APPENDIX A
PROOFS

We prove all propositions and theorems in DATE. We give this proof in the supplementary material to make our paper self-contained.

A. Proof of Proposition 1

Proof. Consider any tuple t with $t \models e_1$. If $t \models e_2$, there exist some conjunction $C_{r_2} \in r_2$ having $t \models C_{r_2}$. Given $r_2 \vdash r_1$, for any $C_{r_2} \in r_2$, there exists conjunction $C_{r_1} \in r_1$ s.t. $C_{r_2} \in C_{r_1}$. Thus, for any $C_{r_2} \in r_2$, if $t \models C_{r_2}$, then $\exists C_{r_1} \in r_1, t \models C_{r_1}$. It follows $t \models r_1$. According to $t \models e_1$, we have $error(m, T_{r_1}) \leq \rho$, i.e., $t \models e_2$ as well. Otherwise, for $t \not\models r_2$, it naturally has $t \models e_2$. To sum up, e_1 implies e_2. □

Proposition 1 guarantees that for any tuple t, if t satisfies e_1, then t also satisfies e_2 under refined r_2 with $r_2 \vdash r_1$.

B. Proof of Proposition 2

Proof. Consider any tuple t with $t \models e_1$ and $t \models e_2$. If $t \models r_3 = r_1 \vee r_2$, there exists some conjunction $C \in r_3$ having $t \models C$. If $C \in r_1$ then $t \models r_1$. It follows $error(m, T_{r_1}) < \rho$ since $t \models e_1$. Thus, we have $t \models e_3$. Otherwise, when $C \in r_2$, following a similar proof, we also have $t \models e_3$. To conclude, e_1, e_2 imply e_3. □

C. Proof of Proposition 3

Proof. Consider any tuple t with $t \models e_1$ and $t \models e_2$. If $t \models r_3 = r_1 \vee r_2$, there exists some conjunction $C \in r_3$ such that $t \models C$. If $C \in r_1$, then $t \models r_1$. Thus, we have $t \models e_3$. Otherwise, when $C \in r_2$, following a similar proof, we also have $t \models e_3$. To conclude, e_1, e_2 imply e_3. □ □

Fusion means that for any tuple t satisfies e_1 and e_2, if the condition $r_3 = r_1 \vee r_2$, then t satisfies $e_3 : (m, \rho, r_3, T_{r_3})$. That is, Fusion is binary for combining the conditions of two DGRs sharing the same regression model. In this way, Fusion helps the reduction of the total number of CRRs for modeling data within the heterogeneous T.

D. Proof of Proposition 4

Proof. Consider any tuple t with $t \models e_1$. If $t \models r$, Line 14 in Algorithm 1 must be satisfied. Thus, we have $error(m(t)) \leq \rho_1 \leq \rho_2$. It follows $t \models e_2$. Otherwise, it is natural to reach $t \models e_2$ when $t \not\models r$. To conclude, e_1 implies e_2. □

E. Proof of Proposition 5

Proof. Consider the events ξ_d for each pre-trained MAB defined by

$$\begin{aligned}\xi_{d_1} = \{j \in \{1, ..., K\},\\ |\frac{1}{n_k}\sum_{s=1}^{n_k} X_{s,d_1} - f_{j,d_1}| \leq \frac{1}{2}\Delta_{K+1-k}\}\\ \xi_{d_2} = \{j \in \{1, ..., K\},\\ |\frac{1}{n_k}\sum_{s=1}^{n_k} X_{s,d_2} - f_{j,d_2}| \leq \frac{1}{2}\Delta_{K+1-k}\}\\ ...\\ \xi_{d_{|D|}} = \{j \in \{1, ..., K\},\\ |\frac{1}{n_k}\sum_{s=1}^{n_k} X_{s,d_{|D|}} - f_{j,d_{|D|}}| \leq \frac{1}{2}\Delta_{K+1-k}\}\end{aligned}$$

Also, consider the event ξ defined by

$$\xi = \{j \in \{1, ..., K\}, |\frac{1}{n_k}\sum_{s=1}^{n_k} X_s - f_j| \leq \frac{1}{2}\Delta_{K+1-k}\} \quad (1)$$

where f_j is defined as:

$$\begin{aligned} f_j &= \frac{\sum_{sim_j \in t, A' \in D} cos < \vec{D}, \vec{D'} > *p(t, A')}{\sum_{sim_j \in t, A' \in D} cos < \vec{D}, \vec{D'} >} \\ &= \frac{\sum_d cos < \vec{D}, \vec{d} > *p(t, d)}{\sum_d cos < \vec{D}, \vec{d} >} \\ &= \frac{\sum_d cos < \vec{D}, \vec{d} > *f_{j,d}}{\sum_d cos < \vec{D}, \vec{d} >} \end{aligned}$$

Setting $w_d = cos < \vec{D}, \vec{d} >$, we can rewrite 1 as:

$$\begin{aligned}\xi = \{1 \leq j \leq K,\\ |\frac{1}{n_k}\sum_{s=1}^{n_k} \frac{\sum_d w_d * X_{s,d}}{\sum_{d \in D} w_d} - \frac{\sum_d w_d * f_{j,d}}{\sum_{d \in D} w_d}| \leq \frac{1}{2}\Delta_{K+1-k}\}\\ = \{1 \leq j \leq K,\\ |\frac{1}{n_k}\sum_{s=1}^{n_k}\sum_d w_d * X_{s,d} - \sum_d w_d * f_{j,d}| \leq \frac{1}{2}\sum_{d \in D} w_d \Delta_{K+1-k}\}\end{aligned}$$

Suppose (12) is true. Using the absolute value inequality, for any $1 \leq j \leq K$, we have:

$$\begin{aligned} &|\frac{1}{n_k}\sum_{s=1}^{n_k}\sum_d w_d * X_{s,d} - \sum_d w_d * f_{j,d}| \\ &= |\sum_d (\frac{1}{n_k}\sum_{s=1}^{n_k} w_d * X_{s,d} - w_d * f_{j,d})| \\ &\leq \sum_d w_d * |\frac{1}{n_k}\sum_{s=1}^{n_k} X_{s,d} - f_{j,d}| \\ &\leq \frac{1}{2}\sum_{d \in D} w_d \Delta_{K+1-k} \end{aligned}$$

This means when $\xi_{d_1} \cap ... \cap \xi_{d_{|D|}}$ is true, ξ has to hold true regardless of w_d. Its contrapositive implies that when $\overline{\xi}$

is true, $\overline{\xi_{d_1}} \cup ... \cup \overline{\xi_{d_{|D|}}}$ has to hold true regardless of w_d. By full probability law, we have

$$\begin{aligned} P(\overline{\xi_{d_1}} \cup ... \cup \overline{\xi_{d_{|D|}}}) &= P(\overline{\xi_{d_1}} \cup ... \cup \overline{\xi_{d_{|D|}}}|\xi) * P(\xi) + \\ & P(\overline{\xi_{d_1}} \cup ... \cup \overline{\xi_{d_{|D|}}}|\bar{\xi}) * P(\bar{\xi}) \\ &\geq P(\overline{\xi_{d_1}} \cup ... \cup \overline{\xi_{d_{|D|}}}|\bar{\xi}) * P(\bar{\xi}) = P(\bar{\xi}) \end{aligned}$$

In algorithm 2, for each $MAB_i \in \{MAB_1, ..., MAB_n\}$ we have

$$P(\bar{\xi_d}) \leq 2K^2 exp(-\frac{\frac{n}{a} - K}{2\overline{log}K * H(i)})$$

By union bound, we come to the conclusion that

$$\begin{aligned} P(\bar{\xi}) \leq P(\overline{\xi_{d_1}} \cup ... \cup \overline{\xi_{d_{|D|}}}) &\leq \sum_{i=1}^{a} 2K^2 exp(-\frac{\frac{n}{a} - K}{2\overline{log}K * H(i)}) \\ &\leq 2aK^2 exp(-\frac{\frac{n}{a} - K}{2\overline{log}K * H(a)}) \end{aligned}$$

Thus, it suffices to show that, by probability of $2aK^2 exp(-\frac{\frac{n}{a}-K}{2\overline{log}K*H(a)})$, the algorithm does not make any error. □

F. Proof of Theorem 1

Proof. Given $ind(r)$ as the sharing probability of DGR r, we adopt a worst-case assumption: each predicate covers at most one point in T_r. Then, at most $|T_r| \cdot ind(r)$ points can produce shared DGRs, while $(1 - ind(r))|T_r|$ cannot. Therefore, $|P'|$ must exceed this latter count, i.e., $|P'| > (1-ind(r))|T_r|$. □

G. Proof of Theorem 2

Proof. Suppose that the greedy algorithm first selects a subset H_i that maximizes Δ_i. However, the local gain Δ_k for each specific distribution $\mathcal{D}_k$ is evaluated on its local partition T_m, while the overall model performance is non-additive across heterogeneous distributions. Therefore, even if H_i is a component of the global optimum, selecting the best combination from the other distributions given H_i, does not necessarily preserve optimality. To conclude, the greedy-choice property does not hold for the generated data selection. □

APPENDIX B
ADDITIONAL EXPERIMENTS

We provide a detailed discussion of our observations in the experiments. The details are as follows.

A. Comparison of all ablation baselines

We test the accuracy ranking of all compared baselines as a heatmap, as shown in Figure 8.

In Figure 8, a lighter cell color corresponds to a more significant discrepancy between the two methods. We observe that DATE significantly outperforms other baselines. It is worth noting that there are no statistical differences between DATE (w/o gen.) and DATE (w/o m.s., gen.), where $p = 0.441 > \alpha = 0.05$. This further demonstrates that the model sharing strategy can maintain accuracy while substantially accelerating runtime.

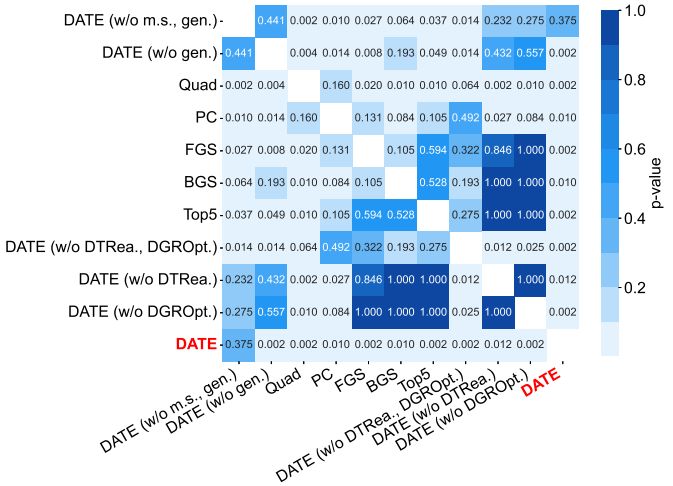

Fig. 8. Heatmap for DATE against ablation baselines in terms of accuracy ranking. The top-ranked method is highlighted in **red**. .

B. Why DATE Can Beat Finetuning-based Solutions?

We attempt to prove the effectiveness of our design of prompts and generation process, answering the question: can our prompt-based solution DATE beat finetuning-based solutions? On the one hand, traditional finetuning-based solutions often rely on prior knowledge and struggle to learn the current distribution. DATE addresses this issue by reflecting generated data with decision tree reasoning and continuing to optimize new DGRs. In this way, the generation of DATE can be closely aligned with the current distribution. On the other hand, real-world datasets do not always include clear language descriptions of the prediction task. For example, feature names and values in financial datasets are often obfuscated with arbitrary symbols to protect confidentiality [49]. In such scenarios, finetuning-based methods lose their primary advantage as they heavily rely on learning from the semantic meaning of the features. In contrast, our prompt-based DATE can achieve superior performance by carefully designing high-quality DGR-based examples for in-context learning. DATE can focus on specific distribution directly within the prompt, leading to more robust performance when semantics are absent.