ReFuGe: Feature Generation for Prediction Tasks on Relational Databases with LLM Agents

Kyungho Kim, Geon Lee, Juyeon Kim, Dongwon Choi, Shinhwan Kang, and Kijung Shin

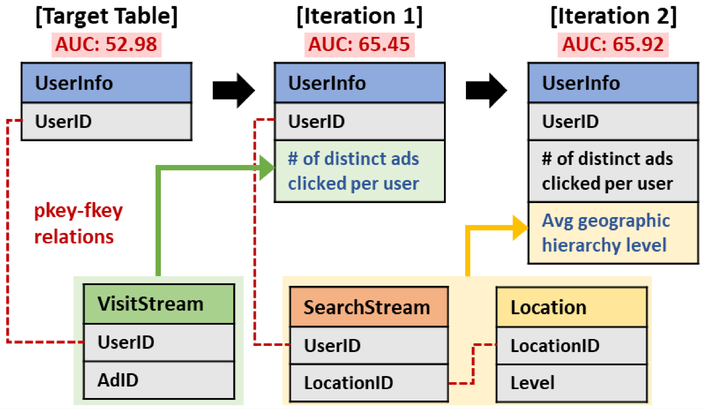
{kkyungho, geonlee0325, juyeonkim, cookie000215, shinhwan.kang, kijungs}@kaist.ac.kr

KAIST AI
Kim Jaechul Graduate School

THE WEB CONFERENCE ACM

Example: Relational Feature Generation

Prediction Task: *Predict whether each customer will click on more than one ad in the next 4 days.*



Summary

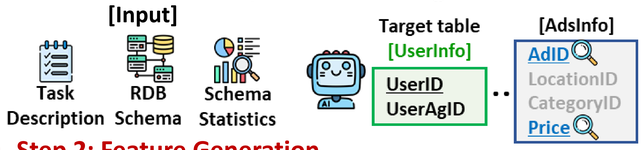
- **Goal:** to improve prediction tasks on relational databases (RDB) by generating features that aggregate predictive signals from multiple tables.
- **Challenges:**
  - **[Relational structure]** Feature generation requires understanding and extracting predictive signals from the relational structure.
  - **[Large feature space]** The space of potential features is extremely large.
  - **[No-ground-truth supervision]** Feature generation lacks ground-truth supervision, making direct supervised learning infeasible.
- **Proposed Method: ReFuGe**
  - An LLM-agent framework for relational feature generation.
  - It consists of three agents with an iterative feedback loop: **(1) schema selection agent**, **(2) feature generation agent**, and **(3) feature filtering agent.**
- **Experiments**
  - Evaluated on 7 RDB benchmark datasets and 11 prediction tasks.
  - ReFuGe achieves the best average AUROC and average rank.

Paper

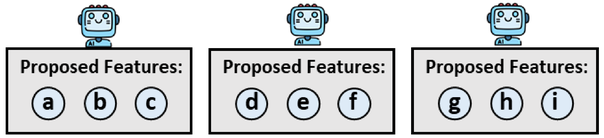
Code

Proposed Method: ReFuGe

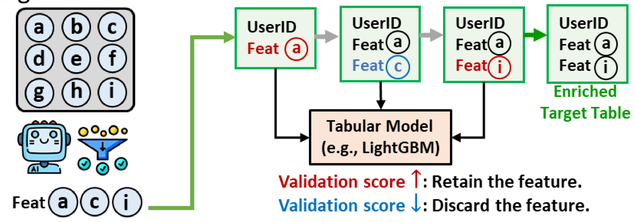
- **Step 1: Schema Selection**
  - Identifies task-relevant tables, columns, and relations.
  - Reduces the search space for feature generation.



- **Step 2: Feature Generation**
  - Generates candidate features ***by aggregating information across selected tables*** and columns from step 1.
  - Uses multiple LLM instances to produce diverse candidate features.



- **Step 3: Feature Filtering**
  - Retains promising features through two-stage filtering:
  - **[Sub-step 1] Reasoning-based filtering:** selects features likely to be useful based on the task and schema context.
  - **[Sub-step 2] Validation-based filtering:** keeps features that improve validation performance when added to the target table.



Reasoning-based filtering ***Validation-based filtering***

- **Iterative feedback loop**
  - ReFuGe repeats the above step until no additional features are selected.
  - Feedback from validation results is passed to the agents to guide the next iteration.

Experimental Results

- **RQ1. Accuracy:** Ours achieves the best average AUROC and the best average rank.

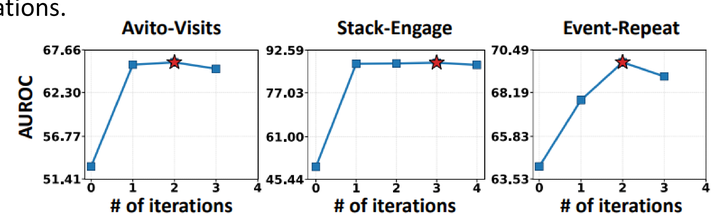
	Data & Task →	rel-avito		rel-stack		rel-f1		rel-event		rel-trial	rel-hm	rel-amazon	Average	
	Method ↓	visits	clicks	engage	badge	dnf	top3	repeat	ignore	outcome	churn	churn	AUROC	Rank
Single Table	LightGBM [12]	53.05	53.60	63.39	63.43	68.56	73.92	68.04	79.93	70.09	55.21	52.22	63.77	6.8
	XGBoost [4]	53.67	54.59	54.78	61.98	65.32	70.08	56.85	74.25	71.21	55.46	59.06	61.57	7.5
	FeatLLM [9]	51.15	48.69	42.56	-	70.43	52.72	46.67	47.88	57.35	-	55.11	52.51	8.7
RDB	ICL [20]	60.28	61.32	81.01	71.13	65.81	88.47	76.38	78.55	55.72	64.34	60.56	69.42	4.7
	Rel-LLM [19]	56.17	62.28	69.46	62.12	71.84	70.64	68.12	61.32	59.02	55.95	60.07	63.36	6.0
	RT-Gemma [16]	62.70	59.80	78.00	80.00	75.80	**91.40**	-	-	57.20	48.70	50.50	67.12	5.9
	GPT [2]	65.21	58.04	83.82	80.21	65.16	77.04	71.26	76.41	**72.94**	55.50	63.72	69.94	4.3
	Claude [3]	64.90	60.90	80.36	80.37	54.99	79.74	66.39	76.20	68.27	67.64	64.66	69.49	4.7
	LLM-CoT [13]	64.89	60.19	82.30	80.97	61.41	75.92	70.13	82.58	68.03	68.06	64.42	70.81	4.2
	ReFuGe (Ours)	**66.43**	**63.20**	**87.66**	**83.80**	**76.26**	83.04	**77.14**	**83.40**	72.22	**68.10**	**67.00**	**75.30**	**1.3**

- **RQ2. Ablation study:** Removing each key component degrades the overall performance.
  - **ReFuGe-SS**: Generates features directly from the full RDB schema without schema selection.
  - **ReFuGe-FF**: Replaces reasoning-based feature filtering with random feature selection.
  - **ReFuGe-FB**: Removes feedback from previous iterations during feature generation.

Data & Task →	rel-avito		rel-stack		rel-f1		rel-event		rel-trial	rel-hm	rel-amazon	Average	
Method ↓	visits	clicks	engage	badge	dnf	top3	repeat	ignore	outcome	churn	churn	AUROC	Rank
ReFuGe-SS	**66.61**	60.15	82.70	**83.89**	73.37	79.46	73.34	**88.36**	72.19	68.04	63.40	73.77	2.3
ReFuGe-FF	66.56	58.25	**88.51**	82.17	62.42	62.20	70.62	82.47	71.00	68.04	63.94	70.56	3.0
ReFuGe-FB	66.43	61.21	87.61	63.41	62.40	81.32	68.00	83.35	71.65	67.92	66.39	70.88	3.1
ReFuGe	66.43	**63.20**	87.66	83.80	**76.26**	**83.04**	**77.14**	83.40	**72.22**	**68.10**	**67.00**	**75.30**	**1.5**

***ReFuGe-FF** shows the largest performance drop, indicating the importance of the feature filtering step.*

- **RQ3. Effect of iterative feedback:** ReFuGe generally improves validation performance over iterations.



*On average, **ReFuGe** performs **2.4 iterations** across all tasks*