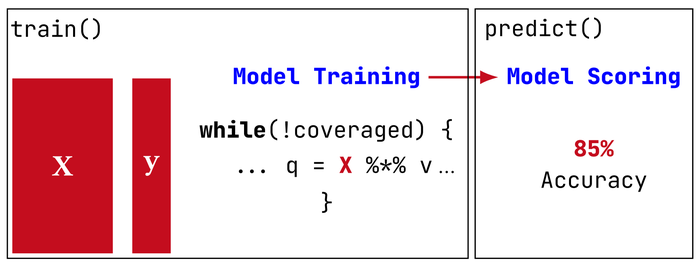
CatDB: Data-catalog-guided, LLM-based Generation of Data-centric ML Pipelines

Saeed Fathollahzadeh [1] Essam Mansour [1] Matthias Boehm [2,3]

1 Concordia University
2 Technische Universität Berlin
3 BIFOLD

1. Motivation

Data Preparation
Data Integration & Data Cleaning
Data Programming & Augmentation
Model and Feature Selection
Hyperparameter Tuning + CV



- Data-centric ML pipelines are crucial ➜ labor-intensive.
- AutoML systems ➜ struggle with large datasets.
- LLMs demonstrate strong capabilities in coding ➜ struggle on unseen data.
- LLM-based pipeline generation ➜ lacks tailored dataset context.

2. Data-centric ML pipelines w/ [Metadata-only vs. CatDB] & LLM

Salary dataset: comma-separated CSV file
Target feature: "**Salary**"

Experience	Skills	Gender	Address	Salary
12 Months	SQL,Java	F	7050 CA	100
two years	JavaScript	Female	TX 7871	150
36 months	C/C++,.Net	M	Texas	300
3 Years	JS,CPP,SQL	Female	7871	310
one Year	Python	Male	CA	200
2 Years	C#,Java	0	TX	175

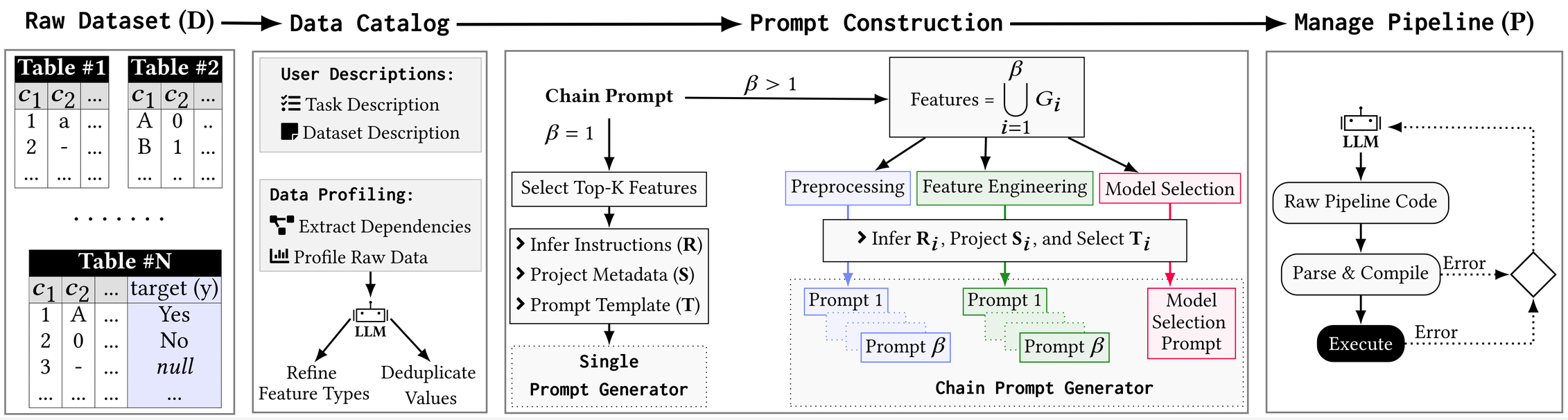
Pipeline DAG with Metadata & LLM

❶ Load Dataset
❷ Hash Categorical Feature
Gender: {M, F, Male, Female}
❸ Train Model
Accuracy = 39.2%

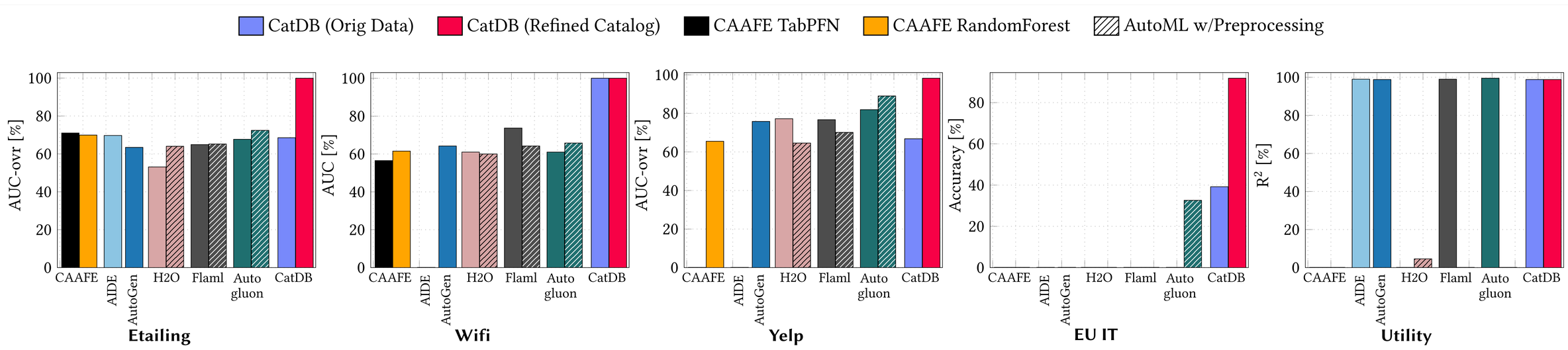
Pipeline DAG with CatDB & LLM

❶ Load Dataset
❷ Deduplicate Categorical Values
Experience: Replace(12 Months:1 year, ..)
Skills: Replace(CPP:C++, .Net:C#, ..)
Gender: Replace(F:Female, M:Male)
❸ Decompose Features
Address: Split to State and ZipCode
❹ Infer Feature Type
Skills: Set as List Type
❺ Feature Selection
Address: Keep ZipCode & Remove State
❻ Feature Hashing
❼ Train Model
Accuracy = 91.8%

3. Data-centric ML pipeline generation in CatDB



4. Performance Comparison (LLM = Gemini-1.5)

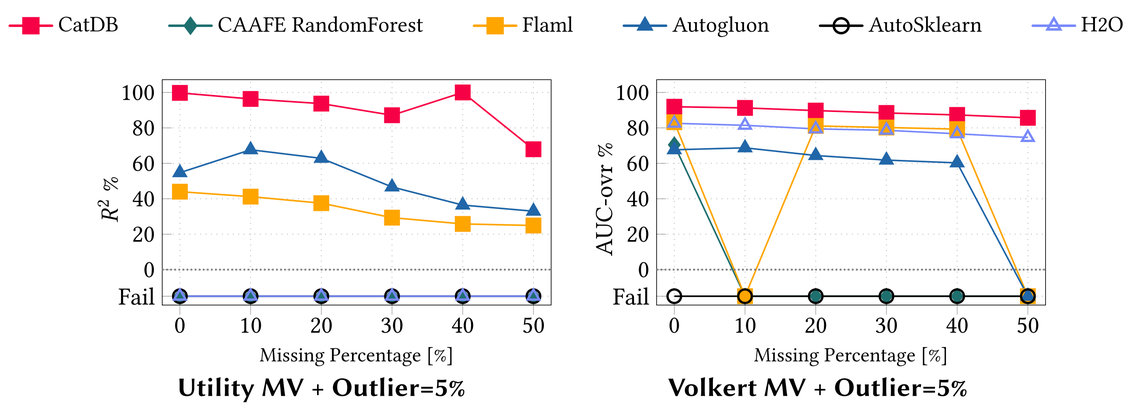


Categorical Features:
› Fixes formatting, transforms complexity.

Deduplication:
› Removes duplicates, balances labels.

Feature Refinement:
› Drops constant/misread features, preserves distribution.

5. Outlier and Missing Value Injection (Gemini-1.5)



Robustness ➜ CatDB maintains high performance even with increasing missing values and 5% outliers.

6. Conclusions

- **Data Catalog Integration** ➜ Use metadata & rules for tailored pipelines.
- **Catalog Refinements** ➜ Enhance catalogs to guide ML pipeline creation.
- **Prompt Chaining** ➜ Sequence prompts to optimize generation.
- **Error Handling** ➜ Validate, fix with knowledge base for reliable pipelines.

Paper

Reproducibility

saeed.fathollahzadeh@concordia.ca