

Eureka: Intelligent Feature Engineering for Enterprise AI Cloud Resource Demand Prediction

Alibaba Cloud

同濟大學
TONGJI UNIVERSITY

復旦大學
FUDAN UNIVERSITY

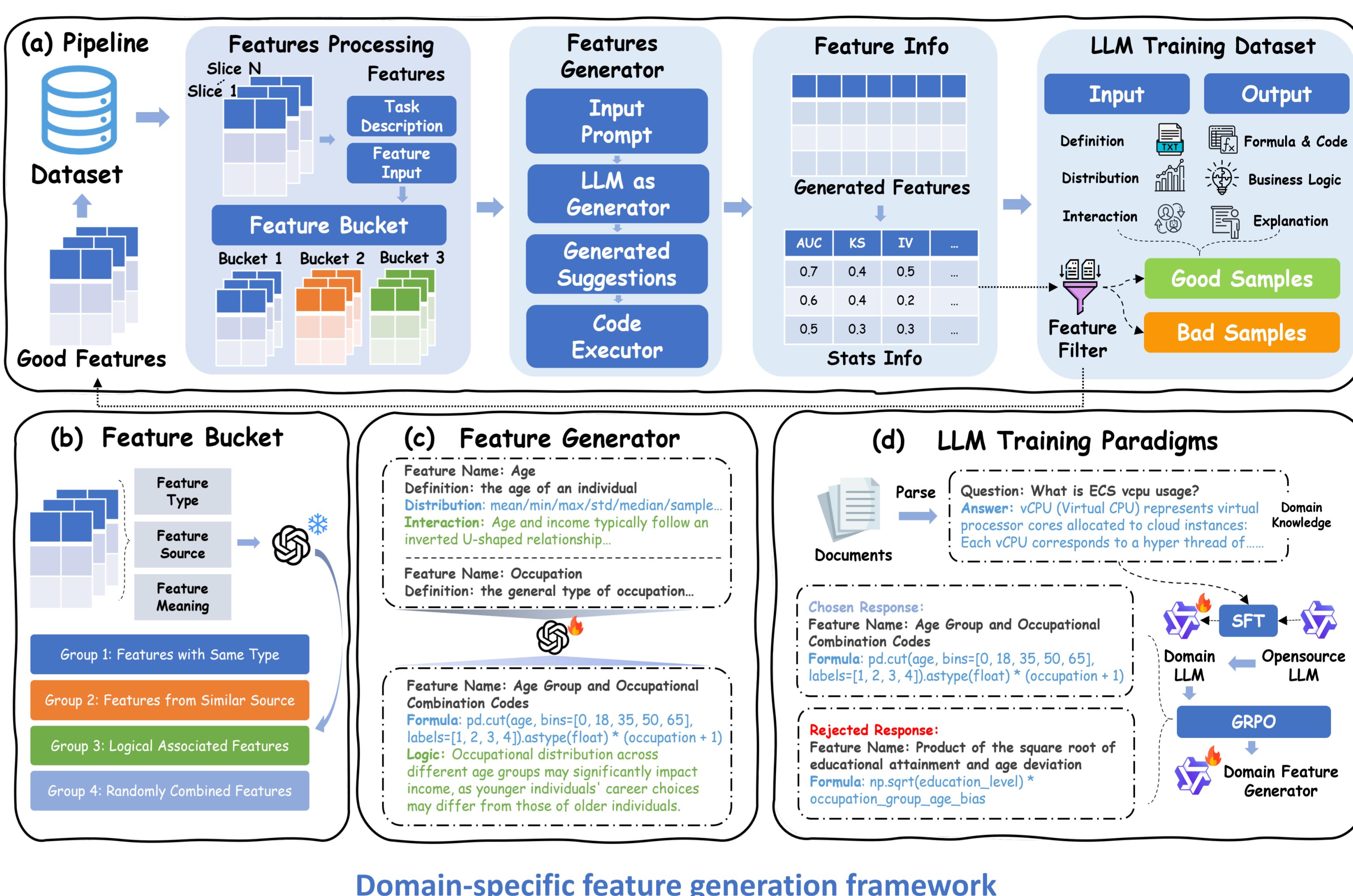
NEURAL INFORMATION PROCESSING SYSTEMS

Hangxuan Li, Renjun Jia, Xuezhang Wu*, Yunjie Qian, Zeqi Zheng, Xianling Zhang

What is Eureka?

Eureka is an LLM-driven agentic framework that automates feature engineering and has three key components:

- **Eureka Expert:** Guides feature exploration using domain knowledge.
- **LLM Feature Factory:** Generates executable features from plans.
- **Self-Evolving Engine:** Improves features with deployment feedback.

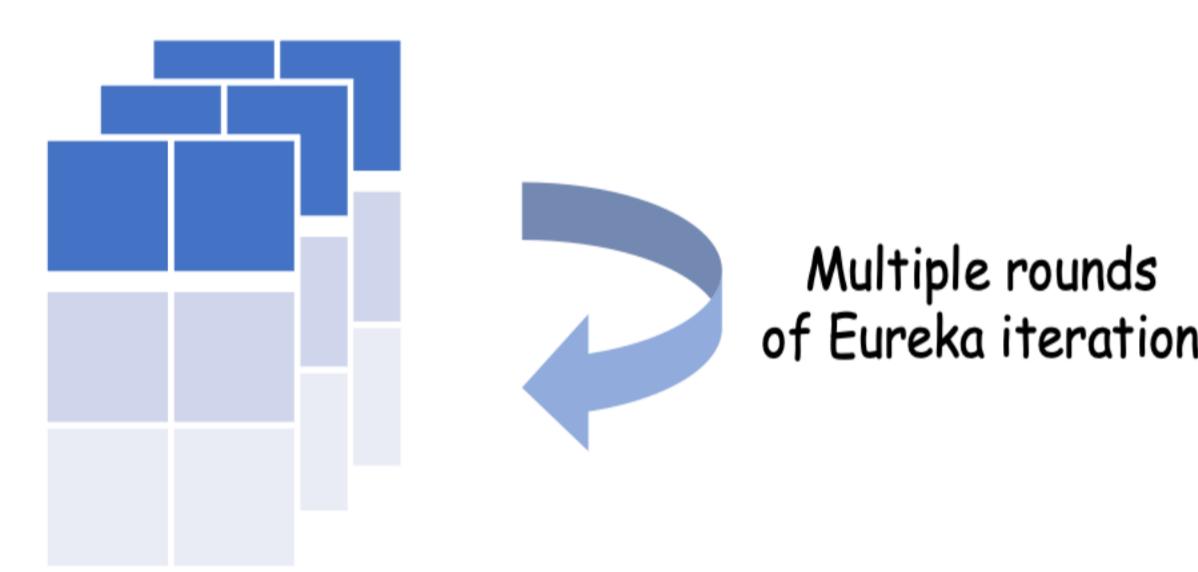


Eureka Expert

Domain-specific feature generation framework

How does Eureka work?

Feature Bucket



storage_lag_3m
gpu_float_range
gcard_range
gpu_float_lag_1m

gpu_volatility_interaction
gpu_float_anomaly_score

scaled_vcpu_gpu_ratio
gpu_anomaly_acceleration_squared

original feature
original feature
original feature
original feature

Iteration 1
Iteration 1

Iteration 2
Iteration 2

Eureka Expert

Thinking Process:

Ori Feature Thinking:
scaled_vcpu_gpu_ratio: The statistical profile of this feature indicates a pronounced right-skewed distribution. This distribution pattern reveals that for the vast majority of samples, the ratio of vCPU growth to GPU volatility is small and stable. ...

New Feature Thinking:
The business logic dictates that high GPU volatility without a corresponding resource imbalance (i.e., "unexplained" volatility) is a strong indicator of non-scheduling anomalies (e.g., hardware failures). By combining these two features, we can effectively isolate this pattern of "anomalous volatility unexplained by resource allocation" to significantly improve predictive power.

LLM Factory

New Feature Des:

Defination: scaled_vcpu_gpu_volatility_ratio is a ratio feature that contrasts resource imbalance risk with anomaly acceleration. A high value signifies a "latent risk"—a major structural misalignment occurring under dynamically stable conditions—which is a key predictor for an impending GPU demand burst.

Formula:
scaled_vcpu_gpu_ratio / (gpu_anomaly_acceleration_squared + 1e-6)

New Feature:
scaled_vcpu_gpu_volatility_ratio

AUC: 0.715
KS: 0.335
IV: 0.782

Business Logic: Experts have assessed that it possesses strong business logic.

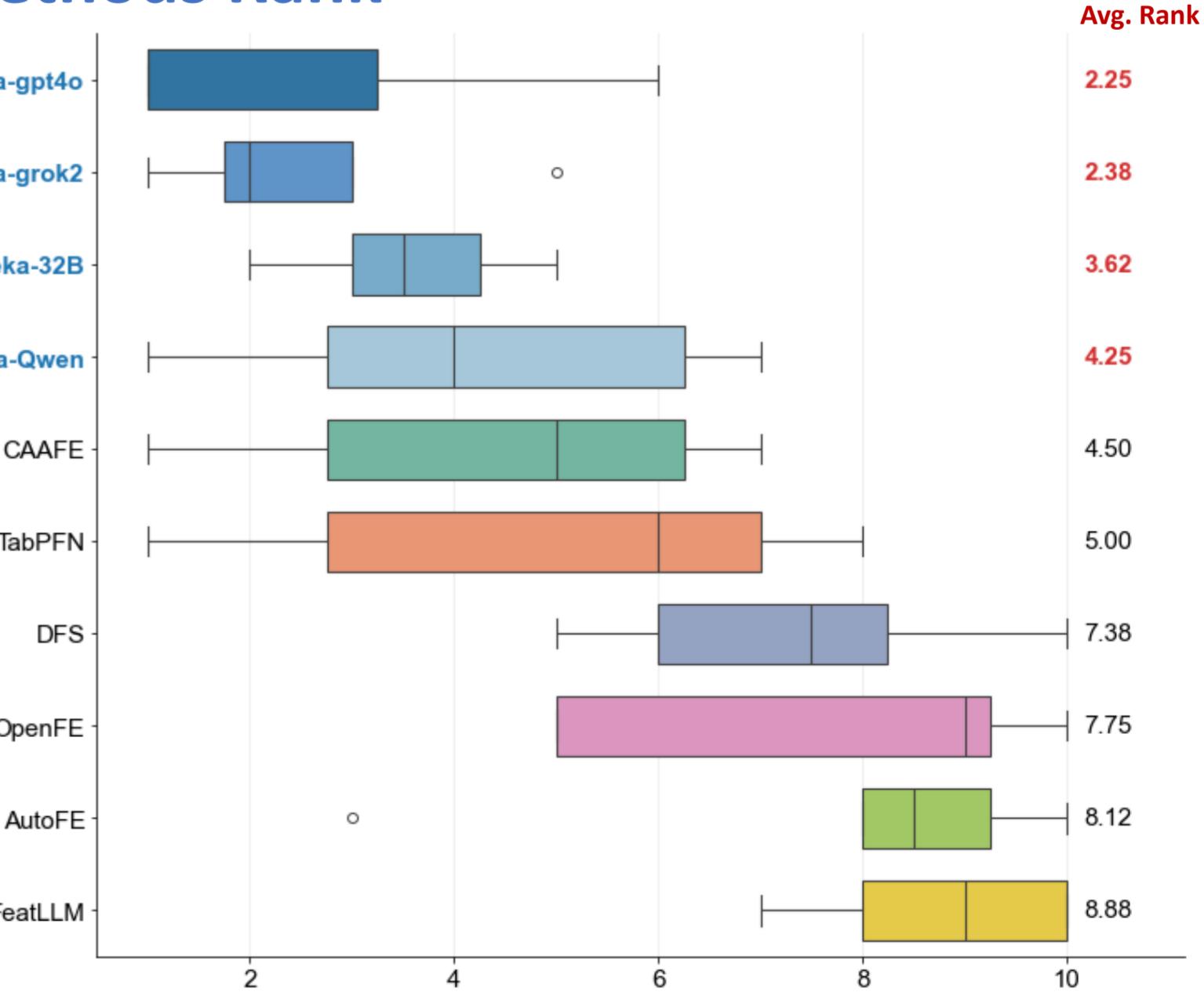
The process of obtaining high-quality features in EGS scenarios

Experimental Results

Eureka performs strongly across multiple datasets:

Method	Adult	Bank	Blood	Credit	Diabetes	Heart	Myocardial	EGS
DFS	0.915	0.897	0.637	0.707	0.882	0.921	0.654	0.68
AutoFE	0.872	0.865	0.735	0.676	0.837	0.903	0.674	0.666
OpenFE	0.92	0.923	0.626	0.701	0.817	0.897	0.697	0.658
TabPFN	0.9	0.904	0.715	0.787	0.888	0.938	0.676	0.694
CAAFE	0.901	0.905	0.713	0.797	0.886	0.941	0.686	0.692
FeatLLM	0.894	0.851	0.678	0.743	0.811	0.881	0.663	0.675
Eureka-Qwen	0.926	0.932	0.738	0.794	0.873	0.926	0.737	0.679
Eureka-32B	0.926	0.932	0.716	0.824	0.884	0.936	0.711	0.688
Eureka-gpt4o	0.928	0.934	0.745	0.838	0.881	0.938	0.699	0.699
Eureka-grok2	0.928	0.934	0.735	0.815	0.886	0.940	0.719	0.686

Methods Rank:



Ablation results

Quantifiable business impact

Configuration	ROC-AUC(%)	Δ vs Eureka(%)
Eureka	69.97	--
w/o Self-Evolving Engine	69.45	-0.52
w/o Eureka Expert	65.91	-4.06
w/o LLM Factory	63.84	-6.13

Impact Area	Key business metric	Improvement
Operational adoption	Adoption rate of generated warnings	91%
Demand fulfillment	Demand fulfillment rate (top-tier customers)	+16%
Resource efficiency	Reduction in server migration loss rate	-33%

Learn More

Scan for more information and contact us!

