
RF-txt	28.66	28.45	31.83	43.48	33.81	1.36	26.26	18.11	43.71	65.26	38.61	26.63	33.11	44.08	34.86	18.82	43.14	66.27	37.85	41.60	27.27	4.73	31.30	31.03
XGB-txt	32.46	33.12	34.38	44.33	34.56	4.40	31.65	33.98	43.59	63.82	38.64	20.88	35.86	48.82	37.35	26.69	44.90	66.31	38.00	42.42	32.82	19.72	36.24	43.75
TabPFN-txt	33.67	14.18	35.04	38.76	41.86	13.60	35.96	27.47	<u>45.36</u>	64.35	37.56	10.62	43.16	43.42	36.07	13.34	49.09	65.61	41.35	40.80	41.79	19.92	19.80	6.10
TabNN-txt	29.64	14.36	31.25	31.16	38.19	8.16	30.24	15.82	44.43	63.40	32.13	12.22	39.20	37.70	34.66	12.87	43.08	59.80	40.37	39.04	35.35	15.57	34.96	30.15
MBERT	33.94	38.68	35.34	50.59	41.84	18.67	39.77	38.66	44.06	<u>70.63</u>	23.57	12.35	38.38	48.66	36.49	19.58	51.58	70.80	40.11	42.02	38.11	<u>33.35</u>	38.75	49.70
MBERT-CORN	28.54	27.34	35.11	55.21	34.88	4.69	32.05	34.84	40.06	70.77	21.06	0.78	32.19	50.62	31.80	19.04	50.92	75.55	35.89	40.95	36.16	29.57	40.53	53.21
LLM-CLS	30.66	24.44	34.58	42.18	38.63	9.63	30.90	29.77	41.36	66.37	32.72	17.47	33.46	44.81	36.66	21.54	37.96	43.92	34.61	22.58	33.85	25.27	32.64	40.84
LLM-RAG	24.10	11.75	33.52	43.32	39.23	10.32	32.57	26.80	41.96	66.08	28.31	8.55	32.85	41.28	33.50	18.74	36.31	42.69	38.74	35.09	35.13	27.00	29.25	30.88
LLM-LoRA	28.02	25.58	35.18	46.96	42.74	17.49	35.35	34.88	42.52	68.80	35.71	23.67	36.72	47.48	38.64	24.95	50.07	68.92	40.84	38.01	35.02	27.09	34.50	50.42

TabPFN-txt (1069.5), LLM-CLS (1007.3), RF-txt (1000.0), TabNN-txt (920.0), and LLM-RAG (911.0) form the lower tier, and the Dummy baseline scores 474.6.

However, Elo combines both metrics, and the two do not always agree. TabPFN-num owes much of its second place to BA, where it is best in 8 of the 12 states. On QWK, the primary metric, its mean of 38.75 falls below that of XGB-num (46.34), RF-num (43.37), MBERT (41.14), LR-num (39.88), and LLM-LoRA (39.52). The QWK ranking is thus led by the two tree ensembles on the *preprocessed* representation. Table 5 shows the same pattern state by state. A tabular classifier on the *preprocessed* representation has the best QWK in 9 of the 12 states and the best BA in all 12. Specifically, XGB-num has the best QWK in five states, and RF-num and TabPFN-num in two each. The three exceptions are MD and WI, where MBERT-CORN is best (70.77 and 53.21), and SC, where XGB-txt is best (42.42). The margins in MD and WI are small. MBERT-CORN leads MBERT by 0.14 points in MD and RF-num by 0.46 points in WI.

The best QWK also varies widely across states. In OK and MD, the strongest models exceed 70. TabPFN-num reaches 77.34 and XGB-num 76.44 in OK, and MBERT-CORN reaches 70.77 and MBERT 70.63 in MD. In CA, CO, KY, NC, SC, and WI, the best QWK lies between 42.42 (SC) and 61.73 (NC). In the remaining four states, even the best model stays below 35, with 33.63 in WA, 31.67 in MT, 29.99 in NE, and 23.33 in GA. Overall, the best QWK per state spans 54 points, and the best model changes from state to state.

5.2.2 Predictive Performance Versus Model Complexity. For the tabular classifiers on the *preprocessed* representation (Figure 6a), the Pareto frontier has only two points. TabPFN-num reaches an Elo score of 1422.2 with zero tuned parameters, because it learns in context. It dominates LR-num, RF-num, and TabNN-num, which all score lower while tuning more parameters. XGB-num is the only model that improves on it. It gains 96.8 Elo points (a 63.6% expected win rate) with about 5.0×10^4 tuned parameters. Figure 6b shows the frontier for the classifiers on the *raw* representation. TabPFN-txt again starts the frontier at zero tuned parameters, with an Elo score of 1069.5. LR-txt adds about 1.8×10^3 tuned parameters and gains 130.6 points (68.0%). LLM-LoRA tunes about four orders of magnitude more parameters than LR-txt but adds only 26.0 points (53.7%). MBERT, the last point on the frontier, finetunes all of its parameters. It tunes about 4.5 times as many parameters as LLM-LoRA and adds another 81.9 points (61.6%). Lastly, MBERT-CORN tunes the same backbone with one fewer output in its head. However, it scores 64.0 points below LLM-LoRA and is not on the frontier.

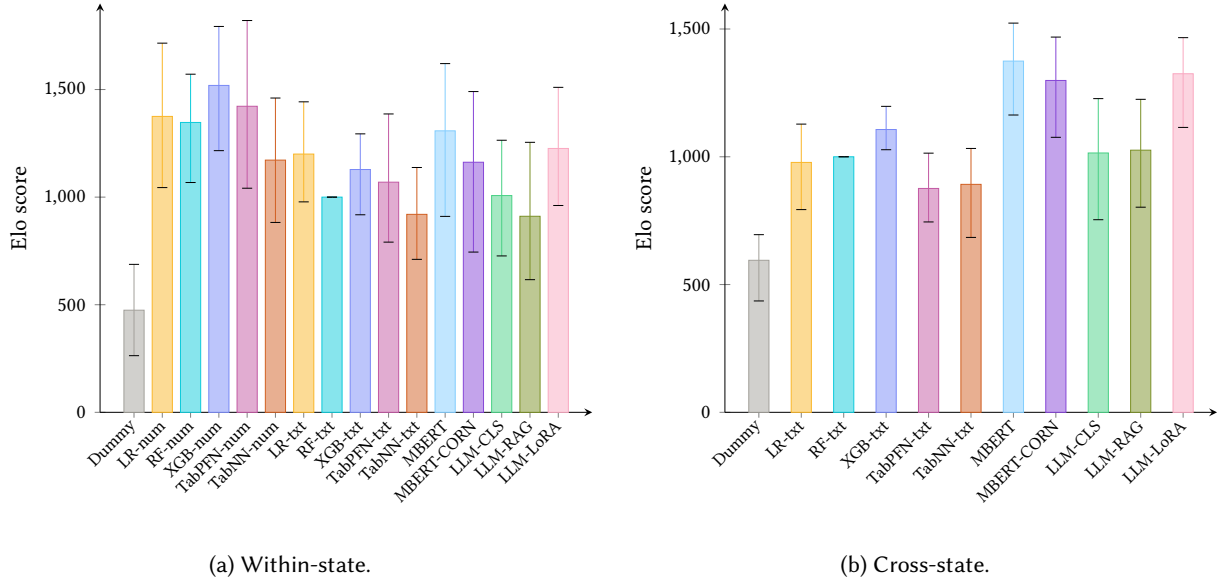


Fig. 5. Elo ratings from the within-state and cross-state experiments. The Elo scale uses a 400-point gap, which corresponds to a 91% expected win rate. Error bars show 95% confidence intervals from 2,000 bootstrap rounds that resample the twelve states. Cross-state scores are averaged over the six supervision levels. RF-txt is the 1,000-point anchor in both panels. The two panels are computed on different sets of comparisons, so scores are comparable within a panel but not across panels.

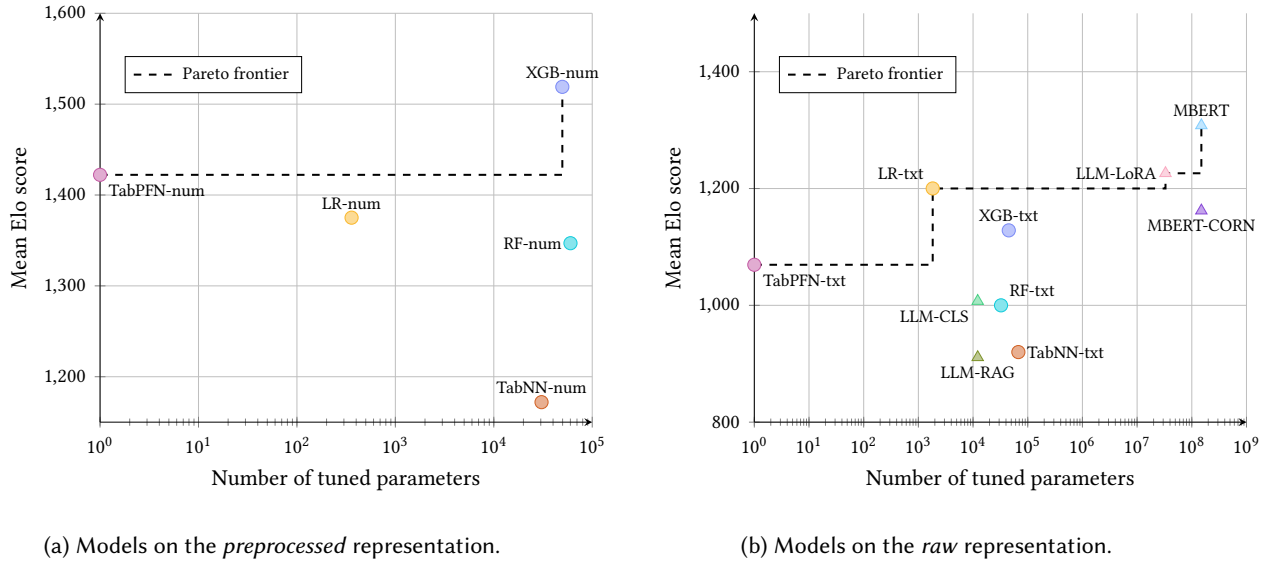


Fig. 6. Pareto frontier of mean Elo score versus number of tuned parameters for within-state QR prediction. The dashed line marks the Pareto-optimal set. Parameter counts are averaged over the 12 per-state models. For tree ensembles we count leaves. TabPFN updates no weights and is plotted at 10^0 .

5.2.3 Discussion. For within-state QR prediction, **tabular classifiers on the *preprocessed* representation are the recommended choice for the best predictive performance.** They hold the four highest Elo scores, the best QWK in 9 of 12 states, and the best BA in all 12. On average, they also gain 303.5 Elo points over the same model on the *raw* representation. These results are consistent with prior work showing that tree-based models remain state of the art on medium-sized tabular data [27]. Prior work also shows that text models that split numbers into sub-word tokens represent numerical values less precisely [79]. The Pareto analysis adds an efficiency perspective to this recommendation. TabPFN-num ranks second overall with no tuned parameters and outperforms both LR-num and RF-num, so it is a strong default when training resources are scarce. XGB-num leads both the Elo ranking and the mean QWK, so it is the better choice when accuracy matters most. When only the *raw* text is available, MBERT is the strongest option. It beats LLM-LoRA in 17 of the 24 state-metric comparisons, even though its backbone is about 27 times smaller. On the other hand, its ordinal variant, MBERT-CORN, is less reliable within a state. It has the best QWK in MD and WI, but its QWK falls close to zero in MT (0.78), the smallest state, and in GA (4.69), the only three-level state. LR-txt is a strong low-cost alternative, with an Elo score close to that of LLM-LoRA at a small fraction of the tuned parameters. Importantly, these recommendations describe a representative set of model families under the single, uniform configuration of our benchmark. Because no method is tuned per state (Section 6.1), they rank the families as evaluated here rather than bound what any of them can reach.

Whether the evaluated models can already **serve as reliable surrogate QR predictors** depends on how well they agree with human raters. In a few states, the strongest models reach what is commonly interpreted as substantial agreement (61–80) [41]. In OK, TabPFN-num attains a QWK of 77.34 and XGB-num 76.44. In MD, MBERT-CORN and MBERT attain 70.77 and 70.63. XGB-num’s 61.73 in NC sits at the lower edge of this band. At these levels, model predictions could already support decision-making, for example as a monitoring tool. However, in WA, MT, NE, and GA, even the best QWK stays below 35 (33.63, 31.67, 29.99, and 23.33, respectively). This is only fair agreement, with limited practical use for high-stakes decisions. Current models are therefore not yet ready to generate reliable surrogate ratings in every state.

5.3 Driving Factor Analysis

QRIS are built not only to rate child care providers but also to help them improve. Predictive accuracy alone does not tell providers what to act on. To connect ratings with improvement, state agencies and providers need to know which features push a provider’s rating up or down. Feature attribution can turn black-box predictions into this kind of actionable guidance. It shows each state which factors drive its quality ratings. States can then check whether those factors are the ones linked to children’s outcomes.

5.3.1 Analysis Setup. We conduct the driving factor analysis with SHapley Additive exPlanations (SHAP) [47, 48]. A SHAP value measures how much a feature contributes to one provider’s prediction, relative to the average prediction in that state. We compute exact SHAP values, average their absolute values across providers, and rank features by this mean. The ranking therefore describes what the model relies on across a state’s entire provider population.

We explain XGBoost, since XGB-num performs consistently well across the within-state experiments (Section 5.2). We make two deliberate changes from the benchmarked configuration. First, the explained model (XGB-reg) is fit with a regression objective that treats the rating as a number, rather than with a multi-class classification objective. It otherwise uses the same *preprocessed* representation and the same configuration as XGB-num (Table 4). A regression model yields one SHAP value per feature that directly reflects its effect on the predicted rating, whereas a multi-class classifier would produce a separate SHAP value for each rating level. Second, the explained model is fit on all rated providers in a state rather than within five-fold cross-validation, so that it represents the full dataset. Because the explained model is not the one whose accuracy is reported in the

within-state benchmark, we also run it through the same five-fold protocol and compare its QWK with that of the benchmarked classifier.

To separate real feature contributions from noise, we implement Boruta’s shadow features [40]. Each shadow feature is a shuffled copy of a real column. It keeps the column’s marginal distribution but breaks any association with the rating. The noise floor is the largest mean absolute SHAP value among all shadow features, and we interpret only the features whose attribution exceeds it. Because tree models tend to favor high-cardinality continuous features over binary ones [73], we sample shadow features by cardinality rather than uniformly at random. For each state, we report how many features clear the noise floor out of all features in its *preprocessed* representation.

5.3.2 Analysis Results. Table 6 reports the fidelity check, which compares the five-fold QWK of XGB-reg with that of XGB-num, the classifier used in the within-state benchmark (Table 5). The differences range from -5.83 (MT) to $+3.86$ (NE) percentage points. XGB-reg scores higher in 9 of the 12 states and by 0.83 points on average. Only MT shows an absolute difference above 5 points.

Table 6. Fidelity of the XGBoost regression model (XGB-reg) used for SHAP analysis. Δ is its QWK difference from the XGBoost classifier (XGB-num). Above floor counts the features that exceed the shadow-feature noise floor, and Total counts the features of the *preprocessed* representation.

State	QWK (%)			#Features	
	XGB-num	XGB-reg	Δ	Above floor	Total
CA	42.59	45.85	+3.25	4	82
CO	57.98	58.95	+0.97	7	117
GA	19.00	18.26	-0.74	5	278
KY	48.10	49.65	+1.55	9	112
MD	65.23	67.95	+2.72	3	34
MT	26.63	20.80	-5.83	0	20
NC	61.73	63.28	+1.55	5	30
NE	29.99	33.86	+3.86	5	87
OK	76.44	77.34	+0.90	7	26
SC	42.18	45.53	+3.36	2	54
WA	33.63	30.58	-3.06	5	107
WI	52.63	54.01	+1.38	6	36

Figure 7 shows the ten highest-attribution features in each state. Across all states, 70 of these 120 features (58.3%) are positive contributors, whose higher values push the predicted rating up. The other 50 (41.7%) are negative contributors. Several states are dominated by positive contributors. CO, SC, MD, OK, and WA have 9, 8, 7, 7, and 7 positive features in their top ten. On the other hand, CA, MT, and WI lean negative, with only 2, 4, and 4 positive features.

Table 6 also reports how many features in each state clear the shadow-feature noise floor, out of the total available. The count ranges from 0 (MT) to 9 (KY), with 58 of 983 features overall. The dashed line in each panel of Figure 7 marks the noise floor. We take the direction of a feature from the correlation between its value and its attribution, which is also what the panel colors encode. Of the 58 features, 37 push the predicted rating up and 21 push it down. Three of them have a correlation below 0.03 in absolute value and are better read as undirected. These are the cost of full-time care at ages 6–8 in KY, licensed capacity in MD, and weekly hours in WA. The 58 features fall into six groups: capacity, operating schedule, and cost (18); program type, licensure, and accreditation (10); geographic and license-administration identifiers (9); services and population served (9); compliance and monitoring (7); and subsidy or public-program participation (5).

Institutional formalization is the clearest positive pattern. Center-based program type clears the floor and is positive in four states, and licensed capacity is positive in seven. All five subsidy or public-funding indicators that clear the floor are positive, as are the public pre-K indicator in NC and Head Start programming in WI. The clearest negative signals are rule violations in GA, enforcement actions in WI, weekend care in WI, and family child care homes in WA. The last is the only home-based indicator that clears the floor. A higher compliance rate in OK raises the predicted rating, which fits this pattern. However, monitoring records are mixed. More inspections in KY, more monitoring visits in OK, and a documented injury investigation in CO all raise the predicted rating, while more distinct inspection types in KY lower it. **Signals also do not generalize across states.** Operating hours lower the predicted rating in CA, CO, and GA but raise it in OK. License age is positive in CO but negative in NC. Lastly, MT is the only state in which no feature clears the noise floor. Its top feature, a regional indicator, falls just short of it (0.195 against 0.198). When the MT analysis is repeated with 30 different random seeds, this indicator clears the floor in 18 runs, so we treat MT’s attribution as inconclusive rather than absent.

5.3.3 Discussion. The fidelity check supports reading the SHAP rankings as explanations of the benchmarked classifier. XGB-reg differs from XGB-num by less than 5 QWK points in 11 of 12 states, and it is slightly more accurate on average. MT is the exception, with a gap of -5.83 points and an unstable noise floor. MT has only 181 rated providers and 20 binary features in its *preprocessed* representation. The balanced class weights give each of its six one-star providers about 12 times the weight of a two-star provider. A shuffled column can therefore fit a few heavily weighted providers about as well as a real feature can. Without class weights, the regional indicator clears the floor in 29 of 30 seeded runs. We therefore read MT’s attributions as inconclusive and draw no conclusions from them.

In the other states, a common pattern emerges. Larger, center-based, publicly funded, and accredited providers tend to receive higher predicted ratings, while rule violations, enforcement actions, and some less formal care arrangements lower them. This is a tendency rather than a rule, as it is state specific. Overall, this pattern **aligns with child care regulations and existing studies**. *Compliance* indicators provide the clearest alignment. Prior work finds that higher-quality programs generally have fewer licensing violations, although the pattern does not hold in every study or state [44]. Compliance with key health and safety licensing indicators has also been linked to positive child outcomes [22]. The positive monitoring records in KY, OK, and CO do not necessarily contradict this. They may reflect exposure to oversight rather than lower quality, for example if larger programs are inspected more often. *Program type and accreditation* features also align with well-documented quality differences. Accreditation is a common marker of quality. For example, NAEYC accreditation is widely relied on as a standard of quality and indicates better-than-average care [83]. Similarly, subsidy acceptance is a positive contributor in several states. This is consistent with federal rules that require subsidized providers to meet baseline health and safety standards, and with the higher subsidy rates that many states pay for higher quality [1].

Capacity and operational scale present a more nuanced picture. Larger licensed capacity is generally a positive contributor. This is in line with evidence of economies of scale in center-based care, where centers serving more children operate at a lower cost per child without an apparent loss of quality [30]. However, weekend care and long operating hours are negative contributors in some states. This contrasts with the finding that centers with longer hours of operation had lower costs per child with no apparent ill effects on quality [30]. Lastly, geographic identifiers also surface above the noise floor. This suggests that unmeasured regional factors, such as local economic conditions, child care supply, or differences in how regulations are implemented, may influence ratings in ways that our dataset only partially captures.

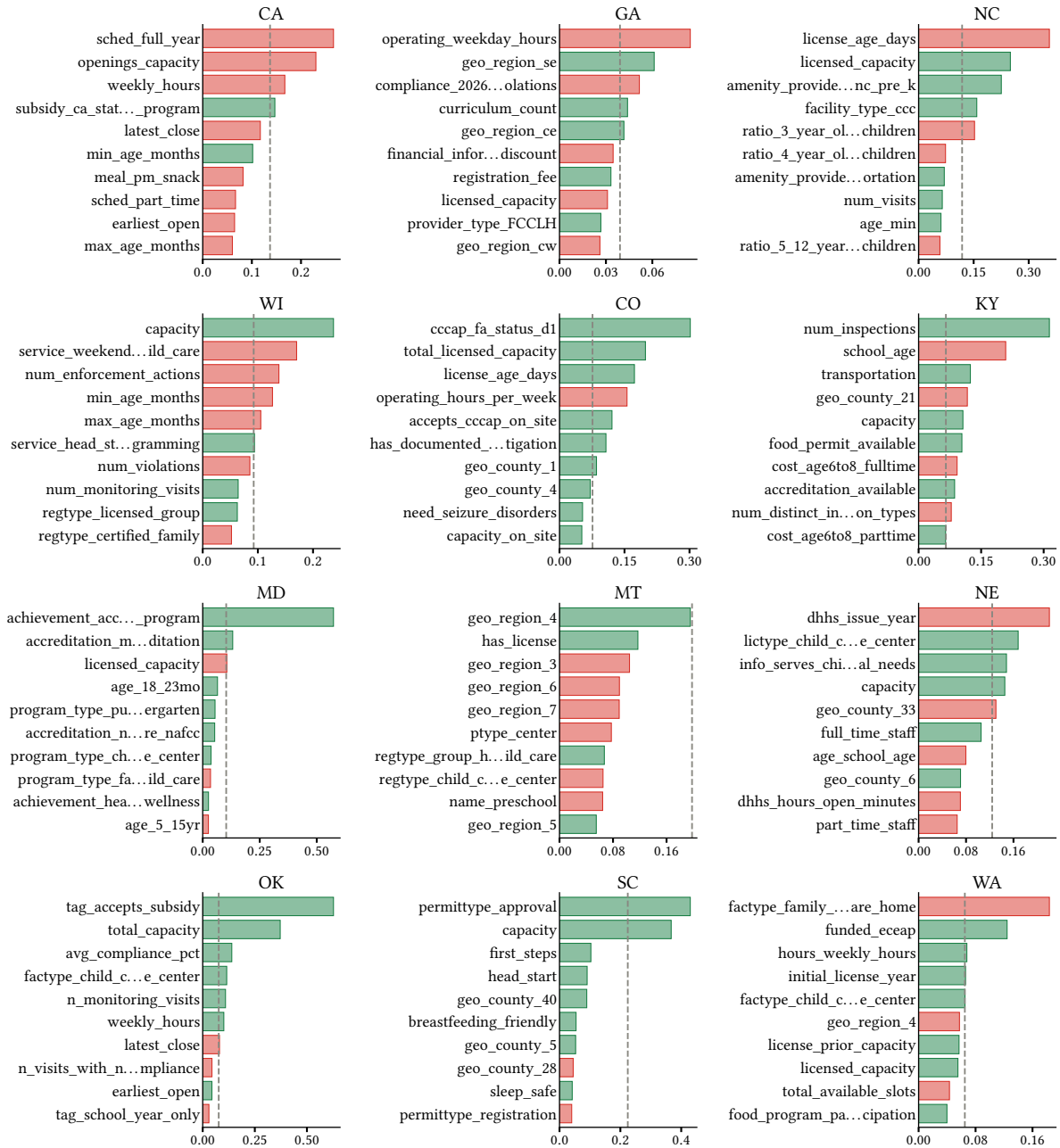


Fig. 7. Ten highest-attribution features per state, ranked by mean |SHAP| over that state's providers. The dashed line is the shadow-feature noise floor, and bars to its left are indistinguishable from shuffled noise. Color gives the direction: green where higher feature values push the predicted rating up, red where they push it down. Feature names are state-specific and horizontal scales differ across panels, so bar lengths are comparable within a panel but not across panels.

5.4 Cross-State QR Prediction Experiment

Table 7 reports the leave-one-state-out transfer results at every target supervision level. Figure 5b summarizes overall cross-state performance with Elo scores computed from BA and QWK across all targets and supervision levels. All models consume the *raw* representation, so every tabular classifier appears in its *-txt* form.

As described in Section 5.1.1, the one-shot-fit models are fit once on the union of the pooled source states and the $p\%$ target draw. The sequentially trained models are first pretrained on the pooled source states and then finetuned on the target draw. Each phase derives its own class weights and early-stopping split. At $p=0$, there is no target data to finetune on, so the zero-shot result of a sequentially trained model is its source-only pretraining result. Lastly, each experiment uses the target state’s native rating scale. Georgia is therefore left out of the five-level source pools. When Georgia is the target, the five-level source states are collapsed onto the reconciled three-level label space as $\{1, 2\} \rightarrow 1$, $\{3\} \rightarrow 2$, and $\{4, 5\} \rightarrow 3$.

5.4.1 Predictive Performance. MBERT attains the highest Elo score (1374.4), followed by LLM-LoRA (1324.9) and MBERT-CORN (1298.6) (Figure 5b). XGB-txt follows at 1106.7. LLM-RAG (1025.8), LLM-CLS (1014.9), RF-txt (1000.0), and LR-txt (977.9) cluster around the anchor, ahead of TabNN-txt (892.1) and TabPFN-txt (876.0). The Dummy baseline scores 594.9. The 267.7-point gap between MBERT and XGB-txt corresponds to an 82.4% expected win rate. In contrast, the 49.5-point gap between MBERT and LLM-LoRA corresponds to only 57.1%.

In the zero-shot setting ($p=0$), most models perform close to chance in 11 of the 12 targets. Across the 120 non-Dummy cells at this level, the median QWK is 2.66, 41 cells are at or below zero, and only 16 exceed 20. MD accounts for 10 of these 16 cells. Every non-Dummy model exceeds a QWK of 20 on MD, with a median of 43.10 and a maximum of 49.08 (TabNN-txt). In the other eleven targets, the median zero-shot QWK lies between -6.69 (CA) and 4.91 (NE). Averaged over the twelve targets, the best zero-shot model is LLM-LoRA with a mean QWK of 9.96, and no model reaches 10.

Once target supervision is available, the finetuned language models take the lead. At $p=16$, MBERT, MBERT-CORN, or LLM-LoRA holds the best QWK in 10 of the 12 targets. The exceptions are MT, where XGB-txt leads, and SC, where RF-txt leads. The three models also hold the best QWK in at least 10 of 12 targets at every higher level. Among them, MBERT-CORN wins most often. It has the best QWK in 28 of the 60 combinations of target and supervision level with $p > 0$, against 12 for MBERT and 11 for LLM-LoRA. It also has the highest mean QWK at $p=80$, with 46.11 against 45.32 for MBERT and 44.13 for LLM-LoRA. The remaining models trail well behind at that level: XGB-txt at 36.27, LLM-RAG at 33.13, RF-txt at 32.11, LR-txt at 29.55, LLM-CLS at 27.75, TabPFN-txt at 26.10, and TabNN-txt at 25.90. However, MBERT-CORN’s lead is specific to QWK. On BA, MBERT beats it in 42 of these 60 combinations, which helps explain why MBERT ranks higher by Elo. This split is consistent with CORN’s ordinal objective [70], which targets the ordering of ratings that QWK rewards rather than per-class accuracy. Lastly, the GA target, which tests transfer across rating scales, stays low throughout. Its best QWK at any supervision level is 21.96 (MBERT-CORN, $p=64$).

5.4.2 Effect of Target Supervision. Figure 8a plots the mean QWK over the twelve targets against the supervision level p for all ten models. For every model, the largest gain occurs between $p=0$ and $p=16$. The three finetuned language models rise the most in this first step: MBERT-CORN from 4.54 to 37.87, LLM-LoRA from 9.96 to 38.42, and MBERT from 8.67 to 37.02. XGB-txt rises from 5.08 to 26.74, RF-txt from 5.67 to 24.98, LR-txt from 8.20 to 21.64, LLM-CLS from 1.62 to 19.99, LLM-RAG from 7.71 to 19.96, TabPFN-txt from 5.93 to 18.61, and TabNN-txt from 6.24 to 15.09. Relative to the total gain from $p=0$ to $p=80$, this first step accounts for 83% of the gain for LLM-LoRA, 80% for MBERT-CORN, 77% for MBERT, 73% for RF-txt, 70% for LLM-CLS, 69% for XGB-txt, and 63% for both LR-txt and TabPFN-txt. LLM-RAG (48%) and TabNN-txt (45%) are the exceptions and gain more gradually.

Table 7. Leave-one-state-out cross-state QR prediction results. At every target supervision level p , each cell is scored on the same stratified 20% test set of the held-out target state. The best value per state and metric within each supervision level, excluding Dummy, is in **bold**, and the second best is underlined. Dummy predicts the majority class of its training set (the pooled source states plus the target draw).

Method	CA		CO		GA		KY		MD		MT		NC		NE		OK		SC		WA		WI	
	BA	QWK	BA	QWK	BA	QWK	BA	QWK	BA	QWK	BA	QWK	BA	QWK	BA	QWK	BA	QWK	BA	QWK	BA	QWK	BA	QWK
0% target supervision																								
Dummy	20.00	0.00	20.00	0.00	33.33	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	0.00	0.00	0.00	0.00
LR-txt	14.77	-3.88	22.38	17.31	32.00	2.06	20.48	5.80	23.04	35.10	24.71	5.19	17.78	-5.89	18.35	11.36	20.51	1.06	26.79	21.66	27.12	1.01	25.92	7.60
RF-txt	13.53	-9.36	17.58	2.62	32.92	-0.55	14.51	2.41	27.73	48.96	19.00	-8.22	23.26	-0.57	19.66	4.57	16.61	-9.08	24.99	20.88	22.41	3.87	23.34	12.52
XGB-txt	41.28	-1.22	18.67	-0.03	29.30	-6.32	17.74	2.75	23.35	40.56	29.36	-1.44	36.48	23.06	18.21	3.28	14.57	-7.02	20.12	-0.70	21.50	-4.65	27.86	3.43
TabPFN-txt	13.85	-11.21	23.20	5.76	38.09	6.20	21.53	-0.72	33.66	33.03	27.00	9.09	31.60	28.63	20.15	4.12	20.30	-0.46	20.00	0.00	22.74	-5.43	24.91	2.15
TabNN-txt	9.91	-7.80	22.15	2.03	33.31	1.42	19.66	-3.15	27.99	49.08	28.29	7.59	18.77	0.19	20.81	-5.18	26.48	27.69	20.00	0.00	22.90	-3.67	14.21	6.65
MBERT	18.18	-1.61	18.72	10.29	39.67	11.61	21.87	1.49	25.19	46.06	14.29	0.00	17.60	5.95	24.44	10.67	21.33	15.49	20.00	2.50	17.16	1.00	25.37	0.63
MBERT-CORN	17.14	-8.00	20.00	1.11	34.94	2.70	18.24	0.54	25.18	20.10	18.43	2.78	11.70	-0.42	24.25	14.26	19.87	-2.14	23.30	7.64	25.26	10.75	27.25	5.17
LLM-CLS	20.88	-6.14	13.81	-6.70	35.28	6.32	20.00	-0.19	25.00	45.63	22.50	-4.42	28.60	3.95	15.94	3.00	18.48	-0.99	20.29	-12.23	23.96	-2.19	11.67	-6.61
LLM-RAG	18.46	-7.24	22.60	11.10	35.33	10.58	21.12	8.74	21.97	46.18	23.93	6.86	17.22	1.07	20.00	5.24	15.75	0.66	20.00	0.00	23.03	1.54	25.12	7.79
LLM-LoRA	19.58	-0.28	18.33	5.85	32.46	-0.13	27.39	9.25	29.67	32.79	11.43	9.02	25.05	13.14	32.39	24.13	20.43	14.38	20.60	0.10	21.63	14.04	21.99	-2.73
16% target supervision																								
Dummy	20.00	0.00	20.00	0.00	33.33	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	0.00	0.00	0.00	0.00
LR-txt	24.44	9.10	27.90	31.28	35.21	1.63	21.33	6.86	41.45	68.77	21.93	-2.41	25.04	19.42	22.44	12.84	34.88	55.42	24.85	11.13	27.36	5.51	38.42	40.08
RF-txt	24.54	1.00	28.65	34.89	34.02	-3.60	20.54	0.69	41.79	67.84	41.57	30.14	31.49	39.28	29.49	12.51	35.67	58.10	37.39	35.95	25.12	-2.08	31.95	25.07
XGB-txt	22.44	3.12	30.41	36.54	34.15	2.62	21.86	3.91	41.52	67.23	45.00	37.84	29.13	34.01	28.79	8.76	38.49	59.08	33.27	33.84	26.47	6.07	33.82	27.90
TabPFN-txt	20.03	2.11	29.89	42.19	34.14	0.13	20.00	0.00	41.64	68.58	28.57	5.63	25.28	18.05	22.97	9.46	33.31	59.45	20.26	-4.20	25.00	0.00	30.35	21.86
TabNN-txt	19.24	2.47	23.66	20.47	28.09	-7.88	22.88	13.73	36.90	62.64	26.29	10.69	25.16	25.50	15.50	1.25	26.39	18.69	20.00	0.00	33.80	19.04	29.80	14.52
MBERT	31.54	30.83	32.50	48.38	40.86	15.78	32.02	23.41	44.72	71.68	33.29	10.94	36.20	47.03	34.23	36.58	42.55	58.19	37.27	29.17	36.32	30.15	39.41	42.07
MBERT-CORN	26.74	19.20	29.77	50.50	38.43	14.49	22.36	19.12	44.25	72.11	15.36	13.79	40.23	52.03	34.59	29.29	45.20	70.33	19.45	31.89	36.82	29.95	41.10	51.70
LLM-CLS	24.61	2.94	27.61	30.94	34.06	0.62	22.40	3.50	43.52	68.72	17.00	0.00	33.85	32.59	30.85	21.78	27.96	24.50	28.75	7.33	33.53	19.33	33.86	27.58
LLM-RAG	26.79	21.91	31.00	31.38	35.40	3.68	22.66	16.51	42.60	68.16	14.00	5.41	31.83	43.66	24.55	2.73	20.44	1.71	28.39	15.04	26.41	5.54	30.65	23.75
LLM-LoRA	33.23	36.44	29.58	37.60	36.80	10.59	38.55	35.21	44.10	70.93	39.29	28.07	35.53	41.56	37.96	31.87	42.30	63.07	37.02	31.66	35.49	34.99	35.14	39.10
32% target supervision																								
Dummy	20.00	0.00	20.00	0.00	33.33	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	0.00	0.00	0.00	0.00
LR-txt	22.81	5.15	29.76	38.89	34.66	0.62	21.32	7.52	41.58	68.95	28.71	10.96	26.00	28.81	25.06	14.75	37.86	60.01	26.94	18.96	28.28	10.21	38.14	37.09
RF-txt	26.42	20.00	30.37	36.74	36.03	5.94	21.96	9.56	42.79	67.97	28.57	24.32	32.13	43.54	27.41	9.05	41.18	61.66	37.60	42.58	26.88	1.05	33.61	27.70
XGB-txt	27.52	20.28	32.34	43.35	35.60	4.50	24.67	18.74	42.67	69.51	30.50	31.58	31.62	41.50	28.80	9.63	41.25	63.19	41.97	50.80	29.88	12.97	35.08	36.45
TabPFN-txt	20.00	0.00	29.74	42.22	34.10	0.05	19.91	0.60	40.03	68.70	18.57	-5.06	25.76	24.53	24.30	-2.11	35.89	61.98	30.50	44.43	25.00	0.00	31.22	24.37
TabNN-txt	24.20	18.39	28.31	24.94	31.02	-0.11	24.38	14.78	38.42	66.46	15.14	-2.63	32.88	16.55	24.32	-1.75	40.40	57.82	19.68	1.09	25.83	4.07	32.06	24.24
MBERT	27.61	32.53	32.28	44.06	44.40	16.46	39.53	38.09	43.92	71.70	19.00	16.44	38.50	56.48	31.86	15.64	52.54	74.39	36.38	37.87	36.07	28.45	41.94	47.85
MBERT-CORN	35.62	41.39	31.67	48.61	41.73	21.43	27.13	30.73	41.25	71.34	26.71	40.74	27.76	45.10	27.90	19.73	50.54	76.88	29.44	39.25	34.02	24.26	44.10	50.71
LLM-CLS	47.90	24.17	27.10	31.26	38.58	12.48	27.66	20.27	44.09	70.92	18.21	-17.91	29.72	29.99	24.55	5.56	30.27	35.93	33.05	28.37	39.26	31.12	35.43	31.00
LLM-RAG	20.98	1.39	29.16	39.16	40.39	10.22	27.14	26.96	41.70	71.23	18.00	6.06	34.72	44.46	27.29	7.54	28.25	38.02	25.16	11.44	35.56	28.12	30.52	22.40
LLM-LoRA	27.49	27.50	34.49	43.68	42.99	16.86	38.34	36.62	43.07	72.11	18.93	14.63	32.33	39.63	33.08	26.47	54.61	72.66	34.02	28.46	35.47	27.54	40.23	44.89
48% target supervision																								
Dummy	20.00	0.00	20.00	0.00	33.33	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	0.00	0.00	0.00	0.00
LR-txt	22.63	8.83	31.19	40.55	35.53	3.24	22.55	10.63	41.26	68.75	27.79	15.19	26.02	30.82	26.28	17.32	38.02	60.44	28.64	25.09	28.95	11.62	38.10	38.20
RF-txt	26.45	15.06	29.21	36.37	36.52	6.86	26.86	24.51	42.75	68.41	29.93	29.27	32.94	39.28	30.52	14.20	39.09	61.54	40.36	51.46	28.38	4.79	34.95	32.50
XGB-txt	30.06	29.90	30.73	44.18	35.69	5.30	26.26	23.81	42.40	69.02	30.43	36.59	30.71	38.66	29.38	20.21	39.79	65.16	43.49	57.15	30.14	10.92	36.54	36.00
TabPFN-txt	21.06	4.42	31.00	42.55	34.38	0.78	21.24	15.07	40.11	68.99	20.00	10.13	26.59	29.82	26.78	19.00	33.95	61.20	33.50	40.80	26.81	3.69	33.49	32.45
TabNN-txt	20.03	11.09	31.71	32.33	35.67	1.84	24.86	21.06	38.25	65.53	12.50	0.00	38.17	43.30	33.60	1.51	39.03	61.59	22.85	3.22	37.07	18.37	36.49	31.50
MBERT	35.68	39.93	33.59	47.87	40.92	16.80	36.20	34.66	44.66	71.30	24.29	33.90	39.98	55.89	36.63	28.50	54.26	72.33	45.13	51.59	40.27	34.88	45.32	52.49
MBERT-CORN	37.97	47.58	31.87	52.99	40.50	16.97	38.81	33.64	40.26	71.12	23.79	19.67	38.63	53.56	29.72	18.91	50.85	75.55	43.87	14.18	40.35	34.35	43.17	57.49
LLM-CLS	48.02	28.04	31.84	37.33	39.95	11.21	32.61	27.44	41.26	70.94	22.64	0.00	33.97	45.90	27.31	17.43	34.66	58.99	28.10	6.49	43.05	28.27	37.96	36.03
LLM-RAG	28.15	26.53	32.16	37.94	36.58	5.42	32.42	26.37	40.33	70.35	20.00	15.05	30.81	40.93	33.12	15.15	31.73	35.68	34.62	12.74	36.89	28.46	34.28	26.00
LLM-LoRA	27.01	26.18	37.30	48.30	40.07	14.71	38.73	37.58	44.67	71.49	32.14	34.67	36.39	49.43	36.19	19.66	57.45	74.23	38.52	35.20	42.02	31.27	40.83	47.15
64% target supervision																								
Dummy	20.00	0.00	20.00	0.00	33.33	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	20.00	0.00	0.00	0.00	0.00	0.00
LR-txt	21.43	17.13	30.69	43.51	33.13	3.28	23.47	14.99	41.44	68.97	32.79	19.75	26.45	31.58	26.12	14.26	38.60	61.43	28.64	26.46	29.78	10.29	38.	

Beyond $p=16$, the other eight curves flatten. Their additional gain from $p=16$ to $p=80$ ranges from 5.71 for LLM-LoRA to 9.53 for XGB-txt. Only MBERT and LR-txt rise at every step. LLM-LoRA dips at $p=32$ and ends below its $p=64$ value. XGB-txt, TabPFN-txt, and LLM-CLS peak at $p=48$, and MBERT-CORN and RF-txt peak at $p=64$. In contrast, LLM-RAG’s curve is the least regular. It falls from 28.40 at $p=48$ to 25.29 at $p=64$ and then rises to 33.13 at $p=80$. TabNN-txt does the reverse, falling from 28.86 at $p=64$ to 25.90 at $p=80$.

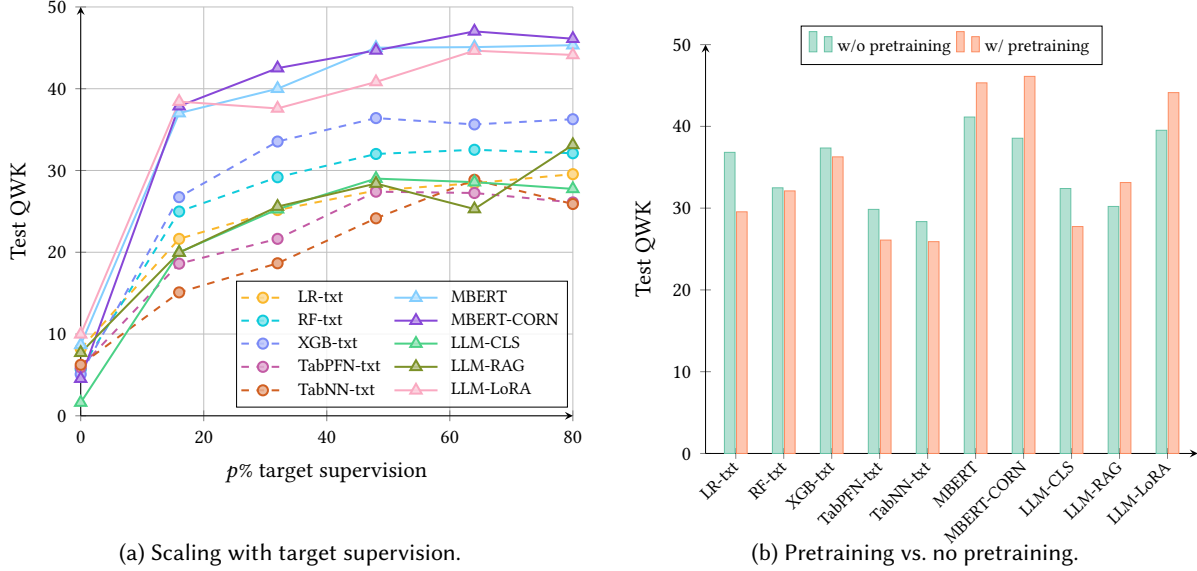


Fig. 8. (a) Mean test QWK over the twelve target states as a function of the target supervision level p for all ten models. (b) Mean test QWK of the same ten models over the twelve states, with and without pretraining on the pooled source states. “w/ pretraining” is the $p=80$ cross-state setting. “w/o pretraining” is the within-state setting.

5.4.3 Discussion. The cross-state results lead to three conclusions. First, **zero-shot transfer did not prove viable for cross-state QR prediction** in our benchmark, with one exception. Without target supervision, the best source-trained model exceeds a QWK of 20 in only five of the twelve targets: MD (49.08), NC (28.63), OK (27.69), NE (24.13), and SC (21.66). In the other seven, it stays below 18. MD is the only target on which transfer works across models. It is also the state whose within-state attribution is dominated by accreditation flags that are tied to the rating by policy (Section 5.3). One plausible reading is that this policy link, rather than a general quality signal, is what transfers to MD. On this evidence, a state without any rated providers should not expect a usable predictor from other states’ data alone.

Second, **the amount of target supervision needed is modest**. Labeling 16% of a target state’s providers delivers 63% to 83% of the gain from full supervision for eight of the ten models, and the returns diminish after that. At this level, the mean QWK of MBERT-CORN, LLM-LoRA, and MBERT already reaches 98%, 97%, and 90% of their within-state means, although the two settings are scored on different test sets (Section 5.5). A state that rates a stratified sample of its providers can therefore bootstrap a predictor for the rest at a fraction of the assessment cost.

Third, **the recommended models reverse relative to the within-state setting**. Since the state-specific *preprocessed* schemas cannot be pooled, cross-state prediction is limited to the *raw* representation. On this representation, the finetuned language models (MBERT, MBERT-CORN, and LLM-LoRA) lead the Elo ranking

and hold the best QWK in at least 10 of 12 targets at every supervision level above zero. Within this group, MBERT-CORN is the strongest choice when ordinal agreement matters most, and MBERT is the most balanced across both metrics. Tree ensembles on sentence embeddings remain competitive in a few targets. They hold the best QWK in SC at every supervised level and in MT at two of five, but they are not the general choice. Lastly, the GA target shows the limit of the reconciled label space. Collapsing five levels onto three keeps GA’s QWK below 22 at every supervision level. A coarse label mapping alone was therefore not sufficient to bridge rating systems with different scales, and transfer across rating scales remains an open problem.

5.5 Transfer Utility Analysis

Cross-state QR prediction measures what a target state can achieve by borrowing data from other states. A complementary question concerns a state that already has a substantial labeled set. Does pretraining on other states add anything beyond training on its own data alone? To answer it, we compare the within- and cross-state settings. For each of the twelve states, the model without pretraining is the within-state model, which sees only the target state’s own providers. The model with pretraining is the same model pretrained on the pooled source states and then adapted on the target state’s full training pool ($p=80$). Both settings therefore train on 80% of the target state’s rated providers. We include both one-shot-fit and sequentially trained models, so the comparison also shows how each training regime uses source data. Georgia is included on its native three-level scale in both settings, with the pooled source states coarsened to that scale. For each model, we report the mean test QWK across the twelve states in the two settings.

5.5.1 Analysis Results. Figure 8b reports the mean test QWK of all ten models over the twelve states, with pretraining (“w/ pretraining”) and without it (“w/o pretraining”). The effect of pretraining depends on the training regime. Four of the six sequentially trained models gain from pretraining. MBERT-CORN gains the most, rising from 38.55 to 46.11 (+7.56). It is followed by LLM-LoRA from 39.52 to 44.13 (+4.61), MBERT from 41.14 to 45.32 (+4.18), and LLM-RAG from 30.21 to 33.13 (+2.92). The gain is most consistent for MBERT-CORN, LLM-LoRA, and MBERT, which improve in 11, 10, and 9 of the 12 states, while LLM-RAG improves in only 6. Part of MBERT-CORN’s gain repairs its weakest within-state results, with rises of 17.4 points in MT and 13.5 in GA. However, sequential training alone does not guarantee a gain. LLM-CLS falls from 32.40 to 27.75 (−4.65), and TabNN-txt from 28.35 to 25.90 (−2.45). For the four one-shot-fit models, which are fit on the union of the pooled source states and the target data, adding source data lowers the mean QWK. LR-txt falls from 36.82 to 29.55 (−7.27), TabPFN-txt from 29.85 to 26.10 (−3.75), XGB-txt from 37.35 to 36.27 (−1.08), and RF-txt from 32.49 to 32.11 (−0.38). Overall, MBERT-CORN, MBERT, and LLM-LoRA hold the three highest mean QWK values after pretraining.

5.5.2 Discussion. The pretraining comparison shows that whether cross-state data help depends less on their availability than on how they are used. The clearest benefit appears for the three language models whose backbone is finetuned: MBERT, MBERT-CORN, and LLM-LoRA. For them, pretraining on the pooled source states before adapting on the target raises the mean QWK by more than four points, even when the target’s full training pool is available. In this setup, **cross-state data add value beyond a state’s own labels**, and the gain holds in most states. For the frozen-backbone variants, the result is mixed. LLM-RAG gains 2.92 points, but LLM-CLS, which differs from it only in lacking the exemplars, loses 4.65. One hypothesis is that LLM-RAG’s head learns to compare an input with the per-class exemplars. This comparison may transfer across states more readily than the fixed class boundaries that LLM-CLS learns on the source. TabNN-txt, the only sequentially trained tabular classifier, also loses 2.45 points. In contrast, for the one-shot-fit models, **mixing the pooled source into a single training set lowers performance**. Without a mechanism to prioritize the target distribution, the larger and differently distributed source data appear to dominate the fit. For states that want to draw on other states’ ratings,

we therefore recommend the sequential pretrain-then-adapt framework with a finetuned language model such as MBERT, MBERT-CORN, or LLM-LoRA.

6 Conclusion

We presented *CCQ*, a de-identified dataset of 59,372 child care providers across 12 U.S. states. *CCQ* comes in two releases: a *preprocessed* release for tabular models and a *raw* release for language models. We also release the LLM-based curation pipeline that produced both. Using *CCQ*, we benchmarked tabular and textual classifiers on within- and cross-state QR prediction. In within-state QR prediction, tabular classifiers on the *preprocessed* representation perform best. XGB-num ranks first, and TabPFN-num ranks second without any tuned parameters. However, agreement with true ratings is substantial in only a few states and fair in others, so predictions should support rather than replace on-site assessment. In cross-state QR prediction, zero-shot transfer is close to chance in every target except MD. On the other hand, labeling 16% of a target state’s providers delivers most of the gain from full target supervision, and finetuned language models then perform best. Pretraining on other states also improves these models, even when the target state’s full training pool is available. Lastly, feature attribution associates higher predicted ratings with larger, center-based, publicly funded, and accredited providers, and lower ones with rule violations and enforcement actions. We hope *CCQ* will support future work on transfer across rating scales, state-specific adaptation, and the drivers of child care quality.

6.1 Limitations

Several limitations of the dataset and benchmark should be kept in mind. On the dataset side, our preprocessing is one of several reasonable ways to structure the data. The curation agents were also evaluated only on Georgia, the reference state (Section 3.2.2), so their measured accuracy may not carry over to the other eleven states. Since the cleaning scripts for these states were not verified line by line, residual cleaning errors can exist. On the benchmark side, every method runs under a single, uniform configuration, with no per-state tuning and a single random seed. The reported numbers are therefore reference baselines rather than the best attainable performance of any model. Cross-state results are single point estimates on a fixed 20% test set of the target state. For the smallest state, MT, with 181 rated providers in total, this test set holds only a few dozen providers. Cross-state results for the GA target also depend on one fixed mapping from five rating levels to three. Furthermore, the feature attributions describe what the explained model relies on, not causal effects on quality. Some high-attribution features, such as accreditation in MD and subsidy acceptance in OK, are themselves tied to state policy. Part of their attribution may therefore reflect administrative requirements linked to QR scores rather than an independent quality signal. Lastly, geographic identifiers likely act as proxies for regional factors that the released features do not capture.

6.2 Ethical Considerations

This study was approved by the Emory University Institutional Review Board (Protocol #: 2025P012045). The dataset contains only aggregated, facility-level information for each child care provider and no individual-level data. To further protect human subjects, we remove or anonymize every feature that could directly identify a specific provider. This minimizes the risk of re-identification and confidentiality breaches. All data handling and analysis followed the approved IRB protocol and relevant ethical guidelines.

We ask users of *CCQ* to follow three guidelines. First, users should not attempt to re-identify providers, for example by linking released records back to the state portals. Second, predicted ratings are research estimates, not official ratings, and should not inform licensing, funding, or enforcement decisions about individual providers. Third, the records reflect the state portals at the time of collection, and ratings may have changed since.

Acknowledgments

This research was partially supported by the 2025-2026 RISE Pilot Grant Program from the *Center for Children’s Health Assessment, Research Translation, and Combating Environmental Risk* (US National Institute of Environmental Health Sciences 1P2CES033430-01).

References

- [1] Administration for Children and Families, U.S. Department of Health and Human Services. 2016. Child Care and Development Fund (CCDF) Program. Final rule, Federal Register, vol. 81, no. 190, pp. 67438–67595 (45 CFR Part 98). <https://www.federalregister.gov/documents/2016/09/30/2016-22986/child-care-and-development-fund-ccdf-program>
- [2] Nikki Aikens, Charles Bush, Philip Gleason, Lizabeth Malone, and Louisa Tarullo. 2016. *Tracking Quality in Head Start Classrooms: FACES 2006 to FACES 2014 (Technical Report)*. OPRE Report 2016-95. Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, Washington, DC.
- [3] American Academy of Pediatrics, American Public Health Association, and National Resource Center for Health and Safety in Child Care and Early Education. 2019. *Caring for Our Children: National Health and Safety Performance Standards; Guidelines for Early Care and Education Programs* (4th ed.). American Academy of Pediatrics, Itasca, IL. <https://nrckids.org/CFOC>
- [4] Anthropic. 2025. *System Card: Claude Opus 4 & Claude Sonnet 4*. System Card. Anthropic. <https://anthropic.com>
- [5] Sercan Ö. Arık and Tomas Pfister. 2021. TabNet: Attentive Interpretable Tabular Learning. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, Vol. 35. Association for the Advancement of Artificial Intelligence, 6679–6687. doi:10.1609/aaai.v35i8.16826
- [6] Australian Children’s Education and Care Quality Authority. 2026. *National Registers of Approved Education and Care Services*. Bulk CSV download, CC BY 3.0 AU; accessed Aug 7, 2026. <https://www.acecqa.gov.au/resources/national-registers>
- [7] Australian Children’s Education and Care Quality Authority. 2026. *NQF Snapshot Q2 2026*. A quarterly report from the Australian Children’s Education and Care Quality Authority. Australian Children’s Education and Care Quality Authority, Darlinghurst, NSW. <https://www.acecqa.gov.au/sites/default/files/2026-08/NQF%20Snapshot%20Q2%202026%20FINAL.pdf>
- [8] Badan Akreditasi Nasional Pendidikan Anak Usia Dini, Pendidikan Dasar, dan Pendidikan Menengah. 2026. *Data Akreditasi Satuan Pendidikan*. Searchable portal; accessed Oct 3, 2026. <https://ban-pdm.id/data-akreditasi-sekolah>
- [9] Daphna Bassok, Anna J. Markowitz, Daniel Player, and Michelle Zagardo. 2018. Are Parents’ Ratings and Satisfaction With Preschools Related to Program Features? *AERA Open* 4, 1 (2018), 1–17. doi:10.1177/2332858418759954
- [10] Carol J. Blaisdell, Christina Park, Manjit Hanspal, Mary Roary, S. Sonia Arteaga, Susan Laessig, Erin Luetkemeier, Matthew W. Gillman, and on behalf of program collaborators for Environmental influences on Child Health Outcomes. 2022. The NIH ECHO Program: investigating how early environmental influences affect child health. *Pediatric Research* 92, 5 (2022), 1215–1216. doi:10.1038/s41390-021-01574-8
- [11] Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2024. Elo Uncovered: Robustness and Best Practices in Language Model Evaluation. In *Advances in Neural Information Processing Systems*, Vol. 37. 106135–106161.
- [12] Leo Breiman. 2001. Random Forests. *Machine Learning* 45 (2001), 5–32.
- [13] Kay H. Brodersen, Cheng Soon Ong, Klaas E. Stephan, and Joachim M. Buhmann. 2010. The balanced accuracy and its posterior distribution. In *2010 International Conference on Pattern Recognition*. IEEE, 3121–3124. doi:10.1109/ICPR.2010.764
- [14] Vincent Char, Linda J. Harrison, and Hui Li. 2023. Macro-structural predictors of Australian family day care quality. *Frontiers in Public Health* 11 (2023), 1114256. doi:10.3389/fpubh.2023.1114256
- [15] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, San Francisco, CA, USA, 785–794. doi:10.1145/2939672.2939785
- [16] Jacob Cohen. 1968. Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin* 70, 4 (Oct. 1968), 213–220. doi:10.1037/h0026256
- [17] Colorado Dept of Early Childhood. 2026. *Colorado Shines*. Online; accessed Jun 7, 2026. <https://www.coloradoshines.com/>
- [18] Jiashan Cui and Luke Natzke. 2021. *Early Childhood Program Participation: 2019*. First Look NCES 2020-075REV. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, DC. <https://nces.ed.gov/pubs2020/2020075REV.pdf>
- [19] Gillian Doherty, Barry Forer, Donna S. Lero, Hillel Goelman, and Annette LaGrange. 2006. Predictors of quality in family child care. *Early Childhood Research Quarterly* 21, 3 (2006), 296–312. doi:10.1016/j.ecresq.2006.07.006
- [20] Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. 2025. TabArena: A Living Benchmark for Machine Learning on Tabular Data. In *Advances in Neural Information Processing Systems (NeurIPS 2025), Datasets and Benchmarks Track*, Vol. 38. arXiv:2506.16791

- [21] Jacob A. Esplin, Brionne G. Neilson, Ann M. Berghout Austin, and Alexander Fronk. 2019. Self-Report QRIS: Challenges With Validation. Author manuscript, Faculty Publications, The Aquila Digital Community, The University of Southern Mississippi. Published in *Early Education and Development* 30(6):817–834, doi:10.1080/10409289.2019.1591045. Published full text not accessed. https://aquila.usm.edu/fac_pubs/16316
- [22] Richard Fiene. 2002. *13 Indicators of Quality Child Care: Research Update*. Technical Report. Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services, Washington, DC. <https://aspe.hhs.gov/reports/13-indicators-quality-child-care-research-update>
- [23] Simson Garfinkel, Joseph Near, Aref N. Dajani, Phyllis Singer, and Barbara Guttman. 2023. *De-Identifying Government Datasets: Techniques and Governance*. NIST Special Publication NIST SP 800-188. National Institute of Standards and Technology. doi:10.6028/NIST.SP.800-188
- [24] Georgia Dept of Early Care and Learning. 2026. *DECAL FAST*. Online; accessed Feb 7, 2026. <https://families.dec.ga.gov/>
- [25] Hillel Goelman, Barry Forer, Paul Kershaw, Gillian Doherty, Donna Lero, and Annette LaGrange. 2006. Towards a predictive model of quality in Canadian child care centers. *Early Childhood Research Quarterly* 21, 3 (2006), 280–295. doi:10.1016/j.ecresq.2006.07.005
- [26] Government of Alberta. 2026. *Child Care Information [2021 to current]*. Quarterly machine-readable CSV; accessed Aug 7, 2026. <https://open.alberta.ca/dataset/childcareinformation>
- [27] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 2022. Why do tree-based models still outperform deep learning on typical tabular data?. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, Vol. 35. 507–520.
- [28] Ishaan Gulrajani and David Lopez-Paz. 2021. In Search of Lost Domain Generalization. In *International Conference on Learning Representations (ICLR)*.
- [29] Linda J. Harrison, Manjula Waniganayake, Jude Brown, Rebecca Andrews, Hui Li, Fay Hadley, Susan Irvine, Lennie Barblett, Belinda Davis, and Maria Hatzigianni. 2024. Structures and systems influencing quality improvement in Australian early childhood education and care centres. *The Australian Educational Researcher* 51 (2024), 297–319. doi:10.1007/s13384-022-00602-8
- [30] Suzanne W. Helburn and Carollee Howes. 1996. Child Care Cost and Quality. *The Future of Children* 6, 2 (1996), 62–82. doi:10.2307/1602419
- [31] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmester, and Frank Hutter. 2025. Accurate predictions on small data with a tabular foundation model. *Nature* 637, 8045 (2025), 319–326. doi:10.1038/s41586-024-08328-6
- [32] Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. Version 2. Published at ICLR 2022; published version not accessed. arXiv:2106.09685v2 [cs.CL] <https://arxiv.org/abs/2106.09685v2>
- [33] Wenhao Huang, Zhouhong Gu, Chenghao Peng, Zhixu Li, Jiaqing Liang, Yanghua Xiao, Liqian Wen, and Zulong Chen. 2024. AutoScraper: A Progressive Understanding Web Agent for Web Scraper Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2371–2389. doi:10.18653/v1/2024.emnlp-main.141
- [34] Lisa P. Jacobson, Corette B. Parker, David Cella, Daniel K. Mroczek, Barry M. Lester, and on behalf of program collaborators for Environmental influences on Child Health Outcomes. 2024. Approaches to protocol standardization and data harmonization in the ECHO-wide cohort study. *Pediatric Research* 95, 7 (2024), 1726–1733. doi:10.1038/s41390-024-03039-0
- [35] Jade Marcus Jenkins, Jennifer K. Duer, and Maia Connors. 2021. Who Participates in Quality Rating and Improvement Systems? HHS Public Access author manuscript, PMC7543882. Published in final edited form in *Early Childhood Research Quarterly* 54:219–227, doi:10.1016/j.ecresq.2020.09.005. Published full text not accessed. <https://pmc.ncbi.nlm.nih.gov/articles/PMC7543882/>
- [36] Jaskaran Singh Kawatra, Sebin Sabu, Pavithra Rajendran, Caroline Baumgartner, Avish Vijayaraghavan, Ewart Jonny Sheldon, John Booth, Neil Sebire, Shiren Patel, Alexandros Zenonos, and Rebecca Pope. 2025. Minimal Data Maximum Impact: Lessons Learned from Real-World Unstructured Data in Paediatric Care. In *AAAI Bridge Program on AI for Medicine and Healthcare (Proceedings of Machine Learning Research, Vol. 281)*. PMLR, 70–78.
- [37] Kentucky Cabinet for Health and Family Services. 2026. *Kentucky Kynect Child Care Search*. Online; accessed Jun 7, 2026. <https://kynect.ky.gov/>
- [38] Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*. arXiv:1412.6980v9 [cs.LG]
- [39] Ron Kohavi. 1995. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95)*. 1137–1143.
- [40] Miron B. Kursat and Witold R. Rudnicki. 2010. Feature Selection with the Boruta Package. *Journal of Statistical Software* 36, 11 (2010), 1–13. doi:10.18637/jss.v036.i11
- [41] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174. doi:10.2307/2529310
- [42] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*. Vancouver, Canada.

- [43] Deming Li. 2022. Machine Learning Based Preschool Education Quality Assessment System. *Mobile Information Systems* 2022, 1 (2022), 2862518. doi:10.1155/2022/2862518
- [44] Ying-Chun Lin, Diane M. Early, James Fuller, Tracy Gebhart, Brenda Miranda, and Kelly Maxwell. 2024. *Licensing Violations and Program Quality in Child Care and Early Education*. OPRE Report #2024-163. Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, Washington, DC. <https://acf.gov/opre/report/licensing-violations-and-program-quality-child-care-and-early-education>
- [45] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*. arXiv:1711.05101v3 [cs.LG]
- [46] Jiaying Lu, Stephanie R. Brown, Songyuan Liu, Shifan Zhao, Kejun Dong, Del Bold, Michael Fundora, Alaa Aljiffry, Alex Fedorov, Jocelyn Grunwell, and Xiao Hu. 2025. Early Risk Prediction of Pediatric Cardiac Arrest from Electronic Health Records via Multimodal Fused Transformer. Published at IEEE EMBC 2025, doi:10.1109/EMBC58623.2025.11251673. Published version not accessed. arXiv:2502.07158v3 [cs.LG] <https://arxiv.org/abs/2502.07158v3>
- [47] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From Local Explanations to Global Understanding with Explainable AI for Trees. HHS Public Access author manuscript, PMC7326367. Published in final edited form in *Nature Machine Intelligence* 2(1):56–67, doi:10.1038/s42256-019-0138-9. Published full text not accessed. <https://pmc.ncbi.nlm.nih.gov/articles/PMC7326367/>
- [48] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, Vol. 30. 4765–4774.
- [49] Anna J. Markowitz, Daphna Bassok, and Daniel Player. 2020. Simplifying quality rating systems in early childhood education. *Children and Youth Services Review* 112 (2020), 104947. doi:10.1016/j.childyouth.2020.104947
- [50] Maryland State Department of Education. 2026. *Maryland EXCELS: Find a Program*. Online; accessed Jun 7, 2026. <https://findaprogram.marylandexcels.org/>
- [51] Kelly Maxwell, Diane Early, Brenda Miranda, Nina Johnson, and Ying-Chun Lin. 2024. *Improving the Quality of Licensing Data*. OPRE Report #2024-123. Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, Washington, DC. <https://acf.gov/opre/report/improving-quality-licensing-data>
- [52] Naci Mocan. 2007. Can consumers detect lemons? An empirical analysis of information asymmetry in the market for child care. *Journal of Population Economics* 20, 4 (2007), 743–780. doi:10.1007/s00148-006-0087-6
- [53] Montana Department of Public Health and Human Services. 2026. *Best Beginnings STARS to Quality Program: STARS Programs List by Level*. Online; accessed Jun 7, 2026. <https://dphhs.mt.gov/assets/ecfsd/childcare/STARS/Participants/STARSPROGRAMSLISTBYLEVEL.pdf>
- [54] MyChildCarePlan. 2026. *My Child Care Plan California*. Online; accessed Jun 7, 2026. <https://mychildcareplan.org/>
- [55] Avanika Narayan, Ines Chami, Laurel Orr, and Christopher Ré. 2022. Can Foundation Models Wrangle Your Data? *Proceedings of the VLDB Endowment* 16, 4 (2022), 738–746. doi:10.14778/3574245.3574258
- [56] National Center on Early Childhood Quality Assurance. 2019. *Program Participation in QRISs in 2017*. Quality Compendium: 2017 QRIS Fact Sheet. National Center on Early Childhood Quality Assurance, Fairfax, VA. https://childcareta.acf.hhs.gov/sites/default/files/qrisc_program_participation_2017.pdf
- [57] National Institute of Child Health and Human Development Early Child Care Research Network. 2000. The Relation of Child Care to Cognitive and Language Development. *Child Development* 71, 4 (July/August 2000), 960–980. doi:10.1111/1467-8624.00202
- [58] Nebraska Department of Education. 2026. *Step Up to Quality: Find a Provider*. Online; accessed Jun 7, 2026. <https://stepuptoquality.ne.gov/>
- [59] Noripansyah, Abdul Kadir, Dewi Kusumaningsih, and Haderiansyah. 2023. Accreditation Prediction of Early Childhood Education Institutions Using Machine Learning Techniques. *Jurnal Teknik Informatika (JUTIF)* 4, 3 (June 2023), 647–655. doi:10.52436/1.jutif.2023.4.3.999
- [60] North Carolina Division of Child Development and Early Education. 2026. *NC Child Care Search*. Online; accessed Jun 7, 2026. <https://ncchildcare.ncdhhs.gov/>
- [61] Oklahoma Department of Human Services. 2026. *OKDHS Child Care Locator*. Online; accessed Jun 7, 2026. <https://ccl.dhs.ok.gov/>
- [62] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. Vancouver, Canada.
- [63] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (Oct. 2011), 2825–2830.
- [64] Ellen S. Peisner-Feinberg, Margaret R. Burchinal, Richard M. Clifford, Mary L. Culkin, Carolee Howes, Sharon Lynn Kagan, and Noreen Yazejian. 2001. The Relation of Preschool Child-Care Quality to Children's Cognitive and Social Developmental Trajectories through Second Grade. *Child Development* 72, 5 (2001), 1534–1553. doi:10.1111/1467-8624.00364

- [65] Ben Peters, Vlad Niculae, and André F. T. Martins. 2019. Sparse Sequence-to-Sequence Models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1504–1519.
- [66] Pusat Data dan Teknologi Informasi. 2026. *Statistik Pendidikan Anak Usia Dini 2025/2026 (Statistics of Early Childhood Education 2025/2026)*. Statistical publication Volume 7. Sekretariat Jenderal, Kementerian Pendidikan Dasar dan Menengah. ISSN 3123-5824; data from Dapodik as of 30 November 2025. <https://data.kemendikdasmen.go.id/publikasi/p/buku-statistik/statistik-pendidikan-anak-usia-dini-paud-tahun-2025-2026-2026-paud-sederajat>
- [67] Qwen Team. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388* (2025).
- [68] Alibaba Group Qwen Team. 2025. Qwen3-Coder-30B-A3B-Instruct. <https://huggingface.co/Qwen/Qwen3-Coder-30B-A3B-Instruct>
- [69] Nils Reimers and Iryna Gurevych. 2019. 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*. Association for Computational Linguistics, Hong Kong, China, 3982–3992.
- [70] Xintong Shi, Wenzhi Cao, and Sebastian Raschka. 2023. Deep Neural Networks for Rank-Consistent Ordinal Regression Based On Conditional Probabilities. Published in *Pattern Analysis and Applications* 26(3):941–955, doi:10.1007/s10044-023-01181-9. Published full text not accessed. arXiv:2111.08851v5 [cs.LG] <https://arxiv.org/abs/2111.08851v5>
- [71] Iram Siraj, Steven J. Howard, Denise Kingston, Cathrine Neilsen-Hewett, Edward C. Melhuish, and Marc de Rosnay. 2019. Comparing regulatory and non-regulatory indices of early childhood education and care (ECEC) quality in the Australian early childhood sector. *The Australian Educational Researcher* 46 (2019), 365–383. doi:10.1007/s13384-019-00325-3
- [72] South Carolina Department of Social Services. 2026. *ABC Quality: Find Child Care*. Online; accessed Jun 7, 2026. <https://abcquality.org/>
- [73] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8 (2007), 25. doi:10.1186/1471-2105-8-25
- [74] The Build Initiative and Child Trends. 2019. *A Catalog and Comparison of Quality Initiatives (Quality Compendium)*. Data system; accessed Aug 7, 2026. <https://qualitycompendium.org/>
- [75] United States Department of Health and Human Services. Administration for Children and Families. Office of Planning, Research and Evaluation. 2014. Head Start Family and Child Experiences Survey (FACES), United States, 2014–2017. Public-use files; classroom quality measured with ECERS-R, CLASS, and Arnett Scale. doi:10.3886/ICPSR36643.v7
- [76] U.S. House of Representatives. 2024. H. RES. 1443: Expressing the sense of the House of Representatives that all young children and families should have access to high-quality, affordable child care and early education. <https://www.govinfo.gov/content/pkg/BILLS-118hres1443ih/xml/BILLS-118hres1443ih.xml>. House Resolution, 118th Congress, 2d Session; introduced September 12, 2024, and referred to the Committee on Education and the Workforce.
- [77] Deborah Lowe Vandell, Jay Belsky, Margaret Burchinal, Nathan Vandergrift, Laurence Steinberg, and on behalf of NICHD Early Child Care Research Network. 2010. Do Effects of Early Child Care Extend to Age 15 Years? Results From the NICHD Study of Early Child Care and Youth Development. NIH Public Access author manuscript, PMC2938040. Published in final edited form in *Child Development* 81(3):737–756, doi:10.1111/j.1467-8624.2010.01431.x. Published full text not accessed. <https://pmc.ncbi.nlm.nih.gov/articles/PMC2938040/>
- [78] Deborah Lowe Vandell and Zehra Gülseven. 2023. The Study of Early Child Care and Youth Development (SECCYD): Studying Development from Infancy to Adulthood. *Annual Review of Developmental Psychology* 5, 1 (2023), 331–354. doi:10.1146/annurev-devpsych-120621-035345
- [79] Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP Models Know Numbers? Probing Numeracy in Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 5307–5315. doi:10.18653/v1/D19-1534
- [80] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. In *Advances in Neural Information Processing Systems*, Vol. 33.
- [81] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2526–2547.
- [82] Washington State Department of Children, Youth, and Families. 2026. *Find Child Care WA*. Online; accessed Jun 7, 2026. <https://www.findchildcarewa.org/>
- [83] Marcy Whitebook, Laura M. Sakai, and Carollee Howes. 2004. Improving and Sustaining Center Quality: The Role of NAEYC Accreditation and Staff Stability. *Early Education and Development* 15, 3 (2004), 305–326. doi:10.1207/s15566935eed1503_4
- [84] Wisconsin Department of Children and Families. 2026. *Wisconsin Child Care Finder*. Online; accessed Jun 7, 2026. <https://childcarefinder.wisconsin.gov/>
- [85] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language

- Processing. In *Proceedings of the 2020 EMNLP (Systems Demonstrations)*. Association for Computational Linguistics, 38–45.
- [86] Stephen Wood and Rosemary Kendall. 2013. *Parents and the High Cost of Child Care: 2013 Report*. Technical Report. Child Care Aware of America, Arlington, VA. <https://files.eric.ed.gov/fulltext/ED559908.pdf>
- [87] John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024. SWE-agent: Agent-Computer Interfaces Enable Automated Software Engineering. In *Advances in Neural Information Processing Systems*, Vol. 37. arXiv:2405.15793