Generalization or Memorization? Multi-Agent vs. Baseline LLMs and AutoML Models for Tabular Classification

Aida Sanatizadeh
Northern Illinois University
asanatizadeh@niu.edu

Sorouraladat Fatemi
California State University, Monterey Bay
sfatemi@csumb.edu

Reza Mousavi
University of Virginia
mousavi@virginia.edu

Ahmed Abbasi
University of Notre Dame
aabbasi@nd.edu

Abstract

Large Language Models (LLMs) are increasingly used for structured tabular data, yet it remains unclear whether their performance reflects genuine reasoning or memorization of pre-training corpora. We investigate this question through a rigorous, contamination-aware evaluation of a representative modular Multi-Agent LLM (MALLM) framework against state-of-the-art AutoML systems and established baselines (TABLET, TABLLM). We evaluate eleven binary classification tasks: five pre-cutoff benchmarks likely seen during LLM pre-training and six post-cutoff datasets released after the LLM knowledge cutoff. Results show a sharp performance dichotomy: MALLM achieves competitive or superior performance on pre-cutoff datasets but substantially underperforms AutoML on post-cutoff data, exhibiting poor calibration and high variance, especially on hard-to-classify instances. By contrast, AutoML models generalize consistently and align confidence more closely with instance hardness. These findings suggest that, despite agentic scaffolding, current LLMs cannot yet replace production-grade discriminative models for tabular classification, underscoring the need for contamination-free benchmarks to accurately assess tabular reasoning capabilities.

1 Introduction

Organizations increasingly rely on tabular (structured) data for high-stakes decisions in healthcare triage, credit risk, fraud detection, HR analytics, and operations. In these settings, analytical systems must do more than optimize average accuracy; they must also generalize under distribution shift, remain well calibrated, and behave predictably on difficult or atypical cases. The rapid adoption of large language models (LLMs) has therefore raised a critical question for both researchers and practitioners: *can LLM-based systems credibly replace production-grade AutoML frameworks for tabular analytical tasks?* Rising investment in startups and research on structured-data LLMs suggests that industry increasingly views this area as a core application (Capital, 2025; Dillet, 2025). However, despite early deployments in domains where reliability, fairness, and auditability are critical (Caruana et al., 2015; Rudin, 2019), the empirical evidence remains fragmented and inconclusive (Fang et al., 2024; Bordt et al., 2024).

To address this uncertainty, we implement a modular multi-agent LLM (MALLM) framework using GPT-4o and LangGraph, utilizing specialized agents for feature analysis, engineering, retrieval, and prediction. We conduct a head-to-head evaluation of this system (illustrated in Figure 1) across zero-shot, few-shot, and RAG settings against two distinct baselines: specialized tabular LLMs (TABLET and TABLLM) and state-of-the-art AutoML systems (AutoGluon, MLJAR, and H2O). To rigorously test generalization versus memorization, our experimental design spans eleven binary classification tasks split between five widely circulated datasets predating the LLM knowledge cutoff and six novel post-cutoff datasets from healthcare, education, housing, and workplace contexts.

Our evaluation framework prioritizes operational reliability over simple aggregate accuracy. We move beyond standard performance indicators (e.g., AUC, F1, log loss) to provide a holistic view of deployment readiness. This includes assessing probability calibration, sensitivity to class imbalance, and resilience to benign schema perturbations—specifically feature-name edits relevant to mitigating prompt-injection risks. Furthermore, we analyze instance-level hardness and typicality to determine if model confidence aligns with genuine competence on hard-to-classify and atypical instances.

To guide this investigation, we pose three research questions:

Findings of the Association for Computational Linguistics: ACL 2026, pages 40099–40132

- **RQ1:** Can a multi-agent LLM system match state-of-the-art AutoML frameworks on tabular classification tasks across diverse domains?
- **RQ2:** To what extent is reported LLM effectiveness on tabular data driven by memorization of pre-training corpora rather than genuine generalization?
- **RQ3:** How do the characteristics of the data set, such as hardness, typicality, and class balance, differentially affect the reliability of LLMs versus AutoML?

In answering these questions, our study makes four contributions that resolve the limitations of previous benchmarks. First, we introduce a contamination-aware evaluation protocol that explicitly separates memorization from reasoning. By splitting tasks between widely circulated pre-cutoff benchmarks and novel post-cutoff datasets, we disentangle the conflated factors of recall and generalization. Second, we move beyond headline metrics by implementing a suite of instance-level diagnostics to characterize behavior across the instance hardness and typicality spectra. Third, we operationalize a transparent agentic architecture for tabular classification, allowing us to isolate failure modes specific to retrieval, feature engineering, or reasoning. Fourth, we provide actionable evidence for practitioners: we demonstrate that while LLMs offer rapid prototyping benefits, mature AutoML frameworks remain the superior choice for reliability.

Our analysis yields three primary findings:

Memorization drives pre-cutoff success: On datasets likely present in pre-training corpora, the MALLM achieves performance competitive with AutoML, consistent with advantages derived from prior exposure.

Generalization failure on novel data: On post-cutoff datasets, MALLM performance degrades substantially. Predictions skew toward the majority class, decision boundaries blur, and confidence becomes poorly calibrated, particularly on hard or atypical instances.

The "AutoML Advantage" is stability, not just accuracy: Across all settings, AutoML baselines (AutoGluon, MLJAR, H2O) generalize more reliably, demonstrate greater resilience to instance hardness and atypicality, and exhibit lower variance across runs.

These results suggest that, for tabular classification, MALLMs are useful for interpretability scaffolding and few-shot adaptation, but they do not yet match the calibration and stability required for high-stakes predictive decision-making. Progress will likely depend on hybrid architectures that use language-based reasoning for feature- and context-level inference while relying on tabular specialists to learn robust decision boundaries, together with evaluation protocols that explicitly distinguish dataset recall from genuine predictive reasoning.

2 Related Work

Recent research on LLMs for tabular data spans serialization methods for classification (Hegselmann et al., 2023; Dinh et al., 2022; Slack and Singh, 2023), table reasoning (Liu et al., 2024; Nahid and Rafiei, 2024), and practical applications (Zhang et al., 2025; Liu et al., 2023). However, critical questions remain regarding whether LLM performance stems from genuine generalization or memorization, and whether these systems can credibly replace AutoML for high-stakes tabular classification.

2.1 Serialization and Prompt Engineering

While gradient-boosted decision trees (GBDTs) remain a leading conventional approach for tabular data (Shwartz-Ziv and Armon, 2021), LLMs offer unique advantages in zero-shot reasoning and interpretability. The primary challenge lies in serializing structured rows into sequential text (Fang et al., 2024). Early approaches, such as LIFT (Dinh et al., 2022) and TABLLM (Hegselmann et al., 2023), demonstrated that simple template-based serializations (e.g., "Feature Name: Value") allow LLMs to exploit semantic knowledge in column headers. These methods have proven competitive with GBDTs in few-shot settings (16–64 examples) but typically lag behind specialized models on large datasets or tasks heavily reliant on continuous numerical features (Manikandan et al., 2023).

To address these limitations, recent work focuses on enriching prompts with domain context. Subsequent work shows that prepending expert instructions, feature definitions, and schema metadata improves performance (Slack and Singh, 2023; Jaitly et al., 2023; Sui et al., 2024), though these approaches rely on static prompting without agentic reasoning.

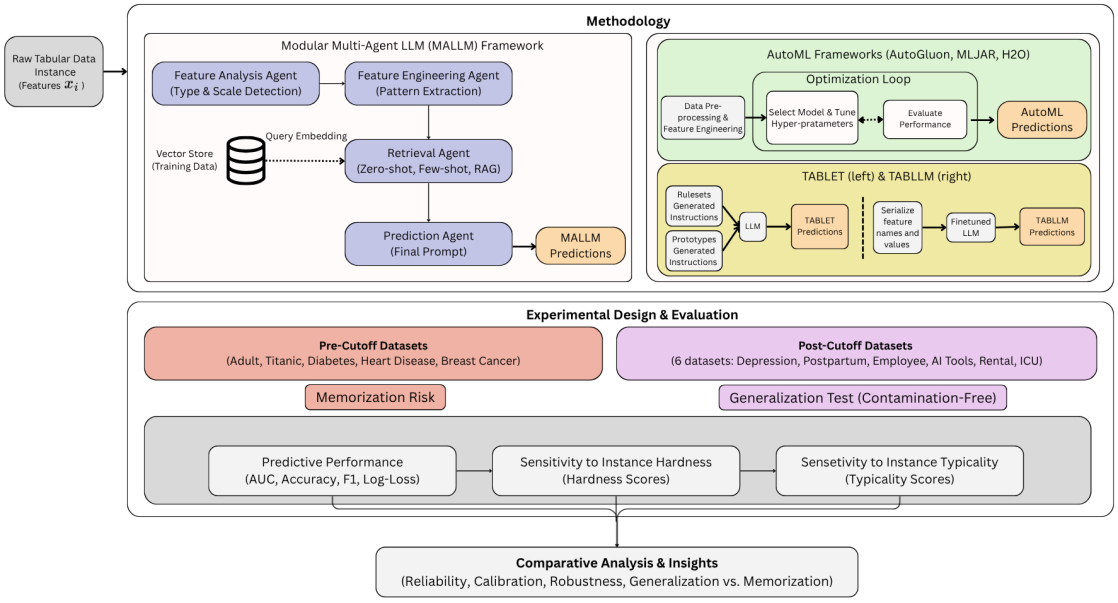

Figure 1: Multi-agent LLM framework (MALLM) and evaluation pipeline

2.2 Generalization vs. Contamination

A key tension in this domain is the risk of data contamination. LLMs may achieve high performance not through reasoning, but by recognizing datasets seen during pre-training. Bordt et al. (2024) demonstrated that GPT-4 can verbatim reproduce classic benchmarks (e.g., Iris, Wine) from memory. While some studies attribute LLM success to "world knowledge" (Hegselmann et al., 2023), distinguishing this from rote memorization is difficult without controlled evaluation. To date, no study has systematically benchmarked LLMs against AutoML using a strict separation of pre-cutoff (contaminated) and post-cutoff (unseen) datasets.

2.3 Retrieval-Augmented and Multi-Agent Systems

To overcome context window limits and improve reasoning depth, the field is moving toward modular architectures. Retrieval-Augmented Generation (RAG) has been adapted for tabular data to supply relevant historical examples as context (Fang et al., 2024). Similarly, agentic frameworks like StructGPT (Jiang et al., 2023) decompose complex tasks—such as table question-answering—into iterative reading and reasoning steps.

2.4 Summary of Open Challenges

Despite rapid progress, three gaps persist. First, evaluations conflate recall with reasoning by relying on benchmarks likely present in pre-training corpora.[1] Second, rigorous head-to-head comparisons between multi-agent LLMs and production AutoML systems are lacking. Third, prior work overlooks reliability on difficult or atypical instances (Cook et al., 2025). We bridge this by moving beyond aggregate accuracy to analyze instance-level hardness and typicality.

3 Data Collection & Measures

We evaluated our framework (Figure 1) on six binary classification datasets released after September 2024, after the knowledge cut-off of GPT-4o-2024-08-06 model, ensuring contamination-free testing. The post-cutoff datasets include Student Depression (predicting depression status among students based on demographics, academic performance, and mental health indicators), Postpartum Depression (predicting postpartum depression risk from self-reported symptoms), Employee Satisfaction (predicting employee job satisfaction in a large U.S. corporation from work-life balance, compensation, and demographic factors), Students AI Tools (predicting impact on grades from AI tool usage patterns among college students), Rental Properties (predicting low- versus high-price categories based on housing and neighborhood characteris-

[1] Descriptions and results for these five datasets are provided in Appendix E

Table 1: Overview of Post-Cutoff Datasets

Dataset	Inst.	Tot. Feat.	Cat.	Num.	Positive Class	Source
Student Depression	1000	18	8	10	58% positive	Kaggle
Postpartum Depression	1491	11	9	2	65% positive	Kaggle
Employee Satisfaction	1000	64	4	60	48% positive	Private
Students AI Tools	2000	15	8	7	55% positive	Kaggle
Rental Properties	482	18	0	18	54% positive	Private
ICU Admission	254	18	0	18	59% positive	Private

tics) and ICU Admission (predicting admission to the ICU based on patient health status). Table 1 summarizes the class distributions, feature counts, and sources for each dataset. Complete dataset descriptions, including detailed feature specifications and download links, are provided in Appendix A.

4 Methodology

This study evaluates the performance of our representative multi-agent LLMs alongside established AutoML frameworks (AutoGluon, MLJAR, and H2O) and baseline LLM-based tabular models such as TABLET and TABLLM on binary classification tasks across six diverse tabular datasets. Below, we describe the models and present our framework.

4.1 AutoML Models

We utilize three production-grade AutoML frameworks, selected for their ubiquity in industry and differing architectural strategies.

AutoGluon trains diverse base models and combines them via multi-layer stacking using the best quality preset with a 1-hour time limit (Erickson et al., 2020).

MLJAR automates the pipeline from preprocessing to ensembling, with an emphasis on transparency (Płońska and Płoński, 2021).

H2O AutoML applies random grid search across Gradient Boosting Machines, XGBoost, Deep Learning, and GLMs, combining top models into a Stacked Ensemble optimized by AUC (LeDell and Poirier, 2020).

For all frameworks, we used AUC as the primary optimization metric and enabled automatic class balancing to handle asymmetric distributions. We evaluated three data availability settings: (100% data), S2 (50% subsample), and S3 (20% subsample). To ensure statistical robustness, S2 and S3 experiments were repeated across five distinct random seeds (Bouthillier et al., 2021). Detailed hyperparameter configurations are provided in Appendix F.

4.2 Tabular LLM Baseline Models

TABLET converts structured data into semantically enriched natural-language descriptions that capture column semantics, statistics, and distributions (Slack and Singh, 2023). This model-agnostic framework utilizes instruction-tuned prompts to leverage dataset-level context, offering improved robustness and calibration compared to naïve serialization.

TABLLM is an instruction-tuned LLM trained on a diverse mix of tabular problems using schema-aware textual formats (Hegselmann et al., 2023). By learning column semantics and heterogeneous feature types, it enables calibrated predictions and significantly improves generalization across standard tabular benchmarks.

4.3 Modular Multi-Agent LLM Framework

Standard LLM prompting often struggles with the high-dimensional reasoning required for tabular data (Fang et al., 2024). To address this, we developed a modular Multi-Agent LLM (MALLM) framework orchestrated via LangGraph. Unlike fine-tuning approaches that update model weights, our agents operate via *persona-conditioned in-context learning* (Choi and Li, 2024; Wei et al., 2022). The workflow decomposes the classification problem into four distinct reasoning stages, each handled by a specialized agent instantiated with a specific system prompt (persona); detailed prompt templates are provided in Appendix M:

- **Feature Analysis Agent:** Conditioned as a "Senior Data Scientist," this agent analyzes the raw schema to infer semantic types (e.g., distinguishing nominal from ordinal integers) and scale distributions (Cyrus et al., 2024). It outputs a structured meta-data summary.
- **Feature Engineering Agent:** Acting as a "Domain Expert," this agent proposes transformation logic and identifies potential interaction effects based on the analysis output, leveraging the LLM's world knowledge to enrich feature representation (Kanter and Veeramachaneni, 2015).
- **Retrieval Agent (RAG):** This agent retrieves relevant training examples. For RAG configurations, we utilize dense vector retrieval (Lewis et al., 2020), embedding the test instance using `text-embedding-ada-002` and retrieving the k nearest neighbors ($k \in$

$\{5, 10\}$) from the training set to serve as in-context demonstrations (Choi and Li, 2024).

- **Prediction Agent:** Conditioned as a "Lead Analyst," this agent synthesizes the feature analysis, engineered insights, and retrieved examples. It employs Chain-of-Thought (CoT) prompting (Wei et al., 2022) to articulate a reasoning path before outputting the final classification probability.

All agents utilize GPT-4o-2024-08-06. We evaluate three inference strategies: **Zero-Shot** (reasoning based solely on schema), **Few-Shot** (stratified random examples), and **RAG** (semantic similarity-based examples).

4.3.1 Multi-Agent with Zero-Shot Prompting

In this baseline configuration, the prediction prompt includes only a textual description of the dataset and the test instance's features—without any training examples. This setup evaluates the model's intrinsic capacity to generalize from schema-level understanding and apply domain-agnostic reasoning. Despite its simplicity, this condition serves as a control for assessing the added value of contextual demonstrations in the other configurations.

4.3.2 Multi-Agent with Few-Shot In-Context Learning

To evaluate the impact of contextual examples, we extend the prediction prompt with a small number of labeled training instances using a stratified sampling strategy to preserve class balance. We assess two configurations: three randomly sampled examples per class and five randomly sampled examples per class. These examples are embedded directly into the prompt, alongside the dataset description and the test instance's features. The goal is to enable the LLM to perform analogical reasoning based on examples provided within the prompt.

4.3.3 Multi-Agent with Retrieval-Augmented Generation (RAG)

The RAG-based configuration introduces embedding-driven retrieval of training examples. All training samples are first embedded into a high-dimensional vector space using OpenAI's text-embedding-ada-002 model. At inference time, the test instance is also embedded and compared to training samples via cosine similarity. The top five or top ten most similar examples are retrieved and incorporated into the prompt. Unlike stratified sampling, this method offers greater contextual relevance by selecting examples that are most similar to the test instance. This approach is particularly advantageous for ambiguous test cases near decision boundaries, where personalized contextual guidance may improve predictive accuracy.

5 Results

Table 2 summarizes performance on the six post-cutoff datasets, with complete results reported in Table 13 in the Appendix. The AutoML frameworks (AutoGluon, MLJAR, and H2O) consistently achieve the strongest and most stable results, showing higher AUC, more balanced recall, and greater robustness to class imbalance. AutoGluon and H2O typically lead, particularly in maintaining clearer decision boundaries across datasets (Appendix B).

The LLM-based baselines, TABLET and TABLLM, perform markedly worse than the AutoML systems. TABLET frequently collapses to majority-class prediction, yielding deceptively high accuracy but near-zero negative recall under class imbalance. TABLLM is more stable, yet still trails AutoML in AUC and consistency. The MALLM variants outperform both baselines, with RAG-based configurations (e.g., RAG10) performing best among them. However, these models remain biased toward the positive class, achieving high positive recall at the cost of negative recall, and do not match the calibration of AutoML models.

Instance-level hardness analysis (Appendix C) reveals a substantial calibration gap. AutoML models exhibit strong negative correlations between hardness and confidence (e.g., $r < -0.85$ for AutoGluon on Student Depression) and degrade smoothly from Easy to Very Hard instances. In contrast, the baseline LLMs (TABLET and TABLLM) and the zero-shot MALLM show weaker or unstable correlations, indicating overconfidence on difficult cases. MALLM variants improve on the single-agent baselines but still trail AutoML, with sharp drops and high variance on Hard and Very Hard instances, even for RAG10.

We further assessed how representative data patterns affect performance using typicality scores (Appendix D). Typicality analysis shows AutoML models benefit most from representative instances, with strong positive correlations between typical-

ity and confidence (e.g., $r > 0.87$ for AutoGluon on Employee Satisfaction), while LLM baselines show negligible or inconsistent correlations across datasets. RAG10 partially bridges this gap in high-typicality domains such as Employee Satisfaction, but all models struggle in low-typicality settings like Postpartum Depression, where AutoML remains the most robust.

6 Additional Analysis

To complement the main results, we conducted a set of supplementary analyses on deployment feasibility, robustness, hybrid architectures, pipeline ablations, retrieval quality, and prompt optimization. Across these analyses, the overall pattern is consistent: MALLM shows promise as a supporting component, but its limitations as a standalone tabular classifier appear to be structural.

From a deployment perspective, MALLM incurs substantial inference-time latency and token cost, making it less practical for real-time or high-throughput applications, whereas AutoML systems place most of the computational burden at training time (Appendix G). Robustness tests further show that MALLM predictions are generally stable across repeated runs and temperature settings on well-structured datasets, although variance increases on more complex tasks; this suggests that residual instability is driven more by prompt sensitivity than by stochastic inference itself (Appendix H; Table 16).

We also examined whether LLMs are more effective in complementary than end-to-end roles. In hybrid experiments, XGBoost consistently outperformed MALLM as the final predictor, but LLMs contributed some value as upstream feature-analysis tools that help surface relevant variables and interactions (Appendix I). Relatedly, the ablation results indicate that the multi-agent pipeline does not uniformly improve performance: retrieval is the most consistently beneficial component, whereas additional agents often add noise and reduce reliability (Appendix J).

Finally, we evaluated whether retrieval quality or prompt design explains the remaining performance gap. Changing embedding models led to only minimal improvement, suggesting that retrieval is not the main bottleneck; instead, the core limitation lies in the model's ability to use retrieved examples effectively for tabular reasoning (Appendix K). Likewise, automated prompt optimization via DSPy produced only marginal gains over manually designed prompts, indicating that prompt refinement alone is unlikely to overcome the broader weaknesses of LLM-based tabular modeling (Appendix L).

7 Discussion and Conclusion

7.1 The Mirage of Generalization: Memorization vs. Reasoning

A central debate in recent NLP research is whether LLMs generalize to novel distributions or merely interpolate within their massive pre-training corpora (Bordt et al., 2024).

As shown in Table 2, the performance gap is stark. The difference is most pronounced on Postpartum Depression: AutoGluon achieves an AUC of **0.997** versus **0.809** for MALLM, with a six-fold Log Loss gap (0.095 vs. 0.621). This systematic collapse on unseen data suggests that the "reasoning" capabilities reported in prior work (Hegselmann et al., 2023; Slack and Singh, 2023) may instead reflect latent familiarity with widely circulated benchmarks rather than genuine inference from prompt context.

7.2 Operational Reliability and the Calibration Gap

A fundamental calibration crisis emerges from our instance-level diagnostics.

As illustrated in Figure 2, AutoML models exhibit desirable bimodal probability distributions that cleanly separate classes. In contrast, MALLMs produce diffuse distributions with substantially greater uncertainty. More importantly, our hardness analysis (see Table 3 in the Appendix) shows that AutoML models possess stronger "self-knowledge" of data complexity. We observe strong negative correlations (ranging from $r = -0.85$ to -0.87) between instance hardness and confidence for AutoGluon and MLJAR on the depression and satisfaction datasets. This indicates that these models appropriately reduce confidence near decision boundaries or in noisier regions.

Conversely, LLM baselines (TABLET) and zero-shot agents frequently exhibit weak or erratic correlations (e.g., $r = -0.17$ on Student Depression), and on Students AI Tools even positive correlations ($r = 0.49$ for TABLLM), implying dangerous overconfidence on hard instances. This overconfidence on difficult cases represents a critical failure mode in any deployment setting where reliable uncertainty quantification is essential.

7.3 The Limits of Retrieval-Augmented Generation (RAG)

While RAG is often proposed as a remedy for hallucination (Lewis et al., 2020), our results qualify its utility for tabular classification. The RAG10 configuration consistently outperformed Zero-shot and Few-shot baselines (Table 2).

However, our typicality analysis (Appendix D) clarifies the boundary conditions of this improvement. RAG agents perform well only when the test instance is highly typical, that is, statistically similar to the retrieved training examples. As shown in Table 9, when typicality is high, RAG10 achieves strong corrected probabilities (e.g., 0.946 on Student Depression). Yet on low-typicality data—outliers or edge cases—performance deteriorates substantially relative to AutoML. RAG therefore narrows, but does not eliminate, the gap with AutoML on atypical instances.

7.4 Conclusion and Future Directions

Taken together, these findings reframe the central question from whether LLMs can replace AutoML to how the strengths of each paradigm can be productively combined. We conclude with three recommendations for the ACL and ML communities:

1. **Mandatory Contamination-Aware Benchmarking:** Evaluation of tabular LLMs should explicitly exclude datasets potentially present in pre-training corpora. We propose temporally strict post-cutoff datasets as a standard for claims about "reasoning."

2. **Beyond Aggregate Metrics:** Future work should prioritize instance-level diagnostics (hardness and typicality) over global AUC/Accuracy scores. A model that hallucinates confidence on hard instances is unsafe, regardless of its average F1 score.

3. **Hybrid Neurosymbolic Architectures:** The semantic strengths of LLMs (feature engineering, explanation) should be combined with the statistical robustness of GBDTs. Future systems should likely use LLMs as "Data Scientist Agents" that orchestrate and interpret traditional AutoML pipelines, rather than as the inference engine itself.

Ultimately, while LLMs offer a powerful interface for interacting with data, statistical learning remains the stronger paradigm for classifying it.

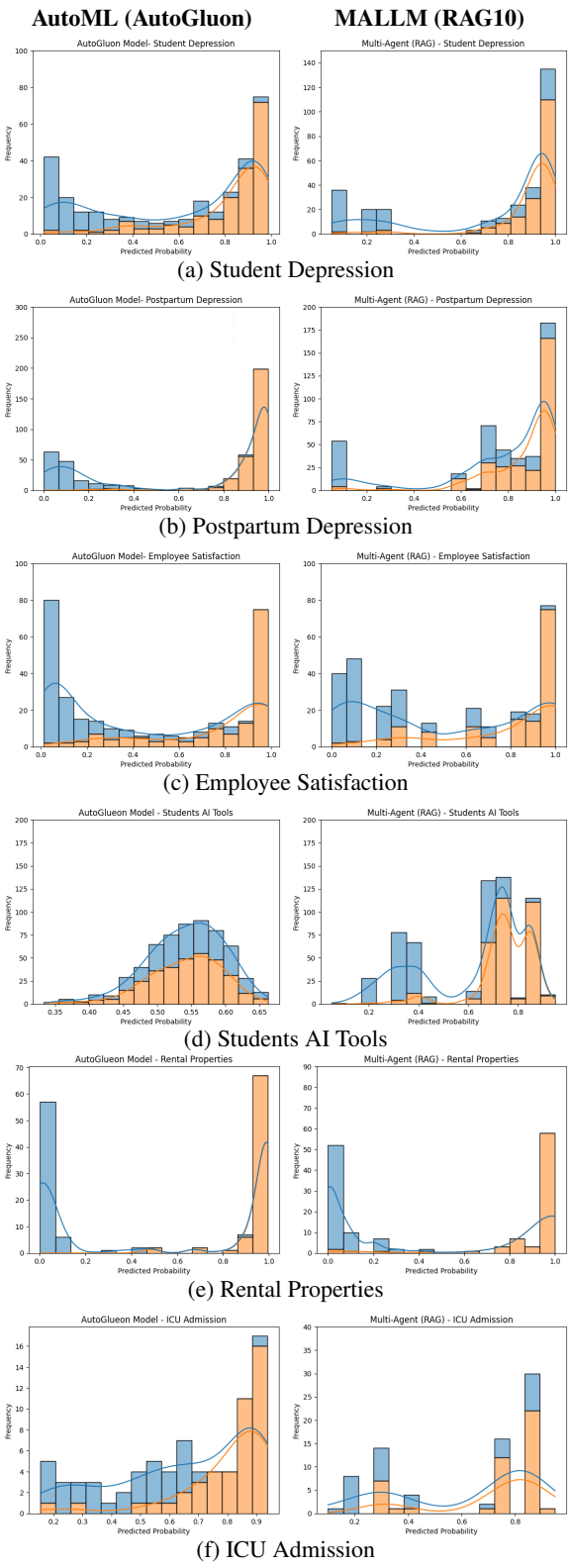

Figure 2: Predicted probability distribution comparison between the best AutoML baseline (AutoGluon) and the best Multi-Agent LLM (MALLM-RAG10) models across all six post-cutoff datasets. AutoGluon consistently achieves sharper bimodal separation (high confidence), whereas the MALLM exhibits broader uncertainty and mid-range overlap.

Table 2: Performance on Post-Cutoff Datasets (Best Configurations), Full results in Appendix Table 13.

Model	AUC	Accuracy	F1 Score	Recall+	Recall–	Precision	Log Loss
STUDENT DEPRESSION							
AutoGluon-S1	**0.912**	0.833	0.861	0.886	0.760	0.838	0.367
MLJAR-S1	0.910	0.833	0.860	0.874	0.776	0.845	0.372
H2O-S1	0.910	0.833	0.860	0.874	0.776	0.845	0.370
TABLET	0.672	0.650	0.734	0.829	0.400	0.659	0.647
TABLLM	0.734	0.750	0.738	0.829	0.640	0.745	0.655
MALLM-RAG10	0.842	0.790	0.835	0.909	0.624	0.801	0.801
POSTPARTUM DEPRESSION							
AutoGluon-S1	0.997	0.980	0.985	0.983	0.975	0.986	0.095
MLJAR-S1	0.987	0.922	0.941	0.962	0.847	0.921	0.145
H2O-S1	**0.998**	0.931	0.944	0.894	1.000	1.000	0.068
TABLET	0.526	0.652	0.751	0.808	0.363	0.702	0.715
TABLLM	0.783	0.833	0.801	0.949	0.618	0.844	0.473
MALLM-RAG10	0.809	0.752	0.838	0.986	0.319	0.827	0.621
EMPLOYEE SATISFACTION							
AutoGluon-S1	**0.946**	0.843	0.833	0.791	0.895	0.880	0.308
MLJAR-S1	0.944	0.853	0.842	0.791	0.915	0.900	0.342
H2O-S1	0.940	0.840	0.836	0.824	0.855	0.847	0.330
TABLET	0.521	0.400	0.571	1.000	0.000	0.400	0.979
TABLLM	0.829	0.830	0.829	0.770	0.888	0.835	0.479
MALLM-RAG10	0.912	0.830	0.834	0.865	0.796	0.832	0.698
STUDENTS AI TOOLS							
H2O-S1	**0.508**	0.538	0.657	0.79	0.216	0.558	0.733
MLJAR-S1	0.507	0.530	0.595	0.624	0.414	0.569	0.688
AutoGluon-S1	0.485	0.530	0.648	0.780	0.220	0.553	0.697
TABLET	0.502	0.556	0.674	0.828	0.220	0.538	2.094
TABLLM	0.502	0.555	0.360	1.000	0.004	0.777	1.083
MALLM-RAG10	0.500	0.553	0.712	1.000	0.000	0.277	0.814
RENTAL PROPERTIES							
AutoGluon-S1	**0.998**	0.979	0.981	0.974	0.985	0.987	0.063
MLJAR-S1	0.989	0.959	0.962	0.962	0.955	0.962	0.167
H2O-S1	0.997	0.979	0.981	0.974	0.985	0.987	0.074
TABLET	0.891	0.869	0.869	0.779	0.971	0.882	1.209
TABLLM	0.828	0.531	0.347	1.000	0.000	0.266	0.670
MALLM-RAG10	0.982	0.965	0.967	0.948	0.985	0.965	0.353
ICU ADMISSION							
AutoGluon-S1	**0.915**	0.833	0.861	0.886	0.760	0.838	0.367
MLJAR-S1	0.906	0.844	0.875	0.933	0.719	0.824	0.517
H2O-S1	0.903	0.870	0.894	0.933	0.781	0.857	0.407
TABLET	0.693	0.662	0.655	0.689	0.625	0.655	1.416
TABLLM	0.517	0.584	0.369	1.000	0.000	0.292	0.685
MALLM-RAG10	0.731	0.701	0.797	1.000	0.281	0.831	0.593

Notes: S1 = full training set, Recall+/Recall- = positive/negative class recall, Bold = best per metric per dataset

8 Limitations

This study has several limitations that should be considered when interpreting our findings on LLM reliability for tabular classification.

First, although we evaluate eleven datasets under a strict pre-/post-cutoff design, the six post-cutoff datasets are concentrated in healthcare, workplace, and education settings. Generalization to other high-stakes domains, such as finance, insurance, or industrial operations, remains untested. In addition, the post-cutoff datasets are relatively small ($N < 2000$), and we do not evaluate MALLM on large-scale tabular problems where context-window constraints and retrieval design may interact differently with model performance.

Second, while we report indicative latency and API-cost comparisons in Appendix G, we do not provide a full systems-level benchmark of deployment efficiency across hardware, concurrency, or throughput settings. The representative MALLM framework relies on multi-step agentic inference and retrieval, which introduces substantially higher serving cost and latency than optimized tree-based ensembles such as AutoGluon and MLJAR. These overheads may further limit the practicality of MALLMs in real-time or resource-constrained production environments.

Third, we evaluate a fixed set of structured prompts and specific zero-shot, few-shot, and RAG configurations rather than performing exhaustive prompt optimization. Because LLM behavior can be sensitive to prompt formulation, stronger domain-specific prompting may improve performance in some settings. However, such gains would likely come at the cost of automation, reproducibility, and portability across tasks.

Fourth, our contamination-aware design reduces, but cannot eliminate, the possibility of benchmark contamination. Because the pre-training corpus of proprietary foundation models is not publicly verifiable, we cannot rule out exposure to similar distributions, pre-release versions, or closely related benchmark descriptions. Our temporal split therefore should be interpreted as a strong mitigation strategy rather than definitive proof of contamination-free evaluation.

Fifth, several post-cutoff datasets contain sensitive attributes and outcomes related to mental health and workplace evaluation. Although acceptable for research benchmarking, API-based LLM pipelines require transmitting data to external servers, which may limit applicability in regulated settings relative to local AutoML systems that can run entirely on-premise.

Sixth, we do not conduct subgroup fairness analysis. We therefore cannot determine whether errors, calibration failures, or confidence misalignment vary systematically across age, gender, race, or socioeconomic status. This is particularly important in the mental health and employment settings studied here, where historical biases in data collection and labeling may produce disparate harms.

Finally, all experiments are restricted to binary classification. We do not evaluate multiclass classification, regression, or time-series forecasting, where the relative strengths and weaknesses of LLM-based systems may differ, particularly when label semantics or more complex numerical reasoning play a larger role.

9 Ethical Considerations

This study raises several ethical considerations that are important for interpreting both the findings and their potential downstream use.

First, the six post-cutoff datasets include sensitive information from vulnerable or potentially affected populations, including students experiencing depression, postpartum mothers at risk of anxiety, and employees in workplace evaluation settings. Our use of these datasets is limited to research and benchmarking. The results should not be interpreted as support for autonomous deployment in clinical, employment, or educational decision-making. In particular, the poor calibration, class imbalance, and high variance exhibited by LLM-based approaches on post-cutoff data suggest that such systems are not suitable for high-stakes use without substantial additional validation, domain oversight, and monitoring.

Second, our study does not include subgroup fairness analysis. We therefore cannot assess whether errors, calibration failures, or confidence misalignment vary systematically across age, gender, race, or socioeconomic status. This omission is especially important in the mental health and employment domains considered here, where historical biases in data collection, diagnosis, and evaluation may already be embedded in the underlying data. Future work should extend contamination-aware evaluation to include subgroup and intersectional fairness analyses.

Third, API-based LLM pipelines may raise pri-

vacy and governance concerns because they require transmitting structured records to external services. Although this may be acceptable for public benchmarking, it can create additional risks in regulated or sensitive settings compared with local AutoML systems that can be deployed fully on-premise. Researchers and practitioners should therefore consider not only predictive performance, but also data handling constraints, legal compliance, and organizational governance when evaluating LLM-based tabular systems.

Finally, the capabilities studied here could be misused for surveillance, profiling, or discriminatory screening, particularly if organizations emphasize performance on familiar benchmarks while ignoring failures on novel data. Commercial incentives may encourage premature deployment despite clear evidence of weaker generalization and calibration. We therefore encourage the community to adopt contamination-aware evaluation, fairness auditing, and domain-specific risk assessment as minimum requirements before considering real-world deployment.

References

Sebastian Bordt, Harsha Nori, Vasco Rodrigues, Besmira Nushi, and Rich Caruana. 2024. Elephants never forget: Memorization and learning of tabular data in large language models. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szuto, Nazanin Mohajer, Vahid Nouriborji, Eugene Cheplygina, Veronika anderos, and 1 others. 2021. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems*, volume 3, pages 729–749.

Balderton Capital. 2025. Prior labs raises €9m to revolutionise how businesses interact with their tabular data. https://www.balderton.com/news/prior-labs-raises-e9m-to-revolutionise-how-businesses-interact-with-their-tabular-data/. Accessed: 24.09.2025.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1721–1730.

Hyeong Kyu Choi and Yixuan Li. 2024. Picle: Eliciting diverse behaviors from large language models with persona in-context learning. In *International Conference on Machine Learning (ICML)*.

Ryan A Cook, John P Lalor, and Ahmed Abbasi. 2025. No simple answer to data complexity: An examination of instance-level complexity metrics for classification tasks. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2553–2573.

David Cyrus, Dany Varghese, and Alireza Tamaddoni-Nezhad. 2024. An inductive logic programming approach for feature-range discovery. In *International Joint Conference on Learning and Reasoning (IJCLR)*, pages 203–217. Springer.

Romain Dillet. 2025. Neuralk-ai is developing ai models specifically designed for structured data. https://techcrunch.com/2025/02/03/neuralk-ai-is-developing-ai-models-specifically-designed-for-structured-data/. Accessed: 24.09.2025.

Tuan Dinh, Yuchen Zeng, Ruiqi Zhang, Zihao Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. 2022. Lift: Language-interfaced fine-tuning for non-language machine learning tasks. *Advances in Neural Information Processing Systems*, 35:11763–11784.

Nicholas Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*.

Xi Fang, Weijie Xu, Fiona Anting Tan, Jianfeng Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. 2024. Large language models(llms) on tabular data: Prediction, generation, and understanding – a survey. *arXiv preprint arXiv:2402.17944*.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. 2023. Tabllm: Few-shot classification of tabular data with large language models. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pages 5549–5581. PMLR.

HopesB. 2023. Student depression dataset. https://www.kaggle.com/datasets/hopesb/student-depression-dataset. Kaggle dataset. Licensed under Apache License 2.0.

Shubham Jaitly, Tanay Shah, Anshul Shugani, and Rajat Singh Grewal. 2023. Towards better serialization of tabular data for few-shot classification with large language models. *arXiv preprint arXiv:2312.12464*.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Xu Zhao, and Ji-Rong Wen. 2023. Structgpt: A general framework for large language model to reason over structured data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9237–9251, Singapore. Association for Computational Linguistics.

James Max Kanter and Kalyan Veeramachaneni. 2015. Deep feature synthesis: Towards automating data science endeavors. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10. IEEE.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines.

Hyunjin Kwon, Matthew Greenberg, Colin Bruce Josephson, Joon Lee, and 1 others. 2024. Measuring the prediction difficulty of individual cases in a dataset using machine learning. *Scientific Reports*, 14(1):10474.

Erin LeDell and Sebastien Poirier. 2020. H2o automl: Scalable automatic machine learning. In *Proceedings of the AutoML Workshop at ICML*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K"uttler, Mike Lewis, Wen-tau Yih, Tim Rockt"aschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Shang-Ching Liu, ShengKun Wang, Tsungyao Chang, Wenqi Lin, Chung-Wei Hsiung, Yi-Chen Hsieh, Yu-Ping Cheng, Sian-Hong Luo, and Jianwei Zhang. 2023. JarviX: A LLM no code platform for tabular data analysis and optimization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 622–630, Singapore. Association for Computational Linguistics.

Tianyang Liu, Fei Wang, and Muhao Chen. 2024. Rethinking tabular data understanding with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 450–482, Mexico City, Mexico. Association for Computational Linguistics.

Hariharan Manikandan, Yiding Jiang, and J. Zico Kolter. 2023. Language models are weak learners. *arXiv preprint arXiv:2306.14101*.

Md Mahadi Hasan Nahid and Davood Rafiei. 2024. NormTab: Improving symbolic reasoning in LLMs through tabular data normalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3569–3585, Miami, Florida, USA. Association for Computational Linguistics.

Pedro Yuri Arbs Paiva, Camila Castro Moreno, Kate Smith-Miles, Maria Gabriela Valeriano, and Ana Carolina Lorena. 2022. Relating instance hardness to classification performance in a dataset: a visual approach. *Machine Learning*, 111(7):3085–3123.

Parvez Al Muqtadir. 2023. Postpartum depression dataset. https://www.kaggle.com/datasets/parvezalmuqtadir2348/postpartum-depression. Kaggle dataset. Licensed under the MIT License.

Aleksandra Płońska and Piotr Płoński. 2021. Mljar: State-of-the-art automated machine learning framework for tabular data (version 0.10.3). https://github.com/mljar/mljar-supervised.

Rakesh Kapilavai. 2025. Ai tool usage by indian college students 2025. https://www.kaggle.com/datasets/rakeshkapilavai/ai-tool-usage-by-indian-college-students-2025. Kaggle dataset. Licensed under CC BY-NC-SA 4.0.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Ravid Shwartz-Ziv and Amitai Armon. 2021. Tabular data: Deep learning is not all you need. *arXiv preprint arXiv:2106.03253*.

Dylan Slack and Sameer Singh. 2023. Tablet: Learning from instructions for tabular data. *arXiv preprint arXiv:2304.13188*.

Yuan Sui, Jiaru Zou, Mengyu Zhou, Xinyi He, Lun Du, Shi Han, and Dongmei Zhang. 2024. TAP4LLM: Table provider on sampling, augmenting, and packing semi-structured data for LLM reasoning. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Gustavo P. Torquette, Victor S. Nunes, Pedro Y. A. Paiva, Lourenço B. C. Neto, and Ana C. Lorena. 2022. Characterizing instance hardness in classification and regression problems. *arXiv preprint arXiv:2212.01897*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Xiaokang Zhang, Sijia Luo, Bohan Zhang, Zeyao Ma, Jing Zhang, Yang Li, Guanlin Li, Zijun Yao, Kangli Xu, Jinchang Zhou, Daniel Zhang-Li, Jifan Yu, Shu Zhao, Juanzi Li, and Jie Tang. 2025. Tablellm: Enabling tabular data manipulation by llms in real office usage scenarios. *Preprint*, arXiv:2403.19318.

Appendix A: Data Collection & Measures

We evaluated our framework on six datasets released after September 2024, following the deployment of the GPT-4o-2024-08-06 model. The post-cutoff datasets used in this evaluation include Student Depression (HopesB, 2023) (Apache License 2.0), Postpartum Depression (Parvez Al Muqtadir, 2023) (MIT License), Employee Satisfaction, Students AI Tools (Rakesh Kapilavai, 2025) (CC BY-NC-SA 4.0 License), Rental Properties, and ICU Admission.

The Student Depression dataset (HopesB, 2023)[2] is released under the Apache License 2.0 and is designed to assess and predict depression status among students. It includes features capturing demographic information (e.g., age and gender), academic performance (e.g., grades and attendance), lifestyle habits (e.g., sleep patterns, exercise, and social activities), mental health history, and responses to standardized depression-related questionnaires. The target variable is *Depression Status*, a binary indicator (Yes/No), as summarized in Table 1.

The Postpartum Depression dataset (Parvez Al Muqtadir, 2023)[3] is distributed under the MIT License and contains self-reported information from postpartum mothers, with the goal of predicting the risk of postpartum depression. Key features include age, emotional states such as sadness or tearfulness, irritability toward the baby and partner, sleep difficulties, and other psychosocial indicators. The target variable is *feeling anxious*, which is commonly used as a proxy for postpartum depression risk.

The Employee Satisfaction dataset is derived from survey data collected within a large U.S. corporation and measures satisfaction across multiple dimensions, including work–life balance, compensation, job security, managerial support, and opportunities for career advancement. Demographic attributes such as age, gender, department, and tenure are also included. The target variable corresponds to overall *job satisfaction*.

The Students AI Tools dataset (Rakesh Kapilavai, 2025)[4] is released under the Creative Commons Attribution–NonCommercial–ShareAlike 4.0 license and captures patterns of AI tool usage (e.g., ChatGPT, Gemini, Copilot) among college students in academic settings. The dataset consists of 16 attributes related to AI tool adoption, trust, perceived usefulness, internet access, and self-reported academic outcomes. The target variable is *impact on grades*, making the dataset suitable for predictive modeling tasks in education analytics.

The Rental Properties dataset focuses on predicting low- versus high-price categories based on housing and neighborhood characteristics(e.g., number of bedrooms, median income, crime rate,.. etc). This dataset captures variation in structural attributes, location-based features, and surrounding neighborhood conditions, providing a realistic setting for evaluating model performance in price classification tasks.

The ICU Admission dataset examines the prediction of ICU admission (Yes/No) based on patient health status and clinical indicators. It includes a range of medical features that reflect patient condition, severity, and risk factors (e.g., age, heart rate, blood pressure and oxygen saturation). This dataset is particularly useful for evaluating model performance in healthcare decision-making contexts, where accurate identification of high-risk patients is critical.

Appendix B: Probability Distribution Analysis

Figures 3 and 4 illustrate the distribution of predicted probabilities for each model across the post-cutoff datasets. AutoML models (AutoGluon, MLJAR, and H2O) produce sharply bimodal distributions, with negatives near 0 and positives near 1, indicating strong calibration and clear class separation.

In comparison, the LLM-based baselines and our representative MALLM framework yield broader and more overlapping distributions. TABLET performs the least reliably, with unstable predictions spread across the entire probability range. TABLLM offers modest improvement but still shows substantial overlap and mid-range clustering. The MALLM variants provide more structure but still fall short of AutoML clarity. Zero-shot produces the flattest distributions, while few-shot5 and RAG10 reduce overlap somewhat, particularly on Postpartum Depression. Although RAG10 is the sharpest among the LLM approaches, it remains

[2] https://www.kaggle.com/datasets/hopesb/student-depression-dataset

[3] https://www.kaggle.com/datasets/parvezalmuqtadir2348/postpartum-depression

[4] https://www.kaggle.com/datasets/rakeshkapilavai/ai-tool-usage-by-indian-college-students-2025

less decisive than any AutoML model.

Appendix C: Instance Hardness Analysis

In machine learning, instance hardness describes how difficult a data point is to classify (Paiva et al., 2022; Torquette et al., 2022; Kwon et al., 2024). Easy instances lie far from decision boundaries, whereas hard instances are ambiguous, noisy, or boundary-adjacent (Cook et al., 2025). We measure instance hardness using the PyHard package, which provides model-agnostic metrics including distance-based measures, neighbor counts, Naive Bayes–based probabilities, and decision tree–based local complexity. These metrics offer a comprehensive, classifier-independent view of instance hardness. Accordingly, we use the full set of PyHard features and report mean instance hardness across models.

Table 3 summarizes the correlations between instance hardness and corrected predicted probability (defined as the likelihood assigned to the ground-truth class) across all post-cutoff datasets. Most models show negative correlations, indicating that confidence decreases as instances become harder (to classify).

Among all methods, AutoGluon, MLJAR, and H2O exhibit the strongest and most stable negative correlations—especially on the Student Depression, Employee Satisfaction, and ICU Admission datasets—demonstrating the clearest alignment between hardness and predictive confidence. Their weaker correlations on Postpartum Depression still remain negative, reflecting meaningful but reduced sensitivity.

The LLM baselines, TABLET and TABLLM, show inconsistent behavior. While both maintain negative correlations on most datasets, they produce positive correlations on Students AI Tools, indicating overconfidence on harder instances—a sign of poor calibration.

The MALLM models consistently yield negative correlations but with smaller magnitudes than the AutoML systems. They track instance hardness moderately well on Postpartum Depression and Students AI Tools, yet their sensitivity remains weaker overall.

Overall, AutoML frameworks show the most coherent and reliable relationship between instance hardness and model confidence, while MALLMs improve over single-agent LLM baselines but still lag behind AutoML in calibration and stability.

Tables 4 and 5 present the analysis of model performance stratified by data hardness levels—Easy, Medium, Hard, and Very Hard—across all datasets. The results reveal clear and consistent trends: as data hardness increases, the average corrected probability systematically declines across all models and datasets.

According to the results, AutoGluon, MLJAR, and H2O show the most stable degradation, with high means on Easy/Medium instances and gradual, predictable declines on Hard and Very Hard cases. Their standard deviations and coefficients of variation remain low, indicating reliable calibration even under challenging conditions.

In contrast, TABLET and TABLLM display much higher variability and weaker alignment with instance hardness. TABLET often shows large standard deviations across all hardness categories, while TABLLM performs well only on easy cases and degrades sharply—sometimes with CV values exceeding 1.0—on hard instances.

The MALLMs perform strongly on Easy and Medium instances, often achieving near-perfect confidence, but they experience steep drops in the Hard and Very Hard categories. Variability increases substantially as hardness rises, signaling reduced robustness to hard or atypical examples. Among them, RAG10 shows the best resilience but still lags behind the AutoML systems.

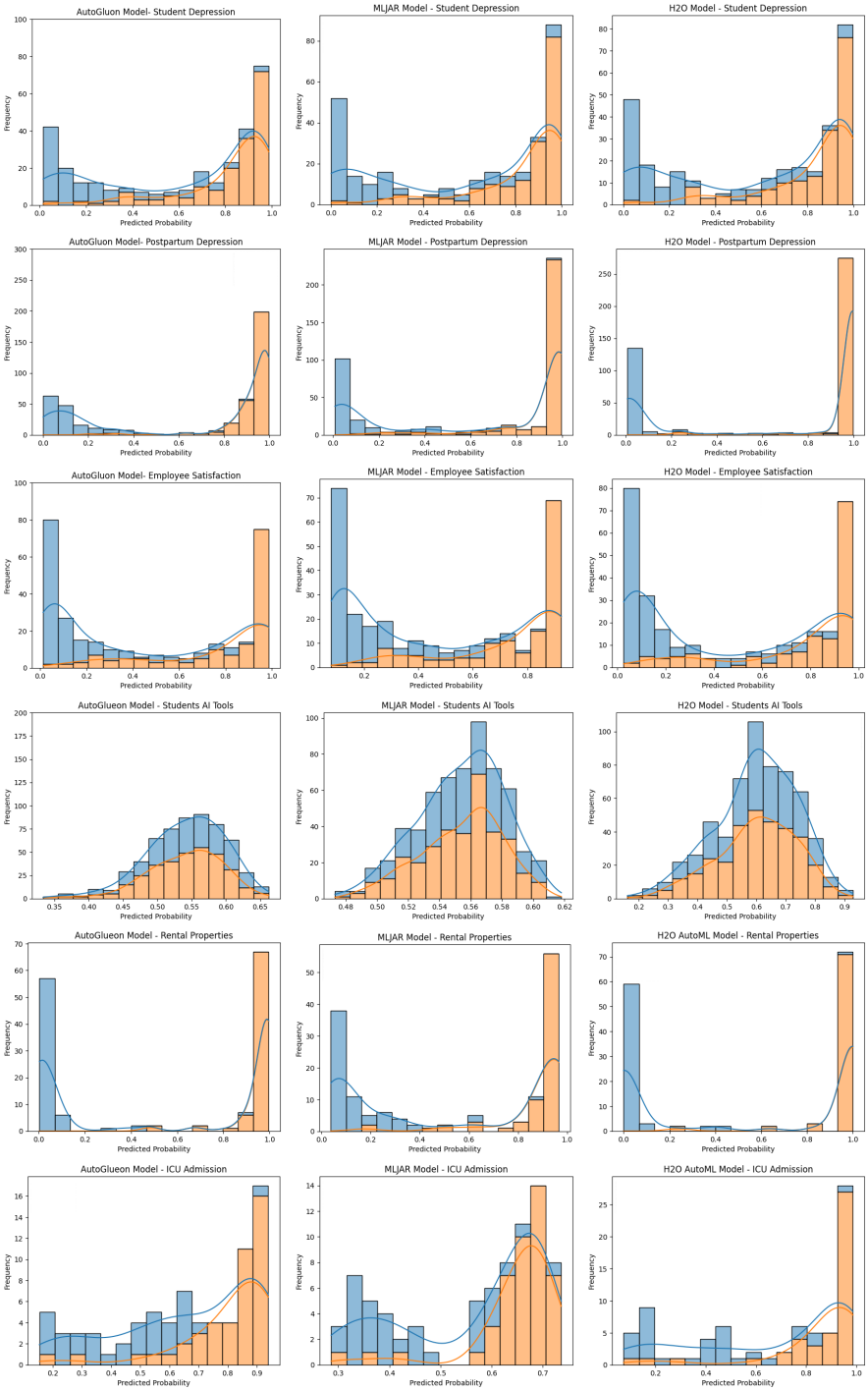

Figure 3: Predicted Probability Distributions for Post-Cutoff Datasets; from top to bottom: Student Depression, Postpartum Depression, Employee Satisfaction, Students AI Tools, Rental Properties, ICU Admission Datasets; from left to right, AutoGlueon, MLJAR, and H2O Models

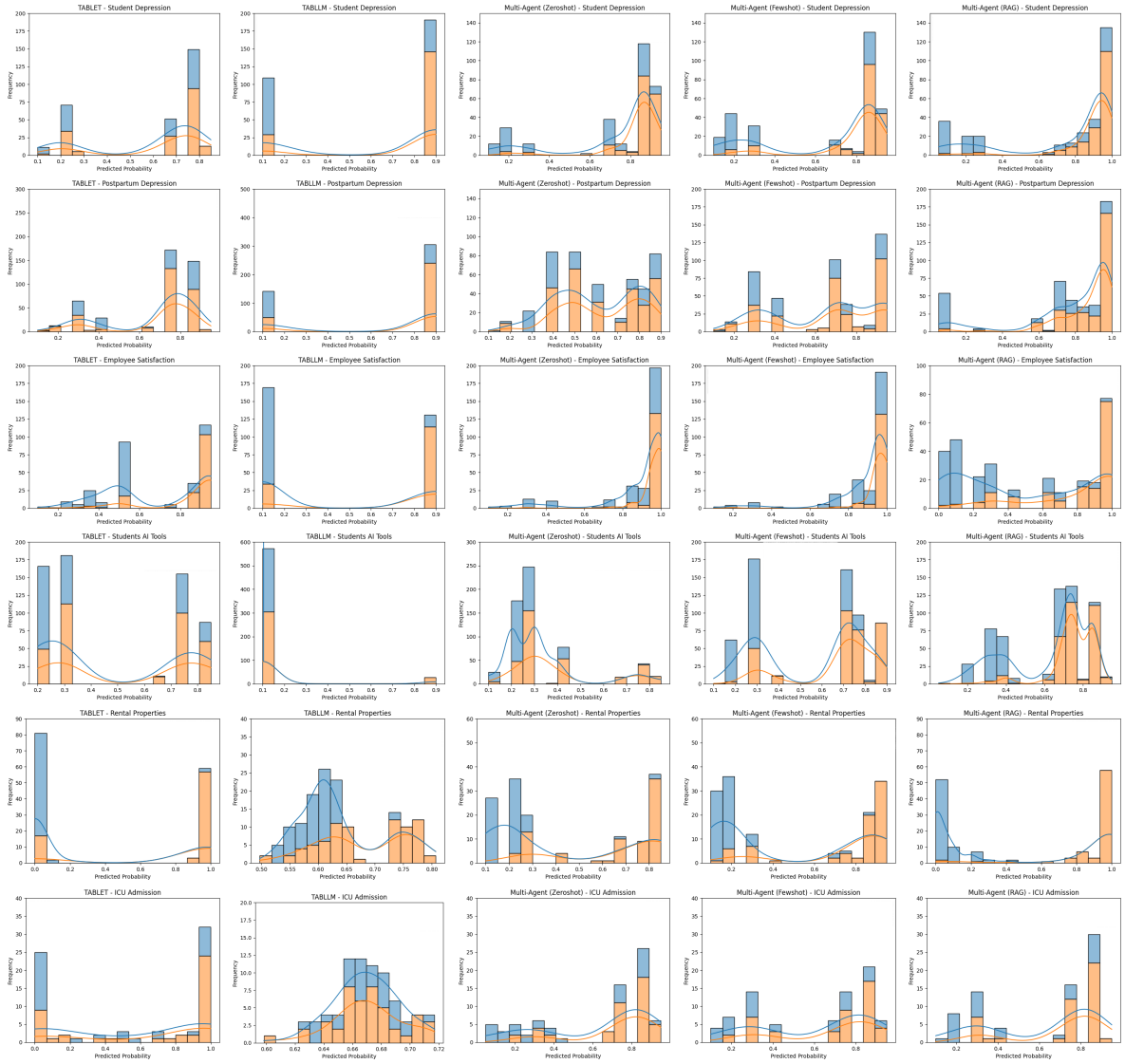

Figure 4: Predicted Probability Distributions for Post-Cutoff Datasets; from top to bottom: Student Depression, Postpartum Depression, Employee Satisfaction, Students AI Tools Datasets, Rental Properties, and ICU Admission; from left to right, TABLET, TABLLM, Zero-shot, Few-shot5, and RAG10 MALLMs Models

Table 3: Correlation between Algorithm Hardness & Corrected Probability for Post-Cutoff Dataset

Dataset	**AutoGluon**	**MLJAR**	**H2O**	**TABLET**	**TABLLM**	**MALLM-Zeroshot**	**MALLM-Fewshot5**	**MALLM-RAG10**
Student Depression	-0.8526	-0.8498	-0.8540	-0.1797	-0.6130	-0.4887	-0.3306	-0.5733
Postpartum Depression	-0.4448	-0.5191	-0.2724	-0.3814	-0.5324	-0.5105	-0.4986	-0.5609
Employee Satisfaction	-0.8690	-0.8814	-0.8531	-0.6867	-0.7007	-0.4575	-0.4371	-0.6968
Students AI Tools	-0.3426	-0.5195	-0.2580	0.05013	0.4922	-0.5124	-0.5505	-0.5467
Rental Properties	-0.4735	-0.7418	-0.3880	-0.6415	-0.4106	-0.4904	-0.5345	-0.3467
ICU Admission	-0.7339	-0.8036	-0.7071	-0.3793	-0.3578	-0.6234	-0.5336	-0.5509

The correlations are statistically significant at the 0.01% level.

Table 4: Models Average Corrected Probability & Various Levels of Hardness — Student Depression & Postpartum Depression Datasets

Dataset	Model	Hardness Category	Hardness Interval	Mean	Std	Coeff Var
Student Depression	AutoGluon	Easy	(0.388, 0.453]	0.9495	0.0344	0.0363
		Medium	(0.453, 0.499]	0.8968	0.0688	0.0767
		Hard	(0.499, 0.571]	0.7935	0.1468	0.1850
		Very Hard	(0.571, 0.738]	0.4400	0.2469	0.5611
	MLJAR	Easy	(0.388, 0.453]	0.9701	0.0327	0.0337
		Medium	(0.453, 0.499]	0.9252	0.0651	0.0704
		Hard	(0.499, 0.571]	0.8159	0.1550	0.1899
		Very Hard	(0.571, 0.738]	0.4301	0.2666	0.6199
	H2O	Easy	(0.388, 0.453]	0.9654	0.0355	0.0368
		Medium	(0.453, 0.499]	0.9170	0.0673	0.0734
		Hard	(0.499, 0.571]	0.8077	0.1521	0.1884
		Very Hard	(0.571, 0.738]	0.4314	0.2574	0.5966
	TABLET	Easy	(0.388, 0.452]	0.7467	0.4378	0.5846
		Medium	(0.452, 0.502]	0.6000	0.4932	0.8220
		Hard	(0.502, 0.573]	0.5600	0.4997	0.8924
		Very Hard	(0.573, 0.738]	0.4933	0.5033	1.0202
	TABLLM	Easy	(0.388, 0.453]	1.0000	0.0000	0.0000
		Medium	(0.453, 0.499]	0.9000	0.1973	0.2055
		Hard	(0.499, 0.572]	0.7333	0.4452	0.6071
		Very Hard	(0.571, 0.738]	0.3200	0.4696	1.4676
	MALLM-Zeroshot	Easy	(0.388, 0.452]	0.9333	0.2511	0.2691
		Medium	(0.452, 0.502]	0.8133	0.3923	0.4823
		Hard	(0.502, 0.573]	0.7333	0.4452	0.6071
		Very Hard	(0.573, 0.738]	0.4000	0.4932	1.2330
	MALLM-Fewshot5	Easy	(0.388, 0.453]	1.0000	0.0000	0.0000
		Medium	(0.453, 0.499]	1.0000	0.0000	0.0000
		Hard	(0.499, 0.572]	0.8000	0.4027	0.5034
		Very Hard	(0.572, 0.738]	0.3600	0.4832	1.3423
	MALLM-RAG10	Easy	(0.388, 0.453]	1.0000	0.0000	0.0000
		Medium	(0.453, 0.499]	0.9467	0.2262	0.2390
		Hard	(0.499, 0.572]	0.8533	0.3562	0.4174
		Very Hard	(0.572, 0.738]	0.3600	0.4832	1.3423
Postpartum Depression	AutoGluon	Easy	(0.401, 0.462]	0.9586	0.1101	0.1149
		Medium	(0.462, 0.505]	0.9431	0.0563	0.0597
		Hard	(0.505, 0.575]	0.9158	0.1196	0.1306
		Very Hard	(0.575, 0.81]	0.8174	0.1722	0.2107
	MLJAR	Easy	(0.404, 0.465]	0.9725	0.1028	0.1057
		Medium	(0.465, 0.513]	0.9735	0.0544	0.0558
		Hard	(0.513, 0.587]	0.9037	0.1344	0.1487
		Very Hard	(0.587, 0.811]	0.7578	0.2642	0.3486
	H2O	Easy	(0.411, 0.459]	0.9766	0.1241	0.1270
		Medium	(0.459, 0.508]	0.9943	0.0099	0.0100
		Hard	(0.508, 0.574]	0.9722	0.0832	0.0855
		Very Hard	(0.574, 0.779]	0.8844	0.2186	0.2471
	TABLET	Easy	(0.401, 0.462]	0.7982	0.4031	0.5049
		Medium	(0.462, 0.505]	0.8636	0.3447	0.3992
		Hard	(0.505, 0.575]	0.5893	0.4942	0.8386
		VeryHard	(0.575, 0.810]	0.3571	0.4813	1.3477
	TABLLM	Easy	(0.401, 0.462]	1.0000	0.0000	0.0000
		Medium	(0.462, 0.506]	0.7727	0.4291	0.5448
		Hard	(0.506, 0.575]	0.3831	0.3770	0.4540
		Very Hard	(0.575, 0.810]	0.3661	0.4839	1.3219
	MALLM-Zeroshot	Easy	(0.401, 0.462]	1.0000	0.0000	0.0000
		Medium	(0.462, 0.506]	0.7364	0.4426	0.6011
		Hard	(0.506, 0.575]	0.5089	0.5022	0.9867
		Very Hard	(0.575, 0.81]	0.3571	0.4813	1.3477
	MALLM-Fewshot5	Easy	(0.401, 0.462]	1.0000	0.0000	0.0000
		Medium	(0.462, 0.505]	0.7455	0.4376	0.5870
		Hard	(0.505, 0.575]	0.5000	0.5022	1.0045
		Very Hard	(0.575, 0.81]	0.3571	0.4813	1.3477
	MALLM-RAG10	Easy	(0.401, 0.462]	0.9386	0.2411	0.2569
		Medium	(0.462, 0.505]	0.8000	0.4018	0.5023
		Hard	(0.505, 0.575]	0.7589	0.4297	0.5661
		Very Hard	(0.575, 0.81]	0.4643	0.5010	1.0790

Table 5: Models Average Corrected Probability & Various Levels of Hardness — Employee Satisfaction & Students AI Tools Datasets

Dataset	Model	Hardness Category	Hardness Interval	Mean	Std	Coeff Var
Employee Satisfaction	AutoGluon	Easy	(0.409, 0.46]	0.9712	0.0158	0.0162
		Medium	(0.46, 0.493]	0.9386	0.0428	0.0456
		Hard	(0.493, 0.593]	0.8235	0.1550	0.1882
		Very Hard	(0.593, 0.792]	0.4742	0.2478	0.5225
	MLJAR	Easy	(0.409, 0.46]	0.9016	0.0227	0.0252
		Medium	(0.46, 0.493]	0.8682	0.0465	0.0536
		Hard	(0.493, 0.593]	0.7678	0.1261	0.1643
		Very Hard	(0.593, 0.795]	0.4750	0.1930	0.4064
	H2O	Easy	(0.409, 0.460]	0.9575	0.0205	0.0214
		Medium	(0.460, 0.493]	0.9210	0.0461	0.0500
		Hard	(0.493, 0.593]	0.8211	0.1595	0.1943
		Very Hard	(0.593, 0.792]	0.4638	0.2641	0.5694
	TABLET	Easy	(0.409, 0.460]	1.0000	0.0000	0.0000
		Medium	(0.460, 0.493]	0.9600	0.1973	0.2055
		Hard	(0.493, 0.592]	0.7867	0.2926	0.3729
		Very Hard	(0.592, 0.792]	0.4267	0.4979	1.1671
	TABLLM	Easy	(0.409, 0.460]	1.0000	0.0000	0.0000
		Medium	(0.460, 0.493]	1.0000	0.0000	0.0000
		Hard	(0.493, 0.593]	0.9200	0.2731	0.2969
		Very Hard	(0.593, 0.792]	0.4000	0.4932	1.2330
	MALLM-Zeroshot	Easy	(0.409, 0.46]	1.0000	0.0000	0.0000
		Medium	(0.46, 0.493]	0.5867	0.4957	0.8450
		Hard	(0.493, 0.592]	0.7600	0.4300	0.5657
		Very Hard	(0.592, 0.792]	0.3333	0.4746	1.4237
	MALLM-Fewshot5	Easy	(0.409, 0.46]	1.0000	0.0000	0.0000
		Medium	(0.46, 0.493]	0.6667	0.4746	0.7119
		Hard	(0.493, 0.593]	0.7467	0.4378	0.5864
		Very Hard	(0.593, 0.792]	0.4133	0.4957	1.1994
	MALLM-RAG10	Easy	(0.409, 0.46]	1.0000	0.0000	0.0000
		Medium	(0.46, 0.493]	0.9867	0.1155	0.1170
		Hard	(0.493, 0.592]	0.9200	0.2731	0.2969
		Very Hard	(0.592, 0.795]	0.4133	0.4957	1.1994
Students AI Tools	AutoGluon	Easy	(0.5, 0.587]	0.5383	0.0545	0.1013
		Medium	(0.587, 0.626]	0.5058	0.0730	0.1443
		Hard	(0.626, 0.659]	0.4928	0.0744	0.1510
		Very Hard	(0.659, 0.717]	0.4760	0.0649	0.1363
	MLJAR	Easy	(0.5, 0.587]	0.5489	0.0359	0.0653
		Medium	(0.587, 0.626]	0.5115	0.0593	0.1159
		Hard	(0.626, 0.658]	0.5013	0.0605	0.1207
		Very Hard	(0.658, 0.72]	0.4626	0.0479	0.1036
	H2O	Easy	(0.500, 0.587]	0.5804	0.1356	0.2337
		Medium	(0.587, 0.626]	0.5222	0.1770	0.3390
		Hard	(0.626, 0.658]	0.4965	0.1815	0.3655
		Very Hard	(0.658, 0.714]	0.4532	0.1583	0.3493
	TABLET	Easy	(0.5, 0.587]	0.5267	0.5090	0.9512
		Medium	(0.587, 0.626]	0.6267	0.4853	0.7744
		Hard	(0.626, 0.659]	0.6133	0.4886	0.7967
		Very Hard	(0.659, 0.792]	0.6000	0.4915	0.8192
	TABLLM	Easy	(0.5, 0.587]	0.1733	0.3798	2.1912
		Medium	(0.587, 0.626]	0.4133	0.4941	1.1954
		Hard	(0.626, 0.659]	0.5400	0.5007	0.9261
		Very Hard	(0.659, 0.723]	0.8400	0.3678	0.4379
	MALLM-Zeroshot	Easy	(0.5, 0.587]	0.9200	0.2722	0.2959
		Medium	(0.587, 0.627]	0.6067	0.4901	0.8079
		Hard	(0.627, 0.658]	0.5267	0.5010	0.9512
		Very Hard	(0.658, 0.717]	0.2267	0.4201	1.8533
	MALLM-Fewshot5	Easy	(0.5, 0.587]	0.9267	0.2616	0.2823
		Medium	(0.587, 0.626]	0.6133	0.4886	0.7967
		Hard	(0.626, 0.658]	0.5200	0.5013	0.9640
		Very Hard	(0.658, 0.723]	0.1600	0.3678	2.2990
	MALLM-RAG10	Easy	(0.5, 0.587]	0.9267	0.2616	0.2823
		Medium	(0.587, 0.626]	0.6000	0.4915	0.8192
		Hard	(0.626, 0.659]	0.5200	0.5013	0.9640
		Very Hard	(0.659, 0.715]	0.1733	0.3798	2.1912

Table 6: Models Average Corrected Probability & Various Levels of Hardness — Rental Properties & ICU Admission Datasets

Dataset	Model	Hardness Category	Hardness Interval	Mean	Std	Coeff Var
Rental Properties	AutoGluon	Easy	(0.301, 0.316]	0.9943	0.0018	0.0018
		Medium	(0.316, 0.388]	0.9632	0.0760	0.0789
		Hard	(0.388, 0.444]	0.9836	0.0236	0.0240
		Very Hard	(0.444, 0.667]	0.8760	0.1978	0.2258
	MLJAR	Easy	(0.301, 0.316]	0.9568	0.0250	0.0261
		Medium	(0.316, 0.388]	0.9319	0.0374	0.0402
		Hard	(0.388, 0.444]	0.8954	0.0807	0.0901
		Very Hard	(0.444, 0.667]	0.7042	0.2301	0.3267
	H2O	Easy	(0.301, 0.316]	1.0000	0.0000	0.0000
		Medium	(0.316, 0.388]	0.9670	0.0899	0.0930
		Hard	(0.388, 0.444]	0.9812	0.0727	0.0741
		Very Hard	(0.444, 0.667]	0.8893	0.2482	0.2791
	TABLET	Easy	(0.298, 0.315]	1.0000	0.0000	0.0000
		Medium	(0.315, 0.413]	1.0000	0.0000	0.0000
		Hard	(0.413, 0.462]	0.9444	0.2323	0.2460
		Very Hard	(0.462, 0.74]	0.5278	0.5063	0.9593
	TABLLM	Easy	(0.298, 0.315]	1.0000	0.0000	0.0000
		Medium	(0.315, 0.413]	0.5000	0.5075	1.0150
		Hard	(0.413, 0.462]	0.0833	0.2803	3.3637
		Very Hard	(0.462, 0.74]	0.4722	0.5063	1.0722
	MALLM-Zeroshot	Easy	(0.298, 0.315]	0.9231	0.2700	0.2924
		Medium	(0.315, 0.413]	0.9706	0.1715	0.1767
		Hard	(0.413, 0.462]	0.9444	0.2323	0.2460
		Very Hard	(0.462, 0.74]	0.5000	0.5071	1.0142
	MALLM-Fewshot5	Easy	(0.298, 0.315]	1.0000	0.0000	0.0000
		Medium	(0.315, 0.413]	0.9706	0.1715	0.1767
		Hard	(0.413, 0.462]	0.9444	0.2323	0.2460
		Very Hard	(0.462, 0.74]	0.5556	0.5040	0.9071
	MALLM-RAG10	Easy	(0.298, 0.315]	1.0000	0.0000	0.0000
		Medium	(0.315, 0.405]	1.0000	0.0000	0.0000
		Hard	(0.405, 0.452]	1.0000	0.0000	0.0000
		Very Hard	(0.452, 0.774]	0.8611	0.3507	0.4073
ICU Admission	AutoGluon	Easy	(0.436, 0.486]	0.8493	0.0735	0.0866
		Medium	(0.486, 0.51]	0.7390	0.1718	0.2324
		Hard	(0.51, 0.591]	0.7305	0.1665	0.2280
		Very Hard	(0.591, 0.754]	0.4509	0.2041	0.4526
	MLJAR	Easy	(0.436, 0.486]	0.6684	0.0329	0.0492
		Medium	(0.486, 0.51]	0.6488	0.0528	0.0814
		Hard	(0.51, 0.59]	0.6598	0.0395	0.0598
		Very Hard	(0.59, 0.754]	0.4569	0.1220	0.2671
	H2O	Easy	(0.436, 0.486]	0.9055	0.0827	0.0913
		Medium	(0.486, 0.511]	0.8091	0.1821	0.2250
		Hard	(0.511, 0.591]	0.8332	0.1318	0.1582
		Very Hard	(0.591, 0.754]	0.4415	0.2836	0.6423
	TABLET	Easy	(0.436, 0.486]	0.8000	0.4104	0.5130
		Medium	(0.486, 0.51]	0.7895	0.4189	0.5305
		Hard	(0.51, 0.59]	0.6316	0.4956	0.7847
		Very Hard	(0.59, 0.754]	0.4211	0.5073	1.2047
	TABLLM	Easy	(0.436, 0.486]	1.0000	0.0000	0.0000
		Medium	(0.486, 0.51]	0.3684	0.4956	1.3452
		Hard	(0.51, 0.59]	0.5789	0.5073	0.8762
		Very Hard	(0.59, 0.754]	0.3684	0.4956	1.3452
	MALLM-Zeroshot	Easy	(0.436, 0.486]	0.9500	0.2236	0.2354
		Medium	(0.486, 0.51]	0.8421	0.3746	0.4449
		Hard	(0.51, 0.59]	0.7895	0.4189	0.5305
		Very Hard	(0.59, 0.754]	0.2632	0.4524	1.7192
	MALLM-Fewshot5	Easy	(0.436, 0.486]	0.9000	0.3078	0.3420
		Medium	(0.486, 0.51]	0.8421	0.3746	0.4449
		Hard	(0.51, 0.59]	0.6316	0.4956	0.7847
		Very Hard	(0.59, 0.754]	0.2632	0.4524	1.7192
	MALLM-RAG10	Easy	(0.436, 0.486]	0.9500	0.2236	0.2354
		Medium	(0.486, 0.51]	0.7895	0.4189	0.5305
		Hard	(0.51, 0.59]	0.6842	0.4776	0.6980
		Very Hard	(0.59, 0.754]	0.3158	0.4776	1.5123

Appendix D: Typicality Analysis

Typicality measures how representative each observation is relative to the dataset's average pattern, with higher scores indicating more homogeneous and aligned data.

As shown in Table 7, Employee Satisfaction exhibits the highest typicality (0.9689), indicating highly uniform data, while Rental Properties shows the lowest (0.7468), reflecting greater heterogeneity. Student Depression and Students' AI Tools fall in the mid-range, suggesting a balance between common patterns and variation.

Next, across models in Table 8, AutoML models consistently leverage typicality to improve performance. They show strong positive correlations between typicality and corrected probabilities, with reliability increasing monotonically as data becomes more typical. This pattern is especially pronounced in Student Depression and Employee Satisfaction. In contrast, table-tuned LLMs (TABLET, TABLLM) exhibit weaker and less stable relationships with typicality. TABLET often shows high variability and even performance declines at higher typicality levels, while TABLLM demonstrates modest improvements but lacks consistent gains.

Among MALLM models, RAG10 stands out as the only approach that effectively benefits from typicality across multiple datasets, achieving performance comparable to AutoML systems in high-typicality settings. However, zero-shot and few-shot variants remain unstable and show limited ability to systematically exploit typicality.

Overall, results indicate that AutoML models are the most reliable in leveraging typicality, while RAG-based approaches offer competitive performance in certain contexts. In contrast, table-tuned and simpler LLM variants exhibit greater volatility and weaker sensitivity to typicality, particularly in low-typicality settings where performance differences across models become more pronounced.

Tables 9 and Table 10 further show that AutoML models exhibit consistent monotonic improvements as typicality increases. Across datasets, corrected probabilities rise steadily from low to very high typicality levels, indicating that these models effectively leverage more representative data to improve reliability.

In contrast, TABLET and TABLLM models display greater variability. TABLET often shows unstable or even declining performance at higher typicality levels, while TABLLM exhibits modest gains but lacks consistent monotonic trends.

Among MALLM approaches, RAG10 demonstrates meaningful gains with increasing typicality, particularly in Employee Satisfaction and Students' AI Tools, where it achieves performance comparable to AutoML models at higher typicality levels. However, zero-shot and few-shot variants remain unstable and fail to systematically benefit from typicality.

At lower typicality levels, performance differences across models become more pronounced. AutoML models maintain relatively higher reliability despite declines, whereas LLM-based approaches—especially TABLET—show substantial variability and occasional performance reversals.

Table 7: Average Typicality Score for Post-Cutoff Datasets

Dataset	Score
Student Depression	0.8144
Postpartum Depression	0.7617
Employee Satisfaction	0.9689
Students AI Tools	0.8149
Rental Properties	0.7468
ICU Admission	0.7719

Appendix E: Analysis of Pre-Cutoff Datasets

Table 12 summarizes the five pre-cutoff benchmark datasets including Adult, Titanic, Heart Disease, Breast Cancer, and Diabetes. These datasets are widely used for evaluating tabular learning models and are drawn from UCI and Kaggle repositories.

As shown in Table 14, MALLMs frequently outperform AutoML baselines on these datasets, achieving higher AUC, accuracy, and F1 scores across several tasks. Their advantage is particularly pronounced in relatively balanced datasets such as Adult, Titanic, and Diabetes. While TABLET and TABLLM remain competitive in some cases, they generally lag behind both AutoML and Multi-Agent configurations. These results suggest that, prior to the cutoff, retrieval-augmented and few-shot prompting strategies can effectively leverage latent knowledge.

However, Table 15 reveals an important limitation. Across all models, correlations between instance hardness and corrected probabilities are negative, indicating appropriate confidence adjustment as task difficulty increases. AutoML systems exhibit the strongest and most consistent negative

Table 8: Correlation between Algorithm Typicality & Corrected Probability for Post-Cutoff Datasets

Dataset	AutoGluon	MLJAR	H2O	TABLET	TABLLM	MALLM-Zeroshot	MALLM-Fewshot5	MALLM-RAG10
Student Depression	0.4467	0.4467	0.4470	-0.0277	0.0781	0.1510	0.3027	0.3051
Postpartum Depression	0.0347	0.0710	0.0345	-0.0050	-0.0541	-0.1058	-0.0982	0.0557
Employee Satisfaction	0.8767	0.8806	0.8483	0.7129	0.7577	0.3083	0.3125	0.6564
Students AI Tools	0.0349	0.0453	0.0162	-0.0049	-0.0473	0.0702	0.0827	0.0824
Rental Properties	0.3850	0.6439	0.2978	0.1106	0.1233	0.0962	-0.0100	0.0658
ICU Admission	-0.2068	-0.0292	-0.1313	0.2976	-0.4137	0.1932	0.1297	0.0345

The correlations are statistically significant at the 0.01% level.

correlations, reflecting robust calibration. In contrast, TABLET and TABLLM show weaker relationships, and MALLMs display the shallowest correlations, suggesting reduced sensitivity to instance difficulty and a tendency toward overconfidence.

Taken together, these findings highlight a key trade-off: although Multi-Agent LLMs can achieve strong predictive performance on pre-cutoff benchmarks, they lack the reliable confidence calibration demonstrated by AutoML models when facing harder instances. One possible explanation is data contamination, as these widely used benchmark datasets may have been included in the pretraining corpus of large language models, potentially inflating performance. This interpretation underscores the importance of evaluating models on post-cutoff or less-exposed datasets to assess true generalization.

Table 9: Models Average Corrected Probability & Various Levels of Typicality — Student Depression & Postpartum Depression Datasets

Dataset	Model	Typicality Category	Typicality Interval	Mean	Std	Coeff Var
Student Depression	AutoGluon	Low	(-0.385, 0.255]	0.6153	0.2935	0.4769
		Medium	(0.255, 0.46]	0.7227	0.2759	0.3818
		High	(0.46, 0.584]	0.8367	0.1556	0.1860
		Very High	(0.584, 0.853]	0.9051	0.1112	0.1228
	MLJAR	Low	(-0.385, 0.255]	0.6198	0.3199	0.5161
		Medium	(0.255, 0.46]	0.7373	0.2870	0.3893
		High	(0.46, 0.584]	0.8541	0.1692	0.1981
		Very High	(0.584, 0.853]	0.9301	0.1192	0.1281
	H2O	Low	(-0.385, 0.255]	0.6181	0.3127	0.5060
		Medium	(0.255, 0.460]	0.7335	0.2809	0.3829
		High	(0.460, 0.584]	0.8461	0.1666	0.1969
		Very High	(0.584, 0.853]	0.9238	0.1198	0.1297
	TABLET	Low	(0.0596, 0.379]	0.6933	0.4642	0.6695
		Medium	(0.379, 0.486]	0.6267	0.4869	0.7770
		High	(0.486, 0.598]	0.6533	0.4791	0.7333
		Very High	(0.598, 0.762]	0.6267	0.4869	0.7770
	TABLLM	Low	(0.0596, 0.379]	0.6133	0.4903	0.7993
		Medium	(0.379, 0.486]	0.7067	0.4584	0.6486
		High	(0.486, 0.598]	0.7733	0.4215	0.5450
		Very High	(0.598, 0.762]	0.9200	0.2731	0.2969
	MALLM-Zeroshot	Low	(0.0596, 0.379]	0.6267	0.4869	0.7770
		Medium	(0.379, 0.486]	0.6933	0.4642	0.6695
		High	(0.486, 0.598]	0.7600	0.4299	0.5657
		Very High	(0.598, 0.762]	0.8000	0.4027	0.5034
	MALLM-Fewshot5	Low	(0.0596, 0.379]	0.5467	0.5012	0.9168
		Medium	(0.379, 0.486]	0.7467	0.4378	0.5864
		High	(0.486, 0.598]	0.8000	0.4027	0.5034
		Very High	(0.598, 0.762]	0.8800	0.3272	0.3718
	MALLM-RAG10	Low	(0.0596, 0.379]	0.6533	0.4791	0.7333
		Medium	(0.379, 0.486]	0.7067	0.4584	0.6486
		High	(0.486, 0.598]	0.8533	0.3562	0.4174
		Very High	(0.598, 0.762]	0.9467	0.2262	0.2390
Postpartum Depression	AutoGluon	Low	(-0.386, 0.106]	0.8941	0.1697	0.1898
		Medium	(0.106, 0.333]	0.9401	0.0805	0.0856
		High	(0.333, 0.617]	0.8997	0.1561	0.1735
		Very High	(0.617, 0.936]	0.9034	0.0979	0.1084
	MLJAR	Low	(-0.386, 0.106]	0.8700	0.2229	0.2562
		Medium	(0.106, 0.333]	0.9418	0.1507	0.1600
		High	(0.333, 0.617]	0.9147	0.1579	0.1726
		Very High	(0.617, 0.936]	0.8814	0.1790	0.2031
	H2O	Low	(-0.386, 0.106]	0.9443	0.1796	0.1902
		Medium	(0.106, 0.333]	0.9797	0.0745	0.0761
		High	(0.333, 0.617]	0.9615	0.1172	0.1219
		Very High	(0.617, 0.936]	0.9424	0.1582	0.1679
	TABLET	Low	(0.357, 0.507]	0.4643	0.5006	1.0790
		Medium	(0.507, 0.541]	0.6518	0.4786	0.7342
		High	(0.541, 0.598]	0.8333	0.3743	0.4492
		Very High	(0.598, 0.719]	0.6545	0.4777	0.7298
	TABLLM	Low	(0.357, 0.507]	0.4196	0.4957	1.1813
		Medium	(0.507, 0.541]	0.7321	0.4449	0.6076
		High	(0.541, 0.598]	0.9123	0.2841	0.3115
		Very High	(0.598, 0.719]	0.9091	0.2888	0.3177
	MALLM-Zeroshot	Low	(0.357, 0.507]	0.5536	0.4994	0.9021
		Medium	(0.507, 0.541]	0.8571	0.3515	0.4101
		High	(0.541, 0.598]	0.8596	0.3489	0.4059
		Very High	(0.598, 0.719]	0.3273	0.4714	1.4403
	MALLM-Fewshot5	Low	(0.357, 0.507]	0.5625	0.4983	0.8859
		Medium	(0.507, 0.541]	0.8571	0.3515	0.4101
		High	(0.541, 0.598]	0.8588	0.3578	0.4205
		Very High	(0.598, 0.719]	0.3273	0.4714	1.4403
	MALLM-RAG10	Low	(0.357, 0.507]	0.5804	0.4957	0.8542
		Medium	(0.507, 0.541]	0.8839	0.3218	0.3640
		High	(0.541, 0.598]	0.8947	0.3082	0.3445
		Very High	(0.598, 0.719]	0.6455	0.4806	0.7446

Table 10: Models Average Corrected Probability & Various Levels of Typicality — Employee Satisfaction & Students AI Tools Datasets

Dataset	Model	Typicality Category	Typicality Interval	Mean	Std	Coeff Var
Employee Satisfaction	AutoGluon	Low	(-0.193, 0.454]	0.4591	0.2344	0.5105
		Medium	(0.454, 0.727]	0.8482	0.1371	0.1616
		High	(0.727, 0.854]	0.9333	0.0668	0.0716
		Very High	(0.854, 0.936]	0.9669	0.0162	0.0167
	MLJAR	Low	(-0.193, 0.454]	0.4696	0.1880	0.4004
		Medium	(0.454, 0.727]	0.7794	0.1207	0.1549
		High	(0.727, 0.854]	0.8651	0.0603	0.0697
		Very High	(0.854, 0.936]	0.8984	0.0203	0.0226
	H2O	Low	(-0.193, 0.454]	0.4585	0.2577	0.5619
		Medium	(0.454, 0.727]	0.8374	0.1558	0.1861
		High	(0.727, 0.854]	0.9177	0.0703	0.0766
		Very High	(0.854, 0.936]	0.9498	0.0267	0.0282
	TABLET	Low	(-0.186, 0.445]	0.3867	0.4903	1.2679
		Medium	(0.445, 0.721]	0.9467	0.2262	0.2390
		High	(0.721, 0.847]	0.9867	0.1155	0.1170
		Very High	(0.847, 0.928]	0.9733	0.1622	0.1666
	TABLLM	Low	(-0.186, 0.445]	0.3733	0.4870	1.3043
		Medium	(0.445, 0.721]	0.9467	0.2262	0.2390
		High	(0.721, 0.847]	1.0000	0.0000	0.0000
		Very High	(0.847, 0.928]	1.0000	0.0000	0.0000
	MALLM-Zeroshot	Low	(-0.186, 0.445]	0.4533	0.5012	1.1055
		Medium	(0.445, 0.721]	0.7733	0.4215	0.5450
		High	(0.721, 0.847]	0.7333	0.4452	0.6071
		Very High	(0.847, 0.928]	0.7200	0.4520	0.6278
	MALLM-Fewshot5	Low	(-0.186, 0.445]	0.4933	0.5033	1.0202
		Medium	(0.445, 0.721]	0.7733	0.4215	0.5450
		High	(0.721, 0.847]	0.7867	0.4124	0.5243
		Very High	(0.847, 0.928]	0.7733	0.4215	0.5450
	MALLM-RAG10	Low	(-0.186, 0.445]	0.4400	0.4997	1.1357
		Medium	(0.445, 0.721]	0.9067	0.2929	0.3230
		High	(0.721, 0.847]	0.9733	0.1622	0.1666
		Very High	(0.847, 0.928]	1.0000	0.0000	0.0000
Students AI Tools	AutoGluon	Low	(-0.409, 0.17]	0.5034	0.0731	0.1451
		Medium	(0.17, 0.384]	0.5029	0.0705	0.1401
		High	(0.384, 0.542]	0.4922	0.0759	0.1543
		Very High	(0.542, 0.814]	0.5143	0.0619	0.1203
	MLJAR	Low	(-0.409, 0.17]	0.5023	0.0593	0.1181
		Medium	(0.17, 0.384]	0.5095	0.0577	0.1133
		High	(0.384, 0.542]	0.4993	0.0643	0.1287
		Very High	(0.542, 0.814]	0.5132	0.0588	0.1145
	H2O	Low	(-0.409, 0.170]	0.5040	0.1786	0.3543
		Medium	(0.170, 0.384]	0.5220	0.1653	0.3166
		High	(0.384, 0.542]	0.5168	0.1753	0.3392
		Very High	(0.542, 0.814]	0.5094	0.1616	0.3173
	TABLET	Low	(0.101, 0.336]	0.5333	0.5006	0.9385
		Medium	(0.336, 0.448]	0.6733	0.4707	0.6989
		High	(0.448, 0.517]	0.5467	0.4995	0.9137
		Very High	(0.517, 0.667]	0.5667	0.4972	0.8774
	TABLLM	Low	(0.101, 0.336]	0.5000	0.5017	1.0034
		Medium	(0.336, 0.448]	0.5533	0.4988	0.9015
		High	(0.448, 0.517]	0.4400	0.4981	1.1319
		Very High	(0.517, 0.667]	0.4733	0.5000	1.0584
	MALLM-Zeroshot	Low	(0.101, 0.336]	0.5467	0.4995	0.9137
		Medium	(0.336, 0.448]	0.5133	0.5015	0.9769
		High	(0.448, 0.517]	0.6267	0.4853	0.7744
		Very High	(0.517, 0.667]	0.5933	0.4929	0.8307
	MALLM-Fewshot5	Low	(0.101, 0.336]	0.5200	0.5013	0.9639
		Medium	(0.336, 0.448]	0.5133	0.5015	0.9769
		High	(0.448, 0.517]	0.6067	0.4901	0.8079
		Very High	(0.517, 0.667]	0.5800	0.4952	0.8538
	MALLM-RAG10	Low	(0.101, 0.336]	0.5200	0.5013	0.9639
		Medium	(0.336, 0.448]	0.5133	0.5015	0.9769
		High	(0.448, 0.517]	0.6067	0.4901	0.8079
		Very High	(0.517, 0.667]	0.5800	0.4952	0.8538

Table 11: Models Average Corrected Probability & Various Levels of Typicality — Rental Properties & ICU Admission Datasets

Dataset	Model	Typicality Category	Typicality Interval	Mean	Std	Coeff Var
Rental Properties	AutoGluon	Low	(-0.285, 0.41]	0.8847	0.1978	0.2236
		Medium	(0.41, 0.613]	0.9492	0.0760	0.0801
		High	(0.613, 0.787]	0.9900	0.0174	0.0176
		Very High	(0.787, 0.909]	0.9957	0.0010	0.0010
	MLJAR	Low	(-0.285, 0.41]	0.7189	0.2327	0.3236
		Medium	(0.41, 0.613]	0.8725	0.0875	0.1003
		High	(0.613, 0.787]	0.9458	0.0271	0.0286
		Very High	(0.787, 0.909]	0.9588	0.0096	0.0100
	H2O	Low	(-0.285, 0.41]	0.9039	0.2409	0.2665
		Medium	(0.41, 0.613]	0.9370	0.1249	0.1333
		High	(0.613, 0.787]	0.9989	0.0034	0.0034
		Very High	(0.787, 0.909]	1.0000	0.0000	0.0000
	TABLET	Low	(-0.281, 0.398]	0.7838	0.4173	0.5325
		Medium	(0.398, 0.659]	0.8919	0.3148	0.3530
		High	(0.659, 0.807]	0.9444	0.2323	0.2460
		Very High	(0.807, 0.909]	0.8571	0.3550	0.4142
	TABLLM	Low	(-0.281, 0.398]	0.4865	0.5067	1.0416
		Medium	(0.398, 0.659]	0.4324	0.5022	1.1614
		High	(0.659, 0.807]	0.5278	0.5063	0.9593
		Very High	(0.807, 0.909]	0.6571	0.4816	0.7329
	MALLM-Zeroshot	Low	(-0.281, 0.398]	0.7568	0.4350	0.5748
		Medium	(0.398, 0.659]	0.8649	0.3466	0.4007
		High	(0.659, 0.807]	0.8889	0.3187	0.3586
		Very High	(0.807, 0.909]	0.8286	0.3824	0.4615
	MALLM-Fewshot5	Low	(-0.281, 0.398]	0.8649	0.3466	0.4007
		Medium	(0.398, 0.659]	0.8919	0.3148	0.3530
		High	(0.659, 0.807]	0.8611	0.3507	0.4073
		Very High	(0.807, 0.909]	0.8571	0.3550	0.4142
	MALLM-RAG10	Low	(-0.281, 0.398]	0.9459	0.2292	0.2423
		Medium	(0.398, 0.659]	0.9459	0.2292	0.2423
		High	(0.659, 0.807]	0.9722	0.1667	0.1714
		Very High	(0.807, 0.909]	1.0000	0.0000	0.0000
ICU Admission	AutoGluon	Low	(-0.225, 0.244]	0.7574	0.2047	0.2703
		Medium	(0.244, 0.403]	0.7154	0.2066	0.2888
		High	(0.403, 0.547]	0.6352	0.2666	0.4196
		Very High	(0.547, 0.871]	0.6665	0.1730	0.2595
	MLJAR	Low	(-0.225, 0.244]	0.6273	0.1163	0.1854
		Medium	(0.244, 0.403]	0.6160	0.0808	0.1312
		High	(0.403, 0.547]	0.5589	0.1512	0.2706
		Very High	(0.547, 0.871]	0.6338	0.0756	0.1193
	H2O	Low	(-0.225, 0.244]	0.8334	0.1917	0.2300
		Medium	(0.244, 0.403]	0.7453	0.2423	0.3251
		High	(0.403, 0.547]	0.6532	0.3556	0.5444
		Very High	(0.547, 0.871]	0.7611	0.1864	0.2449
	TABLET	Low	(-0.225, 0.244]	0.5000	0.5130	1.0260
		Medium	(0.244, 0.403]	0.7368	0.4524	0.6140
		High	(0.403, 0.547]	0.5263	0.5130	0.9747
		Very High	(0.547, 0.871]	0.8947	0.3153	0.3524
	TABLLM	Low	(-0.225, 0.244]	0.7000	0.4702	0.6717
		Medium	(0.244, 0.403]	0.7368	0.4524	0.6140
		High	(0.403, 0.547]	0.7368	0.4524	0.6140
		Very High	(0.547, 0.871]	0.1579	0.3746	2.3727
	MALLM-Zeroshot	Low	(-0.225, 0.244]	0.6500	0.4894	0.7529
		Medium	(0.244, 0.403]	0.6842	0.4776	0.6980
		High	(0.403, 0.547]	0.5789	0.5073	0.8762
		Very High	(0.547, 0.871]	0.9474	0.2294	0.2422
	MALLM-Fewshot5	Low	(-0.225, 0.244]	0.6500	0.4894	0.7529
		Medium	(0.244, 0.403]	0.6316	0.4956	0.7847
		High	(0.403, 0.547]	0.5263	0.5130	0.9747
		Very High	(0.547, 0.871]	0.8421	0.3746	0.4449
	MALLM-RAG10	Low	(-0.225, 0.244]	0.6500	0.4894	0.7529
		Medium	(0.244, 0.403]	0.7895	0.4189	0.5305
		High	(0.403, 0.547]	0.5263	0.5130	0.9747
		Very High	(0.547, 0.871]	0.7895	0.4189	0.5305

Table 12: Overview of Pre-Cutoff Datasets

Dataset	*Classes*	*Instances*	*Tot. Feat.*	*Cat. Feat.*	*Num. Feat.*	*Class Dist.*	*Source*
Heart Disease	2	303	13	5	8	54% No Disease, 46% Disease	UCI
Titanic	2	1309	12	3	4	38% Survived, 62% Not Survived	Kaggle
Breast Cancer	2	569	30	0	30	63% Benign, 37% Malignant	UCI
Diabetes	2	768	8	0	8	35% Positive, 65% Negative	UCI
Adult	2	1000	14	8	6	24% >\$50K, 76% $\leq$\$50K	UCI

Table 13: Models Performance Metrics Across Post-Cutoff Datasets

Models	Log Loss	AUC	Accuracy	Recall	Pos. Recall	Neg. Recall	Precision	Pos. Prec.	Neg. Prec.	F1 Score	Best Thresh.
STUDENT DEPRESSION											
AutoGlueon-S1 (1000)	0.3665	0.9115	0.8333	0.8857	0.8857	0.7600	0.8378	0.8378	0.8261	0.8611	50%
AutoGlueon-S2 (500)	0.3872 (0.0089)	0.9023 (0.0046)	0.8307 (0.0043)	0.8971 (0.0114)	0.8971 (0.0114)	0.7376 (0.0119)	0.8272 (0.0052)	0.8272 (0.0052)	0.8370 (0.0134)	0.8607 (0.0043)	50%
AutoGlueon-S3 (200)	0.4146 (0.0188)	0.8901 (0.0082)	0.8173 (0.0119)	0.8971 (0.0359)	0.8971 (0.0359)	0.7056 (0.0409)	0.8107 (0.0161)	0.8107 (0.0161)	0.8338 (0.0424)	0.8512 (0.0119)	50%
MLJAR-S1 (1000)	0.3718	0.9100	0.8333	0.8743	0.8743	0.7760	0.8453	0.8453	0.8151	0.8596	50%
MLJAR-S2 (500)	0.3865 (0.0105)	0.9044 (0.0031)	0.8313 (0.0084)	0.8823 (0.0118)	0.8823 (0.0118)	0.7600 (0.0226)	0.8375 (0.0120)	0.8375 (0.0120)	0.8220 (0.0131)	0.8592 (0.0066)	50%
MLJAR-S3 (200)	0.4296 (0.0515)	0.8951 (0.0162)	0.8200 (0.0189)	0.8971 (0.0410)	0.8971 (0.0410)	0.7120 (0.0672)	0.8150 (0.0296)	0.8150 (0.0296)	0.8370 (0.0549)	0.8532 (0.0152)	50%
H2O-S1 (1000)	0.3697	0.9103	0.8333	0.8743	0.8743	0.7760	0.8453	0.8453	0.8151	0.8596	50%
H2O-S2 (500)	0.3976 (0.0484)	0.9001 (0.0137)	0.8213 (0.0214)	0.8777 (0.0456)	0.8777 (0.0456)	0.7424 (0.0249)	0.8269 (0.0104)	0.8269 (0.0104)	0.8162 (0.0522)	0.8510 (0.0219)	50%
H2O-S3 (200)	0.4919 (0.1801)	0.8821 (0.0253)	0.8133 (0.0316)	0.8800 (0.0632)	0.8800 (0.0632)	0.7200 (0.0388)	0.8152 (0.0174)	0.8152 (0.0174)	0.8171 (0.0687)	0.8453 (0.0318)	50%
TABLET	0.6470	0.6717	0.6500	0.8286	0.8286	0.4000	0.6591	0.6591	0.6250	0.7342	50%
TABLLM	0.6547	0.7343	0.7500	0.7343	0.8286	0.6400	0.7452	0.7632	0.7273	0.7377	50%
MALLM-Zeroshot	0.5824	0.8011	0.7667	0.7509	0.8457	0.6560	0.7636	0.7749	0.7523	0.8087	0.66
MALLM-Fewshot3	0.5175	0.8053	0.7866	0.7383	0.9200	0.6000	0.8028	0.7630	0.8426	0.8341	0.30
MALLM-Fewshot5	0.5088	0.8113	0.7900	0.7662	0.9085	0.6240	0.8008	0.7718	0.8297	0.8346	0.30
MALLM-RAG5	0.5788	0.8352	0.7900	0.7560	0.9600	0.5520	0.8289	0.7500	0.9078	0.8421	0.65
MALLM-RAG10	0.8005	0.8419	0.7900	0.7663	0.9086	0.6240	0.8008	0.7718	0.8298	0.8346	0.46
POSTPARTUM DEPRESSION											
AutoGlueon-S1 (1491)	0.0950	0.9968	0.9799	0.9828	0.9828	0.9745	0.9862	0.9862	0.9684	0.9845	50%
AutoGlueon-S2 (746)	0.1978 (0.0254)	0.9794 (0.0074)	0.9232 (0.0104)	0.9306 (0.0082)	0.9306 (0.0082)	0.9096 (0.0227)	0.9503 (0.0120)	0.9503 (0.0120)	0.8761 (0.0136)	0.9403 (0.0080)	50%
AutoGlueon-S3 (298)	0.3192 (0.0273)	0.9422 (0.0035)	0.8629 (0.0120)	0.8866 (0.0121)	0.8866 (0.0121)	0.8191 (0.0287)	0.9010 (0.0142)	0.9010 (0.0142)	0.7960 (0.0178)	0.8937 (0.0091)	50%
MLJAR-S1 (1491)	0.1451	0.9874	0.9219	0.9622	0.9622	0.8471	0.9211	0.9211	0.9236	0.9412	50%
MLJAR-S2 (746)	0.2073 (0.0285)	0.9756 (0.0114)	0.9295 (0.0026)	0.9533 (0.0280)	0.9533 (0.0280)	0.8854 (0.0322)	0.9393 (0.0158)	0.9393 (0.0158)	0.9131 (0.0485)	0.9460 (0.0163)	50%
MLJAR-S3 (298)	0.3115 (0.1022)	0.9578 (0.0193)	0.9094 (0.0295)	0.9430 (0.0180)	0.9430 (0.0180)	0.8471 (0.0592)	0.9200 (0.0298)	0.9200 (0.0298)	0.8887 (0.0351)	0.9312 (0.0220)	50%
H2O-S1 (1491)	0.0677	0.9976	0.9308	0.8935	0.8935	1.0000	1.0000	1.0000	0.8351	0.9437	50%
H2O-S2 (746)	0.1869 (0.0294)	0.9802 (0.0069)	0.8714 (0.0464)	0.8213 (0.0881)	0.8213 (0.0881)	0.9643 (0.0490)	0.9794 (0.0265)	0.9794 (0.0265)	0.7546 (0.0833)	0.8904 (0.0471)	50%
H2O-S3 (298)	0.3780 (0.1499)	0.9280 (0.0447)	0.8196 (0.0317)	0.7924 (0.0779)	0.7924 (0.0779)	0.8701 (0.1289)	0.9266 (0.0634)	0.9266 (0.0634)	0.7005 (0.0510)	0.8499 (0.0324)	50%
TABLET	0.7148	0.5256	0.6518	0.8076	0.8076	0.3631	0.7015	0.7015	0.5044	0.7508	50%
TABLLM	0.4732	0.7831	0.8326	0.7831	0.9485	0.6178	0.8438	0.8214	0.8661	0.8008	50%
MALLM-Zeroshot	0.6778	0.5258	0.6496	0.5000	1.0000	0.0000	0.3248	0.6496	0.0000	0.7876	0.10
MALLM-Fewshot3	0.7222	0.5377	0.6496	0.5000	1.0000	0.0000	0.3248	0.6496	0.0000	0.7876	0.10
MALLM-Fewshot5	0.7222	0.5870	0.6517	0.5031	0.0063	0.8255	0.8255	0.6510	1.0000	0.7880	0.11
MALLM-RAG5	0.6321	0.7739	0.7410	0.6745	0.8969	0.4522	0.7276	0.7521	0.7029	0.8181	0.40
MALLM-RAG10	0.6206	0.8093	0.7522	0.6523	0.9862	0.3185	0.8271	0.7284	0.9259	0.8379	0.11
EMPLOYEE SATISFACTION											
AutoGlueon-S1 (1000)	0.3075	0.9459	0.8433	0.7905	0.7905	0.8947	0.8797	0.8797	0.8144	0.8327	50%
AutoGlueon-S2 (500)	0.3354 (0.0173)	0.9379 (0.0048)	0.8380 (0.0051)	0.8054 (0.0175)	0.8054 (0.0175)	0.8697 (0.0164)	0.8579 (0.0128)	0.8579 (0.0128)	0.8214 (0.0112)	0.8306 (0.0062)	50%
AutoGlueon-S3 (200)	0.3702 (0.0306)	0.9253 (0.0127)	0.8280 (0.0206)	0.7919 (0.0193)	0.7919 (0.0193)	0.8632 (0.0261)	0.8496 (0.0265)	0.8496 (0.0265)	0.8099 (0.0175)	0.8196 (0.0210)	50%
MLJAR-S1 (1000)	0.3416	0.9436	0.8533	0.7905	0.7905	0.9145	0.9000	0.9000	0.8176	0.8417	50%
MLJAR-S2 (500)	0.3562 (0.0245)	0.9389 (0.0040)	0.8367 (0.0135)	0.7878 (0.0347)	0.7878 (0.0347)	0.8842 (0.0257)	0.8697 (0.0224)	0.8697 (0.0224)	0.8114 (0.0220)	0.8261 (0.0168)	50%
MLJAR-S3 (200)	0.3658 (0.0265)	0.9317 (0.0085)	0.8273 (0.0249)	0.7392 (0.0855)	0.7392 (0.0855)	0.9132 (0.0487)	0.8981 (0.0484)	0.8981 (0.0484)	0.7860 (0.0430)	0.8064 (0.0427)	50%
H2O-S1 (1000)	0.3295	0.9398	0.8400	0.8243	0.8243	0.8553	0.8472	0.8472	0.8333	0.8356	50%
H2O-S2 (500)	0.3444 (0.0090)	0.9309 (0.0061)	0.8427 (0.0210)	0.8635 (0.0504)	0.8635 (0.0504)	0.8224 (0.0799)	0.8310 (0.0572)	0.8310 (0.0572)	0.8643 (0.0352)	0.8444 (0.0153)	50%
H2O-S3 (200)	0.7367 (0.7862)	0.9154 (0.0155)	0.8340 (0.0121)	0.7608 (0.0533)	0.7608 (0.0533)	0.9053 (0.0410)	0.8899 (0.0400)	0.8899 (0.0400)	0.7971 (0.0295)	0.8183 (0.0194)	50%
TABLET	0.9790	0.5208	0.4000	1.0000	1.0000	0.0000	0.4000	0.4000	0.0000	0.5714	50%
TABLLM	0.4789	0.8292	0.8300	0.8292	0.7703	0.8882	0.8345	0.8702	0.7988	0.8292	50%
MALLM-Zeroshot	1.0349	0.8379	0.7767	0.7777	0.8581	0.6974	0.7844	0.7341	0.8346	0.7913	0.86
MALLM-Fewshot3	1.0943	0.8414	0.7066	0.7096	0.9324	0.4868	0.7599	0.6388	0.8809	0.7582	0.85
MALLM-Fewshot5	1.4228	0.8536	0.7066	0.7095	0.9256	0.4934	0.7561	0.6401	0.8720	0.7569	0.85
MALLM-RAG5	1.1071	0.8789	0.8166	0.8165	0.8040	0.8289	0.8167	0.8206	0.8129	0.8122	0.40
MALLM-RAG10	0.6979	0.9123	0.8300	0.8304	0.8648	0.7960	0.8315	0.8050	0.8581	0.8338	0.30
STUDENTS AI TOOLS											
AutoGlueon-S1 (1000)	0.6970	0.4850	0.5300	0.7801	0.7801	0.2201	0.5534	0.5534	0.4470	0.6475	50%
AutoGlueon-S2 (500)	0.7031 (0.0124)	0.4965 (0.0095)	0.5270 (0.0159)	0.7422 (0.0816)	0.7422 (0.0816)	0.2604 (0.0709)	0.5539 (0.0071)	0.5539 (0.0071)	0.4515 (0.0192)	0.6330 (0.0341)	50%
AutoGlueon-S3 (200)	0.7051 (0.0075)	0.4902 (0.0138)	0.5243 (0.0287)	0.6837 (0.2040)	0.6837 (0.2040)	0.3269 (0.1902)	0.5563 (0.0059)	0.5563 (0.0059)	0.4813 (0.0646)	0.6029 (0.0873)	50%
MLJAR-S1 (1000)	0.6883	0.5072	0.5300	0.6235	0.6235	0.4142	0.5687	0.5687	0.4703	0.5948	50%
MLJAR-S2 (500)	0.7019 (0.0122)	0.4879 (0.0136)	0.5093 (0.0283)	0.6157 (0.1243)	0.6157 (0.1243)	0.3776 (0.0992)	0.5488 (0.0189)	0.5488 (0.0189)	0.4455 (0.0197)	0.5765 (0.0653)	50%
MLJAR-S3 (200)	0.6908 (0.0011)	0.4998 (0.0113)	0.5123 (0.0099)	0.6428 (0.0669)	0.6428 (0.0669)	0.3507 (0.0836)	0.5514 (0.0094)	0.5514 (0.0094)	0.4399 (0.0205)	0.5920 (0.0278)	50%
H2O-S1 (1000)	0.7327	0.5081	0.5383	0.7982	0.7982	0.2164	0.5579	0.5579	0.4640	0.6568	50%
H2O-S2 (500)	1.4195 (1.2549)	0.4816 (0.0195)	0.5160 (0.0299)	0.7500 (0.2040)	0.7500 (0.2040)	0.2261 (0.1867)	0.5439 (0.0092)	0.5439 (0.0092)	0.4030 (0.0396)	0.6223 (0.0790)	50%
H2O-S3 (200)	0.8881 (0.1641)	0.4964 (0.0135)	0.5130 (0.0291)	0.5452 (0.2572)	0.5452 (0.2572)	0.4731 (0.2731)	0.5658 (0.0320)	0.5658 (0.0320)	0.4536 (0.0227)	0.5239 (0.1577)	50%
TABLET	2.0940	0.5020	0.5567	0.5242	0.8283	0.2201	0.5384	0.5682	0.5086	0.6740	50%
TABLLM	1.0831	0.5019	0.5550	0.5019	1.0000	0.0037	0.7771	0.5543	1.0000	0.3603	50%
MALLM-Zeroshot	0.8124	0.4946	0.5533	0.5000	1.0000	0.0000	0.2767	0.5533	0.0000	0.7124	0.10
MALLM-Fewshot3	0.8021	0.5046	0.5533	0.5000	1.0000	0.0000	0.2767	0.5533	0.0000	0.7124	0.10
MALLM-Fewshot5	0.7591	0.4988	0.5533	0.5000	1.0000	0.0000	0.2767	0.5533	0.0000	0.7124	0.10
MALLM-RAG5	0.8277	0.5053	0.5533	0.5000	1.0000	0.0000	0.2767	0.5533	0.0000	0.7124	0.10
MALLM-RAG10	0.8142	0.4996	0.5533	0.5000	1.0000	0.0000	0.2767	0.5533	0.0000	0.7124	0.10
RENTAL PROPERTIES											
AutoGlueon-S1 (482)	0.0633	0.9981	0.9793	0.9744	0.9744	0.9851	0.9870	0.9870	0.9706	0.9806	
AutoGlueon-S2 (241)	0.1374 (0.0138)	0.9884 (0.0018)	0.9586 (0.0084)	0.9667 (0.0115)	0.9667 (0.0115)	0.9493 (0.0133)	0.9570 (0.0109)	0.9570 (0.0109)	0.9609 (0.0128)	0.9617 (0.0079)	50%
AutoGlueon-S3 (96)	0.2000 (0.0302)	0.9787 (0.0071)	0.9366 (0.0236)	0.9385 (0.0278)	0.9385 (0.0278)	0.9343 (0.0359)	0.9439 (0.0297)	0.9439 (0.0297)	0.9294 (0.0294)	0.9409 (0.0220)	50%
MLJAR-S1 (482)	0.1669	0.9889	0.9586	0.9615	0.9615	0.9552	0.9615	0.9615	0.9552	0.9615	
MLJAR-S2 (241)	0.2721 (0.1392)	0.9845 (0.0018)	0.9476 (0.0166)	0.9513 (0.0344)	0.9513 (0.0344)	0.9433 (0.0125)	0.9515 (0.0094)	0.9515 (0.0094)	0.9446 (0.0362)	0.9510 (0.0165)	50%
MLJAR-S3 (96)	0.3555 (0.1769)	0.9691 (0.0299)	0.9186 (0.0255)	0.9026 (0.0421)	0.9026 (0.0421)	0.9373 (0.0324)	0.9444 (0.0265)	0.9444 (0.0265)	0.8937 (0.0435)	0.9224 (0.0245)	50%
H2O-S1 (482)	0.0740	0.9969	0.9793	0.9744	0.9744	0.9851	0.9870	0.9870	0.9706	0.9806	
H2O-S2 (241)	0.1217 (0.0348)	0.9929 (0.0026)	0.9614 (0.0143)	0.9487 (0.0301)	0.9487 (0.0301)	0.9761 (0.0200)	0.9793 (0.0173)	0.9793 (0.0173)	0.9434 (0.0298)	0.9634 (0.0141)	50%
H2O-S3 (96)	0.2899 (0.1185)	0.9719 (0.0209)	0.9090 (0.0302)	0.8795 (0.0638)	0.8795 (0.0638)	0.9433 (0.0679)	0.9522 (0.0562)	0.9522 (0.0562)	0.8747 (0.0541)	0.9118 (0.0295)	50%
TABLET	1.2092	0.8908	0.8690	0.8749	0.7792	0.9706	0.8815	0.9677	0.7952	0.8687	50%
TabLLM	0.6698	0.8281	0.5310	0.5000	1.0000	0.0000	0.2655	0.5310	0.0000	0.3468	50%
MALLM-Zeroshot	0.3869	0.9461	0.9034	0.9005	0.9481	0.8529	0.9075	0.8795	0.9355	0.9125	0.21
MALLM-Fewshot3	0.3619	0.9410	0.8897	0.8892	0.8961	0.8824	0.8892	0.8961	0.8824	0.8961	0.21
MALLM-Fewshot5	0.3224	0.9482	0.8897	0.8884	0.9091	0.8676	0.8900	0.8861	0.8939	0.8974	0.21
MALLM-RAG5	0.1086	0.9929	0.9724	0.9740	0.9481	1.0000	0.9722	1.0000	0.9444	0.9733	0.31
MALLM-RAG10	0.3528	0.9820	0.9655	0.9667	0.9481	0.9853	0.9651	0.9865	0.9437	0.9669	0.31
ICU ADMISSION											
AutoGlueon-S1 (254)	0.4354	0.9146	0.7532	0.9333	0.9333	0.5000	0.7241	0.7241	0.8421	0.8155	
AutoGlueon-S2 (127)	0.4671 (0.0253)	0.8715 (0.0371)	0.7974 (0.0285)	0.9244 (0.0122)	0.9244 (0.0122)	0.6188 (0.0713)	0.7743 (0.0310)	0.7743 (0.0310)	0.8530 (0.0205)	0.8424 (0.0184)	50%
AutoGlueon-S3 (50)	0.5922 (0.0395)	0.7976 (0.0407)	0.7091 (0.0620)	0.8578 (0.1292)	0.8578 (0.1292)	0.5000 (0.2615)	0.7208 (0.0726)	0.7208 (0.0726)	0.7848 (0.1499)	0.7742 (0.0414)	50%
MLJAR-S1 (254)	0.5168	0.9069	0.8442	0.9333	0.9333	0.7188	0.8235	0.8235	0.8846	0.8750	
MLJAR-S2 (127)	0.5504 (0.0358)	0.8594 (0.0326)	0.7948 (0.0296)	0.8667 (0.0875)	0.8667 (0.0875)	0.6938 (0.1046)	0.8034 (0.0407)	0.8034 (0.0407)	0.8021 (0.0820)	0.8303 (0.0314)	50%
MLJAR-S3 (50)	0.5897 (0.0495)	0.8265 (0.0663)	0.7792 (0.0379)	0.8444 (0.1006)	0.8444 (0.1006)	0.6875 (0.1668)	0.8013 (0.0641)	0.8013 (0.0641)	0.7798 (0.0897)	0.8163 (0.0316)	50%
H2O-S1 (254)	0.4069	0.9028	0.8701	0.9333	0.9333	0.7812	0.8571	0.8571	0.8929	0.8936	
H2O-S2 (127)	0.5369 (0.0738)	0.8351 (0.0360)	0.7325 (0.0793)	0.7378 (0.2116)	0.7378 (0.2116)	0.7250 (0.1787)	0.8049 (0.0641)	0.8049 (0.0641)	0.7007 (0.1164)	0.7491 (0.1266)	50%
H2O-S3 (50)	1.5063 (0.9895)	0.7756 (0.0291)	0.6390 (0.1351)	0.4933 (0.2937)	0.4933 (0.2937)	0.8438 (0.1036)	0.7915 (0.0777)	0.7915 (0.0777)	0.5726 (0.1200)	0.5653 (0.2875)	50%
TABLET	1.4162	0.6931	0.6623	0.6569	0.6889	0.6250	0.6546	0.7209	0.5882	0.6553	50%
TabLLM	0.6848	0.5174	0.5844	0.5000	1.0000	0.0000	0.2922	0.5844	0.0000	0.3689	50%
MALLM-Zeroshot	0.6111	0.7149	0.7273	0.6809	0.9556	0.4062	0.7801	0.6935	0.8667	0.8037	0.26
MALLM-Fewshot3	0.6152	0.7073	0.7143	0.6608	0.9778	0.3438	0.7968	0.6769	0.9167	0.8000	0.21
MALLM-Fewshot5	0.6444	0.7094	0.6753	0.6184	0.9556	0.2812	0.7348	0.6515	0.8182	0.7748	0.21
MALLM-RAG5	0.6092	0.7156	0.6753	0.6139	0.9778	0.2500	0.7680	0.6471	0.8889	0.7788	0.21
MALLM-RAG10	0.5932	0.7309	0.7013	0.6406	1.0000	0.2812	0.8309	0.6618	1.0000	0.7965	0.21

Notes: Standard deviations across resamples are in parentheses. All LLM experiments used temperature 0.

Table 14: Models Performance Metrics Across Pre-Cutoff Datasets

Models	Log Loss	AUC	Accuracy	Recall	Pos. Recall	Neg. Recall	Precision	Pos. Prec.	Neg. Prec.	F1 Score	Best Thresh.
Heart Disease											
AutoGlueon-S1	0.4625	0.8968	0.7444	0.5952	0.5952	0.8750	0.8065	0.8065	0.7119	0.6849	50%
AutoGlueon-S2	0.4501 (0.0466)	0.8978 (0.0237)	0.7844 (0.0365)	0.6524 (0.0686)	0.6524 (0.0686)	0.9000 (0.0228)	0.8503 (0.0341)	0.8503 (0.0341)	0.7487 (0.0376)	0.7372 (0.0536)	50%
AutoGlueon-S3	0.4716 (0.0886)	0.8720 (0.0401)	0.7911 (0.0372)	0.7190 (0.1219)	0.7190 (0.1219)	0.8542 (0.0442)	0.8150 (0.0255)	0.8150 (0.0255)	0.7832 (0.0632)	0.7580 (0.0675)	50%
MLJAR-S1	0.5390	0.9152	0.7889	0.6429	0.6429	0.9167	0.8710	0.8710	0.7458	0.7397	50%
MLJAR-S2	0.5845 (0.0360)	0.9056 (0.0229)	0.8133 (0.0440)	0.7429 (0.1500)	0.7429 (0.1500)	0.8750 (0.0706)	0.8489 (0.0598)	0.8489 (0.0598)	0.8063 (0.0794)	0.7813 (0.0800)	50%
MLJAR-S3	0.6042 (0.0438)	0.8966 (0.0242)	0.7889 (0.0294)	0.6952 (0.1207)	0.6952 (0.1207)	0.8708 (0.0632)	0.8345 (0.0574)	0.8345 (0.0574)	0.7717 (0.0550)	0.7498 (0.0619)	50%
H2O-S1	0.4496	0.9023	0.7889	0.6429	0.6429	0.9167	0.8710	0.8710	0.7458	0.7397	50%
H2O-S2	0.5580 (0.3113)	0.9025 (0.0244)	0.7933 (0.0507)	0.6952 (0.1147)	0.6952 (0.1147)	0.8792 (0.0519)	0.8367 (0.0487)	0.8367 (0.0487)	0.7720 (0.0623)	0.7547 (0.0744)	50%
H2O-S3	0.7156 (0.3900)	0.8661 (0.0220)	0.7267 (0.0848)	0.4905 (0.2343)	0.4905 (0.2343)	0.9333 (0.0728)	0.8947 (0.0880)	0.8947 (0.0880)	0.6897 (0.0875)	0.5960 (0.2128)	50%
TABLET	0.6955	0.7006	0.6111	0.8571	0.8571	0.3958	0.5538	0.5538	0.7600	0.6729	–
TABLLM	0.8134	0.6592	0.6778	0.6592	0.3810	0.9375	0.7380	0.8421	0.6338	0.6404	–
MALLM-Zeroshot	0.4873	0.8857	0.7889	0.7961	0.9048	0.6875	0.8044	0.7170	0.8919	0.8000	0.36
MALLM-Fewshot	0.4936	0.8725	0.7778	0.7857	0.9048	0.6667	0.7963	0.7037	0.8889	0.7917	0.36
MALLM-RAG	0.4355	0.9444	0.8667	0.8720	0.9524	0.7917	0.8750	0.8000	0.9500	0.8696	0.43
TITANIC											
AutoGlueon-S1	0.4592	0.8363	0.7818	0.8649	0.8649	0.6111	0.8205	0.8205	0.6875	0.8421	50%
AutoGlueon-S2	0.5049 (0.0178)	0.8024 (0.0211)	0.7527 (0.0163)	0.8432 (0.0586)	0.8432 (0.0586)	0.5667 (0.0913)	0.8017 (0.0233)	0.8017 (0.0233)	0.6471 (0.0552)	0.8204 (0.0190)	50%
AutoGlueon-S3	0.6147 (0.0604)	0.7405 (0.0233)	0.7309 (0.0152)	0.8649 (0.0689)	0.8649 (0.0689)	0.4556 (0.1639)	0.7702 (0.0463)	0.7702 (0.0463)	0.6341 (0.0554)	0.8117 (0.0122)	50%
MLJAR-S1	0.5106	0.8544	0.7818	0.8649	0.8649	0.6111	0.8205	0.8205	0.6875	0.8421	50%
MLJAR-S2	0.5504 (0.0403)	0.8036 (0.0159)	0.7273 (0.0287)	0.8432 (0.1316)	0.8432 (0.1316)	0.4889 (0.2528)	0.7868 (0.0807)	0.7868 (0.0807)	0.6409 (0.1196)	0.8031 (0.0381)	50%
MLJAR-S3	0.6216 (0.0662)	0.7443 (0.0697)	0.7091 (0.0287)	0.8595 (0.0924)	0.8595 (0.0924)	0.4000 (0.1859)	0.7522 (0.0517)	0.7522 (0.0517)	0.6108 (0.1295)	0.7978 (0.0258)	50%
H2O-S1	0.4771	0.8498	0.8182	0.8649	0.8649	0.7222	0.8649	0.8649	0.7222	0.8649	50%
H2O-S2	0.6371 (0.1873)	0.8105 (0.0201)	0.7382 (0.0304)	0.7784 (0.0520)	0.7784 (0.0520)	0.6556 (0.0994)	0.8250 (0.0374)	0.8250 (0.0374)	0.5925 (0.0420)	0.7996 (0.0265)	50%
H2O-S3	1.2524 (0.9417)	0.7620 (0.0848)	0.6364 (0.1483)	0.6000 (0.3183)	0.6000 (0.3183)	0.7111 (0.2976)	0.8640 (0.1181)	0.8640 (0.1181)	0.4992 (0.0985)	0.6385 (0.2787)	50%
TABLET	1.1685	0.7856	0.6927	0.9865	0.9865	0.0889	0.6901	0.6901	0.8267	0.8120	-
TABLLM	0.8245	0.5000	0.6727	0.5000	1.0000	0.0000	0.3364	0.6727	0.0000	0.4022	–
MALLM-Zeroshot	0.6100	0.7395	0.7273	0.5833	1.0000	0.1667	0.8558	0.7115	1.0000	0.8315	0.26
MALLM-Fewshot	0.4614	0.8401	0.8364	0.8356	0.8378	0.8333	0.8130	0.9118	0.7143	0.8732	0.38
MALLM-RAG	0.4777	0.8311	0.8000	0.8371	0.7297	0.9444	0.7970	0.9643	0.6296	0.8308	0.36
BREAST CANCER											
AutoGlueon-S1	0.0434	0.9985	0.9942	0.9844	0.9844	1.0000	1.0000	1.0000	0.9907	0.9921	50%
AutoGlueon-S2	0.0888 (0.0445)	0.9978 (0.0010)	0.9789 (0.0089)	0.9438 (0.0237)	0.9438 (0.0237)	1.0000 (0.0000)	1.0000 (0.0000)	1.0000 (0.0000)	0.9676 (0.0133)	0.9709 (0.0125)	50%
AutoGlueon-S3	0.1835 (0.0811)	0.9888 (0.0088)	0.9485 (0.0162)	0.9000 (0.0713)	0.9000 (0.0713)	0.9776 (0.0314)	0.9643 (0.0470)	0.9643 (0.0470)	0.9444 (0.0386)	0.9283 (0.0251)	50%
MLJAR-S1	0.1007	0.9982	0.9766	0.9375	0.9375	1.0000	1.0000	1.0000	0.9640	0.9677	50%
MLJAR-S2	0.1617 (0.0258)	0.9974 (0.0021)	0.9146 (0.0152)	0.9699 (0.0070)	0.9699 (0.0070)	0.8654 (0.0244)	0.8166 (0.0272)	0.8166 (0.0272)	0.9979 (0.0048)	0.8976 (0.0166)	50%
MLJAR-S3	0.2263 (0.0870)	0.9938 (0.0042)	0.9193 (0.0162)	0.8875 (0.1015)	0.8875 (0.1015)	0.9383 (0.0767)	0.9144 (0.1014)	0.9144 (0.1014)	0.9388 (0.0536)	0.8917 (0.0184)	50%
H2O-S1	0.0728	0.9559	0.9766	0.9375	0.9375	1.0000	1.0000	1.0000	0.9640	0.9677	50%
H2O-S2	0.1012 (0.0545)	0.9972 (0.0211)	0.9731 (0.0106)	0.9313 (0.0237)	0.9313 (0.0237)	0.9981 (0.0042)	0.9966 (0.0076)	0.9966 (0.0076)	0.9605 (0.0132)	0.9627 (0.0150)	50%
H2O-S3	0.1389 (0.0597)	0.9945 (0.0023)	0.9228 (0.0420)	0.8000 (0.1156)	0.8000 (0.1156)	0.9963 (0.0084)	0.9930 (0.0157)	0.9930 (0.0157)	0.8956 (0.0540)	0.8820 (0.0722)	50%
TABLET	0.3254	0.9661	0.8713	0.6562	0.6562	1.0000	1.0000	1.0000	0.8295	0.7925	–
TABLLM	0.6065	0.6953	0.7719	0.6953	0.3906	1.0000	0.8664	1.0000	0.7329	0.7038	–
MALLM-Zeroshot	0.3039	0.9890	0.9415	0.9533	1.0000	0.9065	0.9324	0.8649	1.0000	0.9275	0.36
MALLM-Fewshot	0.3847	0.9915	0.9532	0.9501	0.9375	0.9626	0.9501	0.9375	0.9626	0.9375	0.66
MALLM-RAG	0.2420	0.9983	0.9825	0.9797	0.9688	0.9907	0.9828	0.9841	0.9815	0.9764	0.66
DIABETES											
AutoGlueon-S1	0.4746	0.8369	0.7359	0.5185	0.5185	0.8533	0.6563	0.6563	0.7665	0.5793	50%
AutoGlueon-S2	0.4746 (0.0000)	0.8369 (0.0000)	0.7515 (0.0142)	0.5679 (0.0486)	0.5679 (0.0486)	0.8507 (0.0281)	0.6745 (0.0340)	0.6745 (0.0340)	0.7852 (0.0162)	0.6151 (0.0271)	50%
AutoGlueon-S3	0.4746 (0.0000)	0.8369 (0.0000)	0.7325 (0.0234)	0.5704 (0.0989)	0.5704 (0.0989)	0.8200 (0.0419)	0.6328 (0.0334)	0.6328 (0.0334)	0.7813 (0.0307)	0.5955 (0.0636)	50%
MLJAR-S1	0.4871	0.8304	0.7316	0.5309	0.5309	0.8400	0.6418	0.6418	0.7683	0.5811	50%
MLJAR-S2	0.4867 (0.0087)	0.8344 (0.0086)	0.7368 (0.0071)	0.5654 (0.0519)	0.5654 (0.0519)	0.8293 (0.0311)	0.6436 (0.0238)	0.6436 (0.0238)	0.7800 (0.0141)	0.6001 (0.0236)	50%
MLJAR-S3	0.5388 (0.0495)	0.8032 (0.0335)	0.7229 (0.0279)	0.6123 (0.0764)	0.6123 (0.0764)	0.7827 (0.0663)	0.6105 (0.0539)	0.6105 (0.0539)	0.7902 (0.0217)	0.6067 (0.0354)	50%
H2O-S1	0.5412	0.8179	0.7359	0.8519	0.8519	0.6733	0.5847	0.5847	0.8938	0.6935	50%
H2O-S2	0.4992 (0.0221)	0.8263 (0.0224)	0.7359 (0.0303)	0.7432 (0.0656)	0.7432 (0.0656)	0.7320 (0.0555)	0.6027 (0.0464)	0.6027 (0.0464)	0.8421 (0.0294)	0.6636 (0.0349)	50%
H2O-S3	0.8905 (0.2711)	0.7969 (0.0554)	0.7221 (0.0429)	0.6741 (0.1945)	0.6741 (0.1945)	0.7480 (0.1072)	0.6013 (0.0703)	0.6013 (0.0703)	0.8217 (0.0793)	0.6203 (0.0913)	50%
TABLET	0.5780	0.8118	0.7186	0.7531	0.7531	0.7000	0.5755	0.5755	0.8400	0.6524	–
TABLLM	0.5228	0.7800	0.8100	0.7800	0.7222	0.8377	0.7447	0.5843	0.9052	0.7581	–
MALLM-Zeroshot	0.5335	0.8184	0.7489	0.7499	0.7531	0.7467	0.7323	0.6162	0.8485	0.6778	0.46
MALLM-Fewshot	0.5631	0.8181	0.7273	0.7730	0.9259	0.6200	0.7538	0.5682	0.9394	0.7042	0.36
MALLM-RAG	0.4872	0.8627	0.7965	0.7894	0.7654	0.8133	0.7771	0.6889	0.8652	0.7251	0.46
ADULT											
AutoGlueon-S1	0.3217	0.9092	0.8433	0.5694	0.5694	0.9298	0.7193	0.7193	0.8724	0.6357	50%
AutoGlueon-S2	0.3429 (0.0039)	0.9050 (0.0023)	0.8340 (0.0064)	0.5806 (0.0731)	0.5806 (0.0731)	0.9140 (0.0212)	0.6847 (0.0345)	0.6847 (0.0345)	0.8740 (0.0167)	0.6248 (0.0349)	50%
AutoGlueon-S3	0.4167 (0.0286)	0.8726 (0.0169)	0.8060 (0.0144)	0.6139 (0.1933)	0.6139 (0.1933)	0.8667 (0.0765)	0.6187 (0.0792)	0.6187 (0.0792)	0.8815 (0.0441)	0.5915 (0.0716)	50%
MLJAR-S1	0.3349	0.9098	0.8400	0.5278	0.5278	0.9386	0.7308	0.7308	0.8629	0.6129	50%
MLJAR-S2	0.3633 (0.0277)	0.9080 (0.0029)	0.8293 (0.0076)	0.6389 (0.0501)	0.6389 (0.0501)	0.8895 (0.0114)	0.6463 (0.0147)	0.6463 (0.0147)	0.8866 (0.0130)	0.6417 (0.0264)	50%
MLJAR-S3	0.3874 (0.0425)	0.8843 (0.0195)	0.8013 (0.0407)	0.7028 (0.0813)	0.7028 (0.0813)	0.8325 (0.0785)	0.5852 (0.0721)	0.5852 (0.0721)	0.9003 (0.0184)	0.6314 (0.0196)	50%
H2O-S1	0.3288	0.9066	0.8400	0.8611	0.8611	0.8333	0.6200	0.6200	0.9500	0.7209	50%
H2O-S2	0.3621 (0.0387)	0.8874 (0.0210)	0.8167 (0.0113)	0.6917 (0.1730)	0.6917 (0.1730)	0.8561 (0.0657)	0.6241 (0.0859)	0.6241 (0.0859)	0.9021 (0.0429)	0.6366 (0.0550)	50%
H2O-S3	0.6109 (0.3099)	0.8201 (0.0882)	0.7307 (0.1910)	0.6250 (0.2546)	0.6250 (0.2546)	0.7640 (0.3144)	0.5756 (0.1705)	0.5756 (0.1705)	0.8831 (0.0553)	0.5408 (0.1240)	50%
TABLET	0.4437	0.8787	0.8367	0.6528	0.6528	0.8947	0.6620	0.6620	0.8908	0.6573	–
TABLLM	0.5228	0.7800	0.8100	0.7800	0.7222	0.8377	0.7447	0.5843	0.9052	0.7581	–
MALLM-Zeroshot	0.4571	0.9023	0.7867	0.8454	0.9583	0.7325	0.7566	0.5308	0.9824	0.6832	0.46
MALLM-Fewshot	0.3088	0.9845	0.9600	0.9547	0.9444	0.9649	0.9384	0.8947	0.9821	0.9189	0.46
MALLM-RAG	0.3856	0.9391	0.8633	0.8531	0.9167	0.8289	0.7989	0.6286	0.9692	0.7458	0.36

Table 15: Correlation between Algorithm Hardness & Corrected Probability for Pre-Cutoff Datasets

Dataset	AutoGluon	MLJAR	H2O	TABLET	TABLLM	MALLM-Zeroshot	MALLM-Fewshot	MALLM-RAG
Heart Disease	-0.6722	-0.7353	-0.6344	-0.14809	-0.5304	-0.1940	-0.2153	-0.1844
Titanic	-0.7994	-0.8839	-0.7865	-0.3529	-0.6067	-0.5861	-0.4688	-0.4208
Breast Cancer	-0.4644	-0.6947	-0.5669	-0.6773	-0.6419	-0.2299	-0.3664	-0.2256
Diabetes	-0.8160	-0.8274	-0.7134	-0.4275	-0.6402	-0.5304	-0.2368	-0.5040
Adult	-0.7992	-0.8858	-0.8850	-0.45402	-0.4544	-0.1863	-0.1866	-0.2525

Table 16: Models Performance Metrics Across Post-Cutoff Datasets [Temperature 1.1]

Models	Log Loss	AUC	Accuracy	Recall	Pos. Recall	Neg. Recall	Precision	Pos. Prec.	Neg. Prec.	F1 Score	Best Thresh.
STUDENT DEPRESSION											
MALLM-Zeroshot	0.5666	0.8203	0.7700	0.7342	0.9485	0.5200	0.8064	0.7345	0.8783	0.8279	0.65
MALLM-Fewshot3	0.5387	0.8042	0.7633	0.7297	0.9314	0.5280	0.7901	0.7342	0.8461	0.8211	0.35
MALLM-Fewshot5	0.6470	0.7948	0.7700	0.7251	0.9940	0.4560	0.8508	0.7190	0.9827	0.8345	0.20
MALLM-RAG5	0.5748	0.8433	0.7900	0.7560	0.9600	0.5520	0.8289	0.7500	0.9078	0.8421	0.60
MALLM-RAG10	0.8007	0.8156	0.7900	0.7582	0.9485	0.5680	0.8210	0.7545	0.8875	0.8405	0.65
POSTPARTUM DEPRESSION											
MALLM-Zeroshot	0.7014	0.5685	0.5580	0.5747	0.5189	0.6306	0.5683	0.7224	0.4142	0.6040	0.10
MALLM-Fewshot3	0.9040	0.5756	0.6473	0.4982	0.9965	0.0000	0.3243	0.6487	0.0000	0.7859	0.10
MALLM-Fewshot5	0.8880	0.5941	0.6517	0.5031	1.0000	0.0063	0.8255	0.6510	1.0000	0.7886	0.11
MALLM-RAG5	0.7718	0.7775	0.7410	0.6745	0.8969	0.4522	0.7275	0.7521	0.7029	0.8181	0.40
MALLM-RAG10	0.6117	0.8157	0.7566	0.6616	0.7345	0.9000	0.8172	0.7345	0.9000	0.8394	0.40
EMPLOYEE SATISFACTION											
MALLM-Zeroshot	1.3330	0.8599	0.8466	0.6699	0.9180	0.4210	0.7246	0.6071	0.8421	0.7311	0.85
MALLM-Fewshot3	1.3139	0.8473	0.7033	0.7065	0.9459	0.4671	0.7660	0.6334	0.8987	0.7588	0.85
MALLM-Fewshot5	1.0829	0.8655	0.7100	0.7131	0.9459	0.4802	0.7702	0.6392	0.9012	0.7629	0.88
MALLM-RAG5	1.1956	0.8879	0.8233	0.8231	0.8108	0.8355	0.8234	0.8275	0.8193	0.8191	0.40
MALLM-RAG10	0.8052	0.9094	0.8266	0.8272	0.8716	0.7828	0.8293	0.7962	0.8623	0.8052	0.30
STUDENTS AI TOOLS											
MALLM-Zeroshot	0.8105	0.5298	0.5083	0.5114	0.9798	0.0430	0.5933	0.5025	0.6842	0.6643	0.11
MALLM-Fewshot3	0.8120	0.5308	0.4966	0.5000	1.0000	0.0000	0.2483	0.4966	0.0000	0.6636	0.10
MALLM-Fewshot5	0.8113	0.5313	0.5000	0.5033	1.0000	0.0066	0.7491	0.4983	1.0000	0.6651	0.11
MALLM-RAG5	0.8528	0.5298	0.4983	0.5016	1.0000	0.0033	0.7487	0.7487	1.0000	0.6640	0.10
MALLM-RAG10	0.8143	0.5476	0.4954	0.5046	1.0000	0.0033	0.7523	0.4967	1.0000	0.6789	0.11
RENTAL PROPERTIES											
MALLM-Zeroshot	0.4150	0.9260	0.8690	0.8689	0.8701	0.8676	0.8683	0.8816	0.8551	0.8758	0.26
MALLM-Fewshot3	0.3859	0.9214	0.8690	0.8715	0.8312	0.9118	0.8705	0.9143	0.8267	0.8707	0.31
MALLM-Fewshot5	0.3218	0.9453	0.8966	0.8966	0.8961	0.8971	0.8960	0.9079	0.8841	0.9020	0.26
MALLM-RAG5	0.1318	0.9908	0.9724	0.9740	0.9481	1.0000	0.9722	1.0000	0.9444	0.9733	0.61
MALLM-RAG10	0.1211	0.9894	0.9793	0.9805	0.9610	1.0000	0.9789	1.0000	0.9577	0.9801	0.56
ICU ADMISSION											
MALLM-Zeroshot	0.5918	0.7375	0.7273	0.6944	0.8889	0.5000	0.7381	0.7143	0.7619	0.7921	0.36
MALLM-Fewshot3	0.6274	0.7208	0.6753	0.6139	0.9778	0.2500	0.7680	0.6471	0.8889	0.7788	0.26
MALLM-Fewshot5	0.6775	0.6920	0.7013	0.6497	0.9556	0.3438	0.7590	0.6719	0.8462	0.7890	0.21
MALLM-RAG5	0.6729	0.6663	0.6623	0.5938	1.0000	0.1875	0.8169	0.6338	1.0000	0.7759	0.21
MALLM-RAG10	0.6007	0.7392	0.7143	0.6608	0.9778	0.3438	0.7968	0.6769	0.9167	0.8000	0.23

Appendix F: AutoML Baselines and Hyperparameters

F.1 AutoGluon

We used AutoGluon-Tabular (`TabularPredictor`) with the target label set to "`target`". Model training was performed using `eval_metric = roc_auc` and the `best_quality` preset. Each run was constrained by a time budget of `time_limit = total_time_limit` (default: 60 seconds). In the single-run configuration, the predictor was trained with `verbosity = 2`. For robustness analysis, we additionally conducted multi-seed experiments using seeds $\{1, 123, 42, 0, 110\}$, where for each seed the training data were randomly subsampled to 50% of the original training set (`fraction = 0.5`). In multi-seed runs, AutoGluon was trained with identical hyperparameters but with `verbosity = 0`.

F.2 MLJAR AutoML

We used the `mljar-supervised` AutoML framework in `Compete` mode with `eval_metric = auc` and a total time budget of `total_time_limit` (default: 60 seconds). The search space included LightGBM, XGBoost, CatBoost, Random Forest, and Linear models. Model selection was performed using 5-fold cross-validation with shuffled and stratified folds (`validation_type = kfold`, `k_folds = 5`, `shuffle = True`, `stratify = True`). For robustness, training was repeated across seeds $\{1, 123, 42, 0, 110\}$, using a 50% random subsample of the training data for each seed (`fraction = 0.5`), while keeping all other hyperparameters fixed.

F.3 H2O AutoML

For H2O AutoML, the H2O engine was initialized with `max_mem_size = "4G"` and `nthreads = -1`. All non-target columns were used as input features, and the target variable was converted

to a categorical factor when originally numeric. Model training was performed using `H2OAutoML` with `max_runtime_secs` (default: 60 seconds), `balance_classes = True`, and `sort_metric = AUC`. In the single-run setting, a fixed random seed of 42 was used. For robustness analysis, we conducted multi-seed experiments with seeds $\{1, 123, 42, 0, 110\}$, where each run used a 50% random subsample of the training data (`fraction = 0.5`) and the corresponding seed was passed directly to the AutoML procedure. All code is available at https://github.com/mallm-framework/MALLM.

Appendix G: Cost and Latency Analysis

Table 17 reports cost and latency for all post-cutoff datasets and strategies. Average per-prediction cost ranges from $0.010 (Zero-shot) to $0.028 (Employee Satisfaction Zero-shot), averaging $0.016. RAG-10 incurs a 6% premium over Zero-shot due to longer in-context examples, but latency is largely insensitive to context length (17–22 seconds per prediction on most datasets, 9–11 seconds on Students AI Tools due to shorter feature serializations). A full experimental campaign (150 runs) costs approximately $750 in API fees.

While these costs exceed AutoML's training-time-only computation, MALLM requires no model training and deploys immediately on new datasets—a trade-off favorable for exploratory or low-data settings but less so for production deployment (see Section 5).

Table 17: Cost and latency analysis across post-cutoff datasets using GPT-4o-2024-08-06. Values averaged over 5 runs.

Dataset	Strategy	$/Pred	s/Pred	Total $	Tokens(K)
Stud. Dep. (300)	Zeroshot	0.014	20.4	4.26	833
	Fewshot3	0.016	21.3	4.70	969
	Fewshot5	0.016	21.6	4.91	1044
	RAG5	0.016	21.4	4.84	1046
	RAG10	0.018	20.2	5.32	1236
Postp. Dep. (448)	Zeroshot	0.010	20.3	4.43	802
	Fewshot3	0.010	20.7	4.61	889
	Fewshot5	0.011	20.6	4.81	982
	RAG5	0.015	20.5	6.66	1330
	RAG10	0.016	20.9	7.00	1509
Empl. Sat. (300)	Zeroshot	0.028	31.2	8.26	1693
	Fewshot3	0.018	21.1	5.32	1487
	Fewshot5	0.019	21.1	5.59	1606
	RAG5	0.018	20.4	5.52	1489
	RAG10	0.019	20.2	5.70	1658
Stud. AI (600)	Zeroshot	0.013	9.2	7.66	1543
	Fewshot3	0.016	10.5	9.74	2059
	Fewshot5	0.015	9.8	9.00	2052
	RAG5	0.015	10.8	9.16	1970
	RAG10	0.016	10.5	9.75	2278
Rental Prop. (145)	Zeroshot	0.013	18.4	1.94	380
	Fewshot3	0.015	18.1	2.20	486
	Fewshot5	0.016	18.0	2.35	549
	RAG5	0.015	17.7	2.13	478
	RAG10	0.016	16.7	2.33	562
ICU Adm. (77)	Zeroshot	0.016	20.3	1.22	234
	Fewshot3	0.017	21.3	1.35	274
	Fewshot5	0.018	21.5	1.41	298
	RAG5	0.018	19.1	1.42	303
	RAG10	0.020	18.7	1.55	360
Average across datasets					
	Zeroshot	0.016	20.0	4.63	914
	Fewshot3	0.015	18.8	4.65	1027
	Fewshot5	0.016	18.7	4.68	1088
	RAG5	0.016	18.3	4.96	1103
	RAG10	0.017	17.9	5.27	1267

Appendix H: Robustness and Consistency Analysis

Table 18 reports mean performance and standard deviation across 5 independent inference runs at temperature=0, addressing concerns about LLM consistency. MALLM predictions are generally stable: AUC standard deviations fall below 0.02 for most configurations, and the ranking of strategies is preserved across runs. RAG configurations are the most stable, indicating that retrieval-augmented context acts as a calibration anchor across resamples.

The largest run-to-run variance occurs on datasets where MALLM operates near its discriminative limits—Postpartum RAG-5 ($\pm$0.073) and ICU Admission Fewshot-5 ($\pm$0.046). Students AI Tools remains near-random across all configurations (AUC $\leq$ 0.53) with essentially zero variance, reflecting model collapse to the majority class rather than genuine stability. Overall, these results confirm that while API-level stochasticity produces non-trivial variance even at temperature 0, the variation is modest and does not affect the paper's qualitative conclusions.

Appendix I: Hybrid LLM-XGBoost Analysis

To test whether LLMs add more value as upstream analysts than as end-to-end classifiers, we fed MALLM outputs into XGBoost. From the zero-shot runs we extract (i) features selected by Agent 1 in >50% of instances and (ii) interactions proposed by Agent 2 in >30% of instances, parsed as Python expressions and materialized as new columns. The target is masked from both agents' inputs to prevent leakage. XGBoost is trained with 10 seeds

Table 18: Robustness check: Average performance across 5 independent runs at temperature=0 (standard deviation in parentheses). All experiments use GPT-4o-2024-08-06 with `text-embedding-ada-002` for RAG retrieval.

Models	Log Loss	AUC	Accuracy	Recall	Pos. Recall	Neg. Recall	Precision	Pos. Prec.	Neg. Prec.	F1 Score	Best Thresh.
STUDENT DEPRESSION											
MALLM-Zeroshot	0.5658 (0.0098)	0.8211 (0.0122)	0.7747 (0.0058)	0.7499 (0.0051)	0.8983 (0.0264)	0.6016 (0.0292)	0.7854 (0.0115)	0.7598 (0.0089)	0.8111 (0.0297)	0.8229 (0.0076)	0.70
MALLM-Fewshot3	0.5176 (0.0095)	0.8131 (0.0109)	0.7780 (0.0072)	0.7512 (0.0061)	0.9120 (0.0143)	0.5904 (0.0078)	0.7926 (0.0115)	0.7571 (0.0032)	0.8280 (0.0215)	0.8273 (0.0068)	0.31
MALLM-Fewshot5	0.5004 (0.0109)	0.8238 (0.0085)	0.7840 (0.0108)	0.7621 (0.0131)	0.8937 (0.0272)	0.6304 (0.0418)	0.7920 (0.0139)	0.7727 (0.0170)	0.8114 (0.0308)	0.8283 (0.0089)	0.37
MALLM-RAG5	0.5405 (0.0197)	0.8107 (0.0125)	0.7700 (0.0121)	0.7462 (0.0100)	0.8891 (0.0383)	0.6032 (0.0367)	0.7800 (0.0255)	0.7587 (0.0110)	0.8013 (0.0554)	0.8182 (0.0130)	0.45
MALLM-RAG10	0.5603 (0.1204)	0.8274 (0.0093)	0.7933 (0.0084)	0.7675 (0.0106)	0.9223 (0.0172)	0.6128 (0.0310)	0.8101 (0.0102)	0.7696 (0.0118)	0.8507 (0.0235)	0.8389 (0.0062)	0.43
POSTPARTUM DEPRESSION											
MALLM-Zeroshot	0.7210 (0.0341)	0.5123 (0.0109)	0.6505 (0.0011)	0.5013 (0.0016)	1.0000 (0.0000)	0.0025 (0.0031)	0.5251 (0.2453)	0.6501 (0.0007)	0.4000 (0.4899)	0.7880 (0.0005)	0.12
MALLM-Fewshot3	0.6859 (0.0224)	0.5610 (0.0395)	0.6580 (0.0117)	0.5186 (0.0260)	0.9849 (0.0220)	0.0522 (0.0741)	0.5273 (0.1654)	0.6588 (0.0133)	0.3958 (0.3233)	0.7892 (0.0020)	0.19
MALLM-Fewshot5	0.6864 (0.0199)	0.5730 (0.0369)	0.6500 (0.0009)	0.4832 (0.0336)	0.8013 (0.3975)	0.1651 (0.3302)	0.4249 (0.2003)	0.6498 (0.0006)	0.2000 (0.4000)	0.7876 (0.0002)	0.10
MALLM-RAG5	0.4337 (0.1244)	0.9199 (0.0730)	0.8710 (0.0650)	0.8529 (0.0892)	0.9134 (0.0101)	0.7924 (0.1705)	0.8594 (0.0660)	0.8967 (0.0725)	0.8220 (0.0600)	0.9037 (0.0428)	0.70
MALLM-RAG10	0.5768 (0.0223)	0.8859 (0.0383)	0.8312 (0.0398)	0.8059 (0.0769)	0.8907 (0.0479)	0.7210 (0.2014)	0.8331 (0.0055)	0.8658 (0.0688)	0.8004 (0.0630)	0.8741 (0.0184)	0.59
EMPLOYEE SATISFACTION											
MALLM-Zeroshot	1.0487 (0.0079)	0.8317 (0.0045)	0.7313 (0.0234)	0.7335 (0.0228)	0.8973 (0.0207)	0.5697 (0.0644)	0.7619 (0.0128)	0.6720 (0.0314)	0.8518 (0.0126)	0.7676 (0.0127)	0.86
MALLM-Fewshot3	0.6760 (0.2096)	0.8536 (0.0097)	0.7667 (0.0336)	0.7672 (0.0322)	0.8095 (0.0684)	0.7250 (0.1320)	0.7794 (0.0169)	0.7540 (0.0687)	0.8047 (0.0398)	0.7748 (0.0105)	0.79
MALLM-Fewshot5	0.7253 (0.3491)	0.8583 (0.0068)	0.7733 (0.0347)	0.7737 (0.0334)	0.8000 (0.0656)	0.7474 (0.1288)	0.7839 (0.0171)	0.7671 (0.0658)	0.8007 (0.0377)	0.7779 (0.0143)	0.80
MALLM-RAG5	0.6553 (0.2259)	0.8856 (0.0043)	0.7840 (0.0228)	0.7854 (0.0218)	0.8919 (0.0618)	0.6789 (0.1032)	0.8067 (0.0074)	0.7372 (0.0522)	0.8763 (0.0511)	0.8033 (0.0076)	0.33
MALLM-RAG10	0.5807 (0.0699)	0.8980 (0.0087)	0.8013 (0.0177)	0.8021 (0.0174)	0.8608 (0.0101)	0.7434 (0.0379)	0.8062 (0.0151)	0.7666 (0.0257)	0.8458 (0.0090)	0.8106 (0.0136)	0.38
STUDENTS AI TOOLS											
MALLM-Zeroshot	0.8027 (0.0058)	0.5051 (0.0062)	0.5533 (0.0000)	0.5000 (0.0000)	1.0000 (0.0000)	0.0000 (0.0000)	0.2767 (0.0000)	0.5533 (0.0000)	0.0000 (0.0000)	0.7124 (0.0000)	0.10
MALLM-Fewshot3	0.7902 (0.0110)	0.5231 (0.0142)	0.5533 (0.0000)	0.5000 (0.0000)	1.0000 (0.0000)	0.0000 (0.0000)	0.2767 (0.0000)	0.5533 (0.0000)	0.0000 (0.0000)	0.7124 (0.0000)	0.10
MALLM-Fewshot5	0.7961 (0.0207)	0.5102 (0.0094)	0.5537 (0.0007)	0.5004 (0.0007)	1.0000 (0.0000)	0.0007 (0.0015)	0.3768 (0.2002)	0.5535 (0.0004)	0.2000 (0.4000)	0.7126 (0.0003)	0.10
MALLM-RAG5	0.8039 (0.0126)	0.5272 (0.0110)	0.5537 (0.0007)	0.5004 (0.0007)	1.0000 (0.0000)	0.0007 (0.0015)	0.3768 (0.2002)	0.5535 (0.0004)	0.2000 (0.4000)	0.7126 (0.0003)	0.10
MALLM-RAG10	0.7960 (0.0096)	0.5226 (0.0118)	0.5533 (0.0000)	0.5000 (0.0000)	1.0000 (0.0000)	0.0000 (0.0000)	0.2767 (0.0000)	0.5533 (0.0000)	0.0000 (0.0000)	0.7124 (0.0000)	0.10
RENTAL PROPERTIES											
MALLM-Zeroshot	0.3955 (0.0266)	0.9345 (0.0220)	0.8841 (0.0322)	0.8815 (0.0347)	0.9247 (0.0468)	0.8382 (0.0902)	0.8922 (0.0245)	0.8724 (0.0597)	0.9120 (0.0398)	0.8950 (0.0248)	0.23
MALLM-Fewshot3	0.3493 (0.0137)	0.9398 (0.0060)	0.8841 (0.0119)	0.8834 (0.0131)	0.8961 (0.0217)	0.8706 (0.0399)	0.8851 (0.0102)	0.8883 (0.0274)	0.8820 (0.0195)	0.8916 (0.0096)	0.21
MALLM-Fewshot5	0.3341 (0.0295)	0.9426 (0.0058)	0.8897 (0.0157)	0.8891 (0.0177)	0.8987 (0.0172)	0.8794 (0.0504)	0.8909 (0.0149)	0.8964 (0.0387)	0.8853 (0.0121)	0.8967 (0.0119)	0.23
MALLM-RAG5	0.1600 (0.0942)	0.9901 (0.0037)	0.9738 (0.0080)	0.9752 (0.0079)	0.9533 (0.0104)	0.9971 (0.0059)	0.9734 (0.0079)	0.9973 (0.0055)	0.9496 (0.0109)	0.9747 (0.0078)	0.53
MALLM-RAG10	0.1571 (0.0981)	0.9906 (0.0046)	0.9710 (0.0052)	0.9720 (0.0049)	0.9559 (0.0104)	0.9882 (0.0059)	0.9706 (0.0052)	0.9893 (0.0054)	0.9520 (0.0108)	0.9722 (0.0050)	0.36
ICU ADMISSION											
MALLM-Zeroshot	0.6086 (0.0126)	0.7138 (0.0123)	0.7091 (0.0133)	0.6545 (0.0196)	0.9778 (0.0199)	0.3312 (0.0579)	0.8007 (0.0243)	0.6734 (0.0147)	0.9281 (0.0621)	0.7972 (0.0047)	0.20
MALLM-Fewshot3	0.6183 (0.0593)	0.7164 (0.0422)	0.7143 (0.0184)	0.6716 (0.0273)	0.9244 (0.0458)	0.4188 (0.0918)	0.7530 (0.0298)	0.6931 (0.0274)	0.8129 (0.0685)	0.7908 (0.0102)	0.23
MALLM-Fewshot5	0.6024 (0.0588)	0.7409 (0.0455)	0.7221 (0.0354)	0.6828 (0.0527)	0.9156 (0.0603)	0.4500 (0.1601)	0.7585 (0.0161)	0.7077 (0.0540)	0.8093 (0.0493)	0.7944 (0.0139)	0.29
MALLM-RAG5	0.5977 (0.0080)	0.7244 (0.0067)	0.7117 (0.0301)	0.6640 (0.0451)	0.9467 (0.0479)	0.3812 (0.1361)	0.7778 (0.0281)	0.6867 (0.0391)	0.8688 (0.0840)	0.7938 (0.0104)	0.25
MALLM-RAG10	0.6029 (0.0127)	0.7204 (0.0146)	0.7039 (0.0223)	0.6492 (0.0362)	0.9733 (0.0533)	0.3250 (0.1228)	0.8114 (0.0346)	0.6729 (0.0329)	0.9500 (0.1000)	0.7936 (0.0071)	0.23

(300 estimators, max depth 6, learning rate 0.1, class-balanced) in four configurations: all raw features, LLM-selected features only, and each plus engineered interactions. Configurations involving interactions are reported only when Agent 2 produces valid expressions.

Across five of six datasets, XGBoost with all raw features matches or exceeds the best MALLM (AUC gaps: +6.9pp[5] Student Depression, +7.0pp Postpartum, +4.0pp Employee Satisfaction, +8.3pp ICU). . On Student Depression, adding LLM-suggested interactions (e.g., Financial Stress $\times$ Family Mental Health History) yields the best overall AUC of 0.899. XGBoost on LLM-selected features alone retains 87–100% of full-feature AUC on four datasets, with Rental Properties a striking success: five LLM-chosen features (sqft, baths, beds, latitude, longitude) achieve AUC 0.998, marginally exceeding all 17 raw features. The one clear failure is ICU Admission, where clinically-motivated selection (severity scores, age, BMI) drops AUC from 0.824 to 0.561, indicating that domain-interpretable features are not always the most statistically discriminative.

These results support a hybrid neurosymbolic design: LLMs add value as upstream feature analysts, while gradient-boosted trees remain the reliable inference engine. For Postpartum Depression and Employee Satisfaction, interaction configurations are omitted—Postpartum yielded no interaction meeting the 30% threshold, and Employee Satisfaction's proposed column names (e.g., *engagement_score*) did not exist in the actual schema, illustrating schema hallucination as a failure mode any LLM-driven feature engineering pipeline must guard against.

Appendix J: Agent Pipeline Ablation

Table 20 isolates the contribution of each stage in the multi-agent pipeline by comparing the full 4-agent MALLM against a single Prediction Agent that receives only raw tabular data. The zero-shot vs. pred-only comparison reveals that Agents 1–2 provide inconsistent benefits: they improve AUC on Student Depression (+2.4pp) and Students AI Tools (+17.9pp), but underperform the single-agent baseline on Postpartum Depression (−4.6pp), ICU Admission (−3.7pp), and Rental Properties (−1.1pp), and are essentially neutral on Employee Satisfaction (+0.3pp). Feature analysis and engineering agents therefore do not reliably improve classification when the LLM itself is the final classifier, though Appendix I shows their outputs remain useful when routed to a traditional classifier.

The retrieval agent, in contrast, produces

[5]Throughout the appendices, *pp* denotes percentage points.

Table 19: Hybrid experiment: LLM as feature analyst vs. direct classifier. XGBoost trained with 10 random seeds. LLM features extracted from MALLM zero-shot Agent 1 (feature selection) and Agent 2 (interaction engineering) outputs. Best MALLM selected by highest 5-run average AUC from Table 18.

Dataset	Method	AUC	F1	#Feat
Student Dep.	XGBoost (all)	0.896	0.873	13
	XGBoost (LLM feat.)	0.867	0.836	7
	XGBoost (LLM feat.+int.)	0.867	0.835	10
	XGBoost (all+LLM int.)	**0.899**	0.867	16
	Best MALLM (RAG10)	0.827	0.839	–
Postpartum Dep.	XGBoost (all)	**0.990**	0.981	10
	XGBoost (LLM feat.)	0.985	0.964	7
	Best MALLM (RAG5)	0.920	0.904	–
Employee Sat.	XGBoost (all)	**0.938**	0.847	62
	XGBoost (LLM feat.)	0.920	0.839	8
	Best MALLM (RAG10)	0.898	0.811	–
Students AI	XGBoost (all)	0.421	0.682	14
	XGBoost (LLM feat.)	0.474	0.651	3
	XGBoost (LLM feat.+int.)	0.411	0.635	5
	XGBoost (all+LLM int.)	0.460	0.658	16
	Best MALLM (RAG5)	**0.533**	0.688	–
Rental Prop.	XGBoost (all)	0.997	0.981	17
	XGBoost (LLM feat.)	**0.998**	0.987	5
	XGBoost (LLM feat.+int.)	0.993	0.974	8
	XGBoost (all+LLM int.)	0.994	0.974	20
	Best MALLM (RAG10)	0.991	0.972	–
ICU Admission	XGBoost (all)	**0.824**	0.848	17
	XGBoost (LLM feat.)	0.561	0.696	6
	XGBoost (LLM feat.+int.)	0.599	0.713	10
	XGBoost (all+LLM int.)	0.812	0.831	21
	Best MALLM (Fewshot5)	0.741	0.794	–

Table 20: Ablation: Prediction Agent only vs. full 4-agent pipeline. GPT-4o, temperature 0, 5-run averages. MALLM rows from the robustness table (Table 18).

Dataset	Method	AUC	F1
Student Dep.	Pred-Only	0.797	0.818
	Full (Zeroshot)	0.821	0.823
	Full (RAG10)	0.827	0.839
	Delta (ZS−PO)	*+0.024*	*+0.005*
Postpartum Dep.	Pred-Only	0.558	0.790
	Full (Zeroshot)	0.512	0.788
	Full (RAG5)	0.920	0.904
	Delta (ZS−PO)	*−0.046*	*−0.002*
Employee Sat.	Pred-Only	0.829	0.750
	Full (Zeroshot)	0.832	0.768
	Full (RAG10)	0.898	0.811
	Delta (ZS−PO)	*+0.003*	*+0.017*
Students AI	Pred-Only	0.331	0.696
	Full (Zeroshot)	0.510	0.703
	Full (RAG5)	0.533	0.688
	Delta (ZS−PO)	*+0.179*	*+0.007*
Rental Prop.	Pred-Only	0.945	0.899
	Full (Zeroshot)	0.934	0.895
	Full (RAG10)	0.991	0.972
	Delta (ZS−PO)	*−0.011*	*−0.004*
ICU Admission	Pred-Only	0.751	0.802
	Full (Zeroshot)	0.714	0.797
	Full (Fewshot5)	0.741	0.794
	Delta (ZS−PO)	*−0.037*	*−0.004*

the pipeline's largest and most consistent gains. Adding Agent 3 lifts Postpartum Depression from 0.558 (Pred-Only) to 0.920 (RAG-5), Employee Satisfaction from 0.829 to 0.898 (RAG-10), and Rental Properties from 0.945 to 0.991 (RAG-10). The practical implication is that in-context example retrieval, not intermediate analytical reasoning, is the primary source of value in the multi-agent architecture—directly addressing the concern that error propagation in linear pipelines erodes the benefit of upstream agents.

Appendix K: Embedding Model Comparison

To test whether retrieval quality bottlenecks RAG performance, we replaced `text-embedding-ada-002` with the more capable `text-embedding-3-large` across all six post-cutoff datasets (Table 21). Using RAG $k = 5$ with identical pipeline settings, AUC changes ranged from -1.9pp (Students AI Tools) to $+1.9$pp (Postpartum Depression), with F1 shifts of ≤ 1.7pp on five of six datasets. Postpartum Depression is the only dataset with a meaningful gain (AUC $+1.9$pp, Log Loss -0.254), likely because its clinical embedding text benefits from a higher-capacity encoder. These largely inconsistent and negligible differences across diverse domains confirm that retrieval quality is not the bottleneck; the performance ceiling is set by the LLM's ability to leverage in-context examples for tabular reasoning.

Appendix L: Prompt Optimization with DSPy

To test whether prompt engineering is a limiting factor, we reimplemented the MALLM pipeline with DSPy (Khattab et al., 2024) and ran MIPROv2 Bayesian optimization over 6 instruction candidates per agent on the Student Depression dataset. The unoptimized DSPy baseline already matched the hand-crafted prompts (76.3% validation accuracy vs. 77.9% on the full test set), and the best partially-optimized trial failed to improve over the default (74.3%).

Table 22 shows that on the test set, DSPy Few-

Table 21: Embedding model comparison: `ada-002` vs. `3-large` for RAG retrieval (k=5). Same LLM (GPT-4o), same pipeline.

Dataset	Embedding	AUC	F1	Log Loss
Student Dep.	ada-002	0.810	0.816	0.525
	3-large	0.816	0.827	0.558
	Delta	*+0.006*	*+0.010*	*+0.033*
Postpartum Dep.	ada-002	0.951	0.925	0.520
	3-large	0.971	0.937	0.266
	Delta	*+0.019*	*+0.013*	*−0.254*
Employee Sat.	ada-002	0.887	0.789	0.548
	3-large	0.881	0.791	0.457
	Delta	*−0.006*	*+0.003*	*−0.091*
Students AI	ada-002	0.522	0.696	0.778
	3-large	0.504	0.712	0.822
	Delta	*−0.019*	*+0.017*	*+0.044*
Rental Prop.	ada-002	0.993	0.973	0.109
	3-large	0.995	0.973	0.102
	Delta	*+0.003*	*+0.000*	*−0.006*
ICU Admission	ada-002	0.716	0.779	0.609
	3-large	0.713	0.784	0.640
	Delta	*−0.003*	*+0.006*	*+0.031*

Table 22: DSPy-optimized vs. hand-crafted prompts on Student Depression (RAG k=5, GPT-4o, temperature 0).

Method	AUC	F1
DSPy Fewshot3	0.830 (0.009)	0.824 (0.004)
MALLM-Fewshot3	0.815 (0.010)	0.825 (0.006)
MALLM-Zeroshot	0.828 (0.008)	0.828 (0.004)
MALLM-RAG10	0.826 (0.007)	0.840 (0.006)
AutoGluon-S1	0.896	0.873

shot (k=3) yields AUC 0.830—a marginal 1.5pp gain over the hand-crafted equivalent but well within the envelope of other hand-crafted strategies (Zero-shot 0.828, RAG-10 0.826), and still 6.6pp below the AutoGluon baseline (0.896). Prompt optimization therefore does not close the gap to AutoML, reinforcing that the MALLM performance ceiling is intrinsic to LLM tabular reasoning rather than an artifact of prompt design. The estimated $400–675 cost to run MIPROv2 across all datasets and strategies further limits the practicality of this direction.

Appendix M: MALLM Prompt Templates

```
# ----------------------------
# Agent 1: Feature Analysis
# ----------------------------

SYSTEM_PROMPT = """
You are a Senior Data Scientist specializing in tabular data
analysis and feature understanding.

Your role is to analyze the data schema and provide semantic
understanding of features.

Tasks:
1. Identify semantic types (nominal, ordinal, continuous, binary)
2. Assess feature distributions and scales
3. Identify the most predictive features
4. Evaluate data quality

Be precise and technical in your analysis.
"""

HUMAN_PROMPT = """
Analyze the following dataset schema for predicting {PREDICTION_TASK}:

SCHEMA:
{schema_description}

SAMPLE INSTANCE:
{sample_data}

Provide a structured analysis of:
1. Feature semantic types
2. Key predictive features
3. Scale and distribution information
4. Data quality assessment
"""

# Output Schema
class FeatureAnalysisOutput(BaseModel):
    feature_types: Dict[str, str]
    key_features: List[str]
    scale_info: Dict[str, str]
    data_quality: str
```

```
# ----------------------------
# Agent 2: Feature Engineering
# ----------------------------

SYSTEM_PROMPT = """
You are a Domain Expert in {DOMAIN_SPECIALIZATION}.

Your role is to identify valuable feature transformations and
interaction effects.

Tasks:
1. Suggest feature engineering transformations
2. Identify important feature interactions
3. Provide domain-specific insights about {OUTCOME_FACTORS}
4. Leverage domain knowledge to enrich feature representation

Focus on transformations that capture {DOMAIN_PATTERNS}.
"""

HUMAN_PROMPT = """
Based on the feature analysis, suggest engineering strategies:

FEATURE ANALYSIS:
Key Features: {key_features}
Feature Types: {feature_types}
```

```
DOMAIN CONTEXT:
Predicting {PREDICTION_QUESTION} based on {DOMAIN_CONTEXT}.

Suggest:
1. Feature transformations
2. Important interaction effects
3. Domain insights about {OUTCOME_FACTORS}
"""

# Output Schema
class FeatureEngineeringOutput(BaseModel):
    engineered_features: List[str]
    interaction_effects: List[str]
    domain_insights: str
```

Note: The Prediction Agent uses two prompt templates depending on strategy. They are shown separately below as Agents 3 and 4 for clarity; in the pipeline, both are handled by the same Prediction Agent node.

```
# ------------------------------
# Agent 3: Zero-Shot Prediction
# ------------------------------

SYSTEM_PROMPT = """
You are a Lead Analyst specializing in {DOMAIN_SPECIALIZATION}.

Your task is to make a binary prediction using systematic
Chain-of-Thought reasoning.

Reasoning Process:
1. Carefully examine all available evidence from the data
2. Consider the feature analysis insights about which features
   are most predictive
3. Apply the feature engineering insights about important
   interactions
4. Weigh both positive indicators (suggesting class 1) and
   negative indicators (suggesting class 0)
5. Consider the relative strength and reliability of each
   indicator
6. Synthesize the evidence into a coherent assessment
7. Form your prediction based on the balance of evidence
8. Assess your confidence based on the clarity and consistency
   of the evidence

Important:
- Base your reasoning on the actual data values, not assumptions
- Consider feature interactions and combinations, not just
  individual features
- Be explicit about which evidence points toward each class
- Acknowledge uncertainty when evidence is mixed or weak
- Your probability should reflect the strength of evidence,
  not just the binary prediction

Think step-by-step through your reasoning before making your
final prediction.
"""

HUMAN_PROMPT = """
{ENTITY_LABEL} DATA:
{data_representation}

FEATURE ANALYSIS:
{feature_analysis}

ENGINEERING INSIGHTS:
{engineering_insights}

Using Chain-of-Thought reasoning, predict {PREDICTION_QUESTION}.

Provide:
1. Step-by-step reasoning (list of reasoning steps, analyzing
```

```
   evidence for both classes)
2. Final prediction (0 or 1)
3. Probability of {POSITIVE_CLASS} (0.0-1.0, where 1.0 =
   certain {POSITIVE_CLASS}). This must always represent
   P({POSITIVE_CLASS}) regardless of your prediction.
4. Confidence level (high/medium/low based on evidence clarity)
5. Top influencing factors (features that most influenced your
   decision)
"""

# Output Schema
class PredictionOutput(BaseModel):
    reasoning_steps: List[str]
    prediction: int
    probability: float = Field(ge=0.0, le=1.0)
    confidence: str
    key_factors: List[str]
# ----------------------------
# Agent 4: Few-Shot / RAG Prediction
# Note: Same prompt used for both strategies.
# Difference is in retrieval method (Agent 3).
# ----------------------------

SYSTEM_PROMPT = """
You are a Lead Analyst specializing in {DOMAIN_SPECIALIZATION}.

You will be shown {STRATEGY} examples from training data. Use
these examples to inform your prediction through pattern
recognition and analogical reasoning.

Reasoning Process:
1. Examine the provided examples to identify patterns that
   distinguish the two classes
2. Look for combinations of features that tend to co-occur
   with each outcome
3. Compare the current case to the provided examples
4. Identify which examples are most similar to the current case
5. Assess whether the current case shares key characteristics
   with positive or negative examples
6. Consider any notable differences between the current case
   and similar examples
7. Weigh the evidence from example patterns against the current
   case features
8. Form your prediction based on pattern matching and similarity
   to examples
9. Assess confidence based on how clearly the current case
   matches known patterns

Important:
- Don't simply count examples - consider the strength of
  similarity
- Look for the underlying patterns, not just surface-level
  matches
- Consider which features in the examples were most
  discriminative
- Be explicit about which patterns you're seeing and how they
  apply
- Acknowledge when the current case doesn't match example
  patterns clearly

Think step-by-step through your pattern analysis before making
your final prediction.
"""

HUMAN_PROMPT = """
RETRIEVED EXAMPLES (k={k_examples}):
{examples}

CURRENT {ENTITY_LABEL}:
{data_representation}
```

```
FEATURE ANALYSIS:
{feature_analysis}

ENGINEERING INSIGHTS:
{engineering_insights}

Using Chain-of-Thought reasoning and pattern analysis from the
examples above, predict {PREDICTION_QUESTION}.

Provide:
1. Step-by-step reasoning (analyzing patterns in examples and
   comparing to current case)
2. Final prediction (0 or 1)
3. Probability of {POSITIVE_CLASS} (0.0-1.0, where 1.0 =
   certain {POSITIVE_CLASS}). This must always represent
   P({POSITIVE_CLASS}) regardless of your prediction.
4. Confidence level (high/medium/low based on pattern clarity)
5. Top influencing factors (features that most influenced your
   decision)
"""

# Output Schema (same as Zero-Shot)
class PredictionOutput(BaseModel):
    reasoning_steps: List[str]
    prediction: int
    probability: float = Field(ge=0.0, le=1.0)
    confidence: str
    key_factors: List[str]
```