ICLR
International Conference On Learning Representations

Human-LLM Collaborative Feature Engineering for Tabular Learning

Zhuoyan Li[1], Aditya Bansal[2], Jinzhao Li[1], Shishuang He[3], Zhuoran Lu[1], Mutian Zhang[1], Yiwei Yang[4], Qin Liu[5], Swati Jain[2], Ming Yin[1], Yunyao Li[2]

PurdueUniversity[1] , Adobe[2] , UIUC[3] , University of Washington[4], UC Davis[5]

Adobe

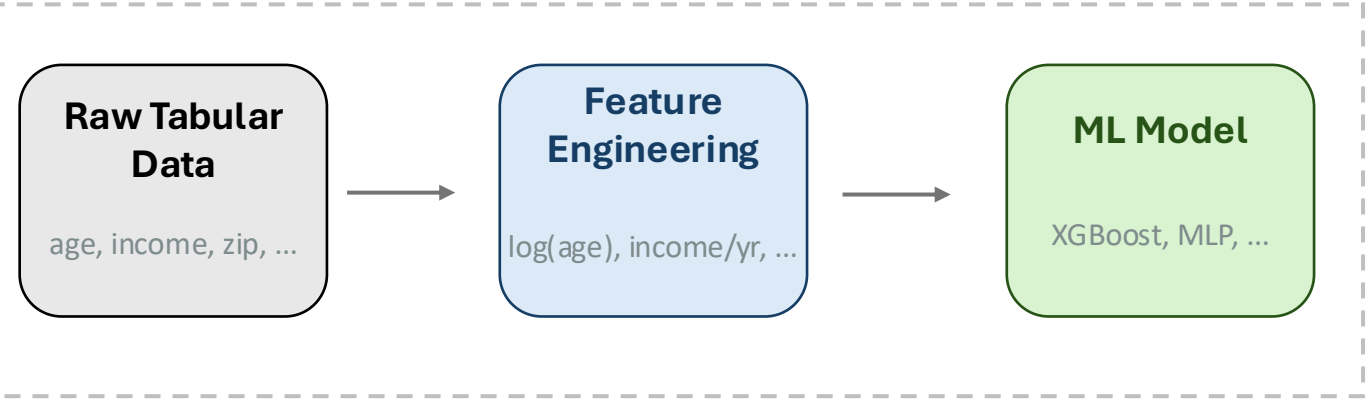
PURDUE UNIVERSITY

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

UCDAVIS UNIVERSITY OF CALIFORNIA

UNIVERSITY of WASHINGTON

Feature Engineering for Tabular Data



LLMs can propose creative feature transformations by leveraging semantic understanding of the task and data.

E.g., CAAFE (Hollmann et al., 2023), OCTree (Nam et al., 2024)

The Problem

Current methods assign the LLM both roles:

1. **Proposing** candidate features
2. **Selecting** which one to evaluate

The LLM acts as a **black-box optimizer**, relying entirely on internal heuristics for both generation and selection.

Why is this a **problem**?

1

No Calibrated Uncertainty

LLM cannot quantify how promising or uncertain a candidate is

2

Wasted Exploration

Repeated low-yield operations without principled search strategy

3

No Human Steerability

No mechanism to incorporate expert domain knowledge

→ Suboptimal feature engineering, especially when iteration budgets are limited.

Our Approach: Decouple Proposal from Selection

Previous Approaches

LLM

Propose + Select
(black box)

- Single model handles both roles
- No uncertainty quantification
- No human interaction mechanism

→

Our Framework

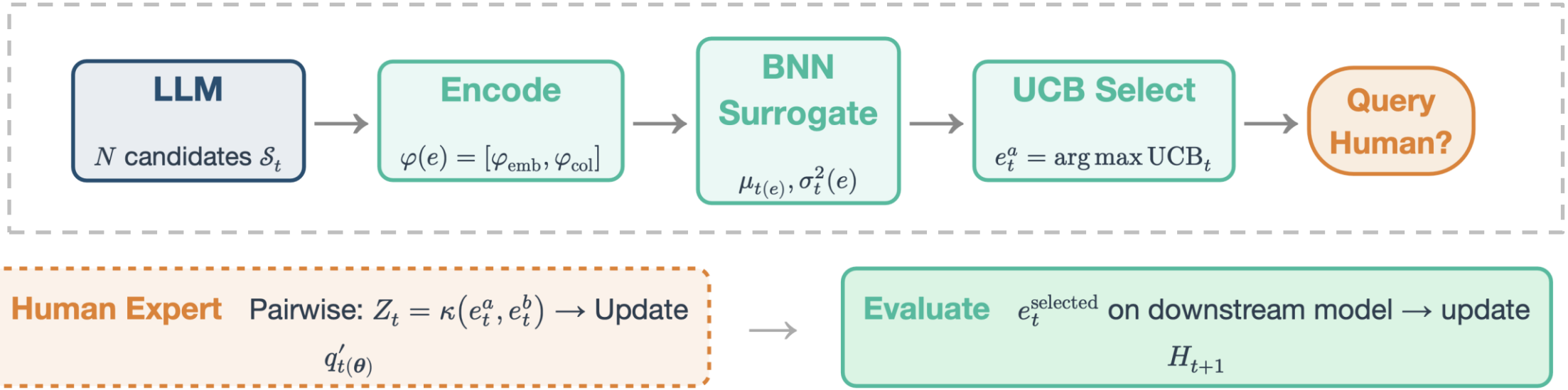
LLM
Propose

BNN+UCB
Select

Human
Selective feedback

- Principled Bayesian selection
- Calibrated uncertainty estimates
- Selective human preference queries

Framework Overview



$$\text{UCB}_{t(e)} = \mu_{t(e)} + \sqrt{\beta_t} \cdot \sigma_{t(e)}$$ Balances exploitation (high μ) with exploration (high σ)

Lemma 3.1: *With probability* $\geq 1 - \delta$, $|g(e) - \mu_{t(e)}| \leq \sqrt{\beta_t} \cdot \sigma_{t(e)}$ *for all* $e \in \mathcal{S}_t$

Surrogate Model: Encoding & BNN

Feature Operation Encoding

$$\varphi(e) = [\varphi_{\text{embedding}}(e), \varphi_{\text{column}}(e)]$$

$\varphi_{\text{embedding}}(e)$

Semantic text embedding via `text-embedding-3-small`

Captures *what* the operation does

$\varphi_{\text{column}}(e)$

Binary column-usage vector $\boldsymbol{m} \in \{0,1\}^d$

Captures *which* columns are used

Why both? Semantic embeddings alone can't distinguish `log(age)` from `log(income)` — column mask resolves this.

Bayesian Neural Network Surrogate

Why BNN over Gaussian Process?

- GP struggles with high-dimensional language-derived embeddings
- BNN provides greater scalability and expressiveness
- Naturally yields calibrated uncertainty via variational inference

Variational Inference

Posterior: $q_t(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}; \boldsymbol{M}_t, \boldsymbol{\Sigma}_t)$

Predictions for candidate e:

$\mu_t(e) = \mathbb{E}_{q_t}[\hat{g}(\varphi(e); \boldsymbol{\theta})]$ *(expected utility)*

$\sigma_t^2(e) = \mathrm{Var}_{q_t}[\hat{g}(\varphi(e); \boldsymbol{\theta})]$ *(uncertainty)*

Unlike a regular NN, BNN learns a distribution over weights → naturally provides calibrated uncertainty.

Selective Human **Preference Feedback**

Human feedback elicited **only** when both conditions hold:

C1 — Overlap $\mathrm{UCB}_{t(e_t^b)} > \mathrm{LCB}_{t(e_t^a)}$

Confidence intervals overlap — outcome uncertain

C2 — Uncertainty ≥ Cost

$$\sqrt{\beta_t}\left(\sigma_{t(e_t^a)} + \sigma_{t(e_t^b)}\right) \geq \gamma_\kappa$$

Potential gain justifies cognitive cost

Posterior update via probit model:

$$\mathcal{P}(Z_t \mid \boldsymbol{\theta}) = \Phi\left(\eta Z_t [\hat{g}(e_t^a) - \hat{g}(e_t^b)]\right)$$

When to ask the human?

No overlap → don't ask

e_t^a : `|-[===]-|`

e_t^b : `|-[===]-|`

Overlap → ask!

e_t^a : `|-[========]-|`

e_t^b : `|-[========]-|`

Overlap = feedback can change decision

Key Results: Classification (AUROC %)

Dataset	OpenFE	AutoGl.	CAAFE	OCTree	Ours (w/o)	Ours (w/)
MLP						
flight	93.3	92.6	92.9	94.8	**96.9**	**97.3**
loan	95.3	95.4	95.7	95.9	**96.0**	**96.1**
conversion	90.7	90.6	90.9	91.1	**92.6**	**92.9**
heart	92.2	92.3	92.6	93.1	**93.4**	**93.6**
XGBoost						
flight	95.7	95.4	95.2	96.4	**97.6**	**98.0**
conversion	91.2	91.9	92.1	92.4	**93.5**	**93.9**
heart	93.6	93.5	93.6	94.3	**95.1**	**94.8**

GPT-4o backbone. 18 datasets total, representative subset shown

Avg Error Reduction vs. Best Baseline

MLP

7.24% w/o Human

8.96% w/ Human

XGBoost

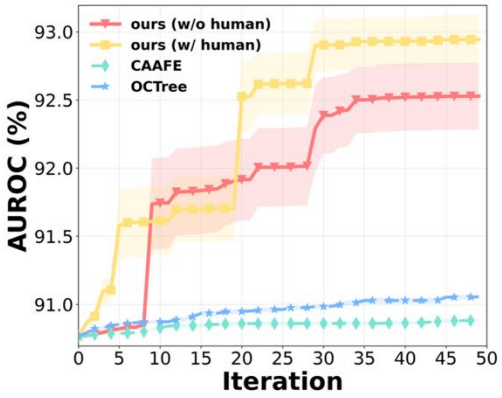
9.02% w/o Human

11.23% w/ Human

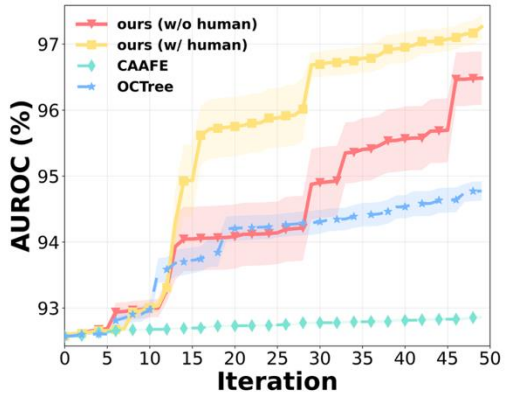
Consistent across 4 backbone LLMs
(DeepSeek-v3, GPT-3.5, GPT-4o, GPT-5)

Performance Trajectories Over Iterations

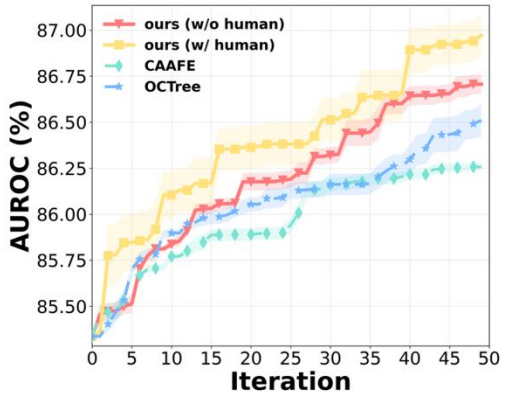
Our method (w/ and w/o human) steadily improves while baselines (CAAFE, OCTree) plateau. 50 iterations, MLP



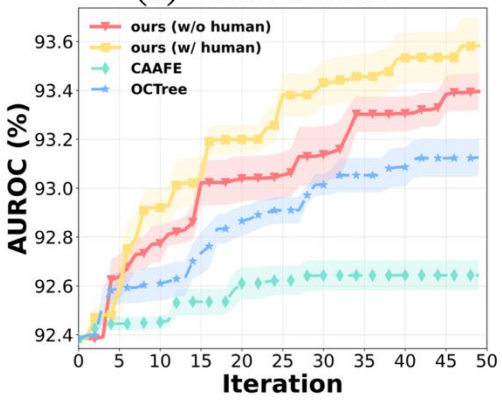
(a) Conversion



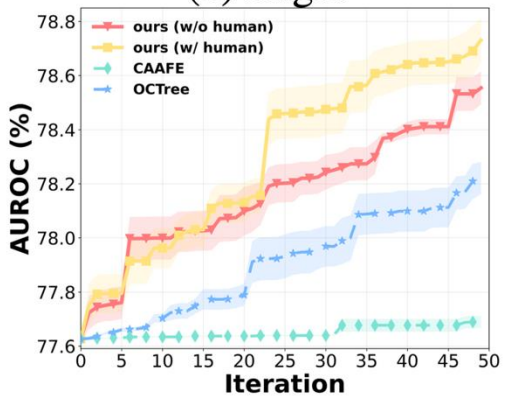
(b) Flight



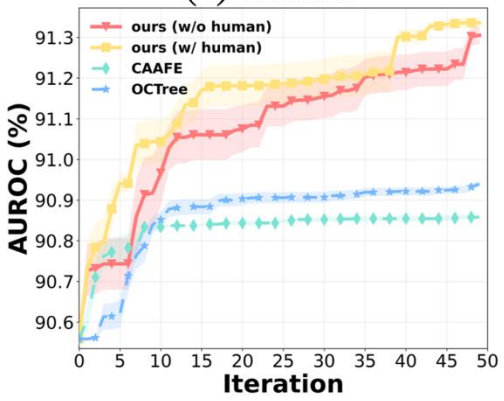
(c) Titanic



(d) Heart



(e) Wine



(f) Adult

Computational **Scalability**

Our BNN surrogate and UCB computation add minimal overhead to the LLM-based pipeline.

Runtime (s) vs. # Initial Features (10K instances)

Features	LLM	Surrogate	UCB	Eval
10	1.82	0.17	0.006	1.79
100	1.82	0.19	0.005	1.78
1,000	1.82	0.20	0.009	8.4
10,000	1.82	0.57	0.018	23.4

Runtime (s) vs. # Instances (100 features)

Samples	LLM	Surrogate	UCB	Eval
1,000	1.82	0.17	0.005	0.28
10,000	1.82	0.23	0.006	1.47
50,000	1.82	0.18	0.006	5.22
100,000	1.82	0.18	0.005	10.65

Surrogate + UCB = only 2.2% of total runtime (at 10K features). Our overhead operates at the feature-operation level and is independent of dataset size. Dominant cost is downstream model evaluation (shared across all methods).

User Study: 31 ML Practitioners

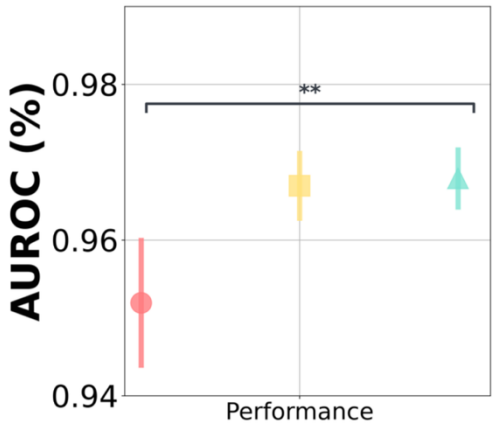
Task: Flight Satisfaction Prediction

3 conditions:

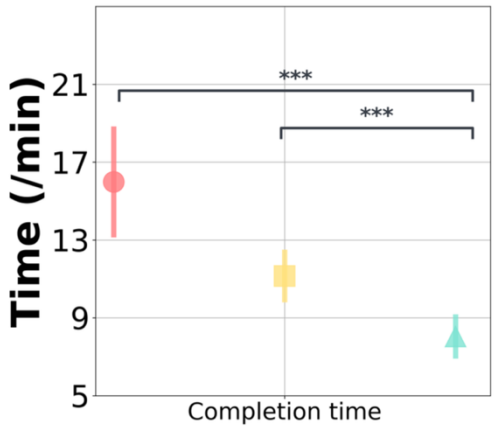
- **Control**
  Human instructs LLM
- **SR:**
  LLM auto-queries human
- **ALG (Ours)**
  Selective elicitation

ALG (Ours) achieves:

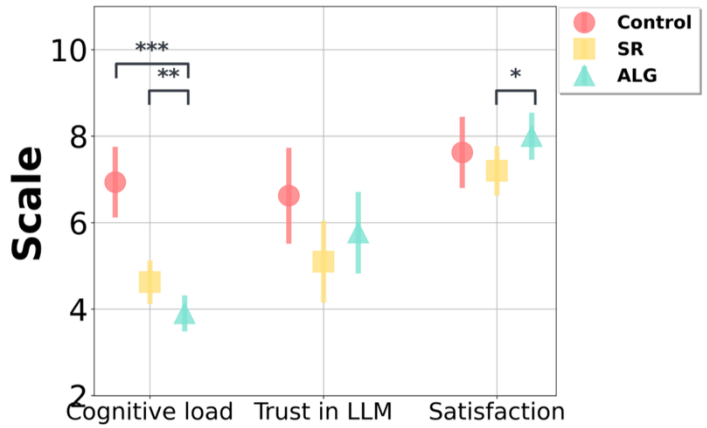
- Higher performance ($p = 0.001$)
- Faster completion ($p < 0.001$)
- Lower cognitive load ($p < 0.001$)
- Higher satisfaction ($p = 0.072$)



(a) Performance



(b) Completion time



(c) User experience

Summary

1. **Decouple** LLM feature proposal from principled Bayesian selection with BNN surrogate
2. **Selective human preference** feedback with theoretical guarantees
3. Consistent improvement across **18 datasets** and **4 backbone** LLMs
4. Real **user study** validates better performance, faster completion, lower cognitive load

Thank you!