Predicting Language Models' Success at Zero-Shot Probabilistic Prediction

Kevin Ren [1*,†], **Santiago Cortes-Gomez**[2], **Carlos Miguel Patiño**[2], **Ananya Joshi**[2], **Ruiqi Lyu**[2], **Jingjing Tang**[2], **Alistair Turcan**[2], **Khurram Yamin**[2], **Steven Wu**[2], **Bryan Wilder**[2]

[1]Cornell Tech, [2]Carnegie Mellon University
[†]Work done while at Carnegie Mellon University
Correspondence: kevinren@cs.cornell.edu

Abstract

Recent work has investigated the capabilities of large language models (LLMs) as zero-shot models for generating individual-level characteristics (e.g., to serve as risk models or augment survey datasets). However, when should a user have confidence that an LLM will provide high-quality predictions for their particular task? To address this question, we conduct a large-scale empirical study of LLMs' zero-shot predictive capabilities across a wide range of tabular prediction tasks. We find that LLMs' performance is highly variable, both on tasks within the same dataset and across different datasets. However, when the LLM performs well on the base prediction task, its predicted probabilities become a stronger signal for individual-level accuracy. Then, we construct metrics to predict LLMs' performance at the *task* level, aiming to distinguish between tasks where LLMs may perform well and where they are likely unsuitable. We find that some of these metrics, each of which are assessed without labeled data, yield strong signals of LLMs' predictive performance on new tasks [1].

1 Introduction

There is increasing interest in using large language models (LLMs) as predictive models, leveraging the world knowledge encoded by their pretraining corpora to make zero-shot predictions in domains without any labeled data. While this predictive capability was first investigated for traditional tasks within Natural Language Processing (NLP), such as text classification or question-answering (Wang et al., 2023b), recent work has utilized LLMs as predictive models in a broader sense. For instance, LLMs have been used to provide medical risk scores (Chung et al., 2024), predict fraud risk in financial applications (Xie et al., 2024) and impute unsurveyed fields in social science surveys (Park et al., 2024; Dominguez-Olmedo et al., 2024). More generally, LLMs can effectively consume text serializations of tabular data; the prevalence of tabular data across many domains likely contributes to this increasing interest across application areas. These applications differ from traditional text-based tasks (Cruz et al., 2024) because the label is not determined fully by the input: people with identical features may have different outcomes. We refer to tasks with this property as *probabilistic prediction*, and the predicted probabilities from the LLM as *risk scores*.

While the zero-shot prediction capabilities of LLMs offer exciting opportunities to scientists and practitioners, it is likely (as we empirically verify) that LLMs' performance varies widely across settings. Then, how can practitioners tell whether an LLM will perform well as a predictive model, prior to observing labeled data? This is a question with no easy answer. The appeal of using a pretrained model in many domains lies in avoiding the cost of collecting labeled data. However, validating conclusions from foundation models without labeled-data confirmation is far from straightforward.

This challenge is especially pronounced in the fully zero-shot case, where users lack access to ground-truth labels altogether. We distinguish performance at two levels of granularity: at the **individual** level, referring to which examples an LLM is likely to predict accurately, and at the **task** level, referring to which overall prediction problems, defined by a dataset and outcome variable, the LLM is likely to perform well on. The ability to quantify uncertainty at both levels allows practitioners to judge which individuals and overall predictive tasks may result in inaccurate predictions.

Previous work has primarily studied uncertainty at the individual level, finding mixed results. Abstention methods use measures of individual-level confidence to flag dubious predictions that should

[1]We release our code at https://github.com/kkr36/llm-eval/tree/camera-ready.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 18337–18363
November 4-9, 2025

be examined manually by a human expert, or ignored altogether (Tomani et al., 2024; Feng et al., 2024). However, both answer-token probabilities and verbalized confidence scores from LLMs have been found to be badly calibrated for probabilistic prediction (Cruz et al., 2024) and also for a variety of question-answering tasks (Xiong et al., 2023), typically due to overconfidence. Despite this, multiple approaches train a post-processing step to improve calibration using only the outputs or last-layer representations of models (Shen et al., 2024; Ulmer et al., 2024). Confidence scores have also been found to be useful in conformal prediction frameworks (Kumar et al., 2023; Mohri and Hashimoto, 2024), suggesting that they can be post-processed to yield informative decisions about when to provide specific information.

Analogously, practitioners may wish to know whether a task is likely suitable for an LLM before using its outputs, via some metric of uncertainty at the task level. Yet, to our knowledge, no previous work considers uncertainty quantification at the task level, at least in the context of probabilistic prediction. This presents a significant challenge, as in many real-world scenarios, practitioners would benefit from heuristics to assess whether LLMs will perform well *a priori*. However, doing so typically requires labeled data—a costly resource that pretrained models are meant to help avoid.

In this work, we conduct a large-scale empirical study on the performance of LLMs for probabilistic prediction on 316 tasks across 31 tabular datasets. The primary question we ask is: **given only unlabeled data, is it possible to anticipate how well the model will perform on a zero-shot prediction task?** We provide the first empirical evidence using task-level strategies to assess signals of LLM performance across prediction tasks. Additionally, we provide more nuanced results about individual-level uncertainty quantification; previous results on LLM calibration for probabilistic prediction (Cruz et al., 2024) are restricted to data from the US Census while we employ a much larger number of tabular datasets across many subject areas. Our empirical study reveals several findings that can inform how LLMs are employed and evaluated in predictive settings:

1. The *distribution* of LLMs' predictions on unlabeled data encodes substantial information about their suitability for a task. We propose simple heuristics and more elaborate model-based strategies that provide a strong signal of LLMs' predictive performance, using only unlabeled data. While we do not suggest that practitioners forgo labeled-data evaluation in high-stakes settings, our results could be useful to provide an initial assessment of which candidates from a set of prediction tasks are more promising for further development—or to screen out applications that have a lower chance of success.

2. At the task level, naive "elicited confidence" strategies (e.g., asking LLMs to rate their skill level given a description of the task) are comparatively unreliable predictors of success.

3. Substantial variation in LLMs' performance on different prediction tasks is *not* explained by broader patterns of "subject matter expertise"; within different tasks defined on the same dataset, predictive performance exhibits very high variance. This implies that attempts to validate LLMs' suitability must be specific to individual predictive tasks, and should not solely utilize information at a dataset or general subject level. For example, validating a social simulator by demonstrating that the LLM predicts observed fields well carries a high degree of risk because success on observed fields often fails to generalize to success on a specific, unobserved field.

4. At the individual level, LLMs' responses to probabilistic prediction tasks are typically poorly calibrated. Beyond overconfidence as reported in previous work (Cruz et al., 2024), we find that LLMs' responses in a given domain are often describable as simply being over- or under-predictions, where risk scores are consistently too large or too small.

5. Despite a lack of calibration in individual-level predictions, in many tasks, individual-level responses still provide an informative signal for abstention decisions because LLMs are more accurate on examples for which they output more extreme risk scores. This conclusion empirically holds even if the numerical scale of the scores is highly distorted. This echoes our first two findings at the task level: LLMs' responses contain considerable latent information about performance at both levels,

but this information often requires postprocessing to elicit meaningful results.

Our results provide a pathway towards more rigorous decisions about which tasks and individual instances are appropriate for LLMs.

2 Related Work

LLMs for Tabular Data: Recent work has shown that LLMs can effectively process tabular data using simple prompting strategies, achieving strong performance (Hegselmann et al., 2023). Pretrained models like TaBERT (Yin et al., 2020), TAPAS (Herzig et al., 2020), and TURL (Deng et al., 2021) focus on tabular data for QA tasks, while others leverage chain-of-thought prompting (Sui et al., 2023; Jin and Lu, 2023) and fact verification (Chen et al., 2020; Eisenschlos et al., 2020). Broader generalization strategies include UniPredict (Wang et al., 2023a) and instruction tuning (Yang et al., 2024). More recent efforts highlight LLMs' ability to perform zero-shot tabular predictions (Shi et al., 2024; Wen et al., 2023; Gardner et al., 2024). As opposed to developing methods to optimize LLMs for the purposes of understanding tabular data, our work seeks to empirically distinguish general factors predicting LLMs' success and failure across prediction tasks.

Elicited Confidence Scores From LLMs: LLM predictions on tabular data can suffer from pretraining-induced biases (Liu et al., 2024), and their uncertainty estimates are often poorly calibrated (Cruz et al., 2024). Methods like multicalibration and prompt-based scoring (Xiong et al., 2023; Detommaso et al., 2024) aim to improve calibration. In contrast to prior work, we primarily study uncertainty estimation at the task level. En route, we also provide a more nuanced picture of individual-level uncertainty on a wider range of tasks than previous work.

3 Methods

We describe our experimental setup, the problem of predicting LLM performance, and the set of proxy methods that we assess for performance prediction.

3.1 Experimental Setup

We conduct experiments on 31 tabular datasets spanning domains such as social surveys, finance, medicine, and transportation (see Appendix A.1 for details). Each dataset is associated with a binary classification task. Using the `folktexts` library (Cruz et al., 2024), we serialize 1,000 randomly sampled rows per dataset (or the full dataset if smaller) into text prompts, followed by a multiple-choice question requesting the label. Predicted probabilities (risk scores) are derived from the token-level output distribution. We evaluate four models that expose token-probability APIs: GPT-4o-Mini, GPT-4o, Mistral-7B-Instruct-v0.1, and Llama-3.1-8b-Instruct. Each model also generates a verbalized confidence score per row (see Appendix A.12 for details). Final evaluations use the ground-truth labels to compute accuracy, AUC, and expected calibration error (ECE).

Beyond the designated "label" column for each dataset, we also treat other features as additional prediction targets, expanding the number of tasks substantially. In each case, one feature is treated as the prediction target while the others serve as inputs. For continuous features, we define binary labels relative to the median to standardize outputs, while for categorical features, we predict whether the value equals the mode. Features with >70% missing values, or categorical features where >99% of rows equal the mode or <10% equal the mode, are excluded. We sample 10 features per dataset to construct auxiliary prediction tasks. For example, given features A, B, C, and an outcome D, we remove D and create three tasks, (A,B→C), (B,C→A), and (A,C→B). We then compute zero-shot predictions on each task and average the AUCs, yielding 285 additional proxy evaluations.

3.2 Predicting task-level performance

We define and empirically evaluate metrics for predicting LLMs' zero-shot performance over domains. Many of these are intuitive extensions of individual-level uncertainty quantification strategies to the task level, and part of our goal is to give practitioners guidance about which extensions perform well empirically and which do not. We group our strategies into several broad categories.

Task-level confidence elicitation: Perhaps the simplest strategy to predict LLMs' performance at a new task is to ask the LLM itself whether it will perform well, analogous to verbalized confidence strategies at the individual level (Tian et al., 2023). We provide the LLM with a text description of the dataset and its target variable (see Appendix A.12 for the exact prompt). We assess several strategies that prompt the LLM to output assessments of its

own expected performance, given that LLMs are sensitive to the manner in which information is elicited. **Direct AUC prediction** asks the LLM to output a prediction of its own AUC at the task. **Integer scoring** asks the LLM to rate its confidence at the task as a number between 1 (no confidence) and 5 (full confidence). Finally, **Decimal scoring** asks the LLM for a continuous rating between 0.0 (no confidence) and 1.0 (full confidence).

Aggregating individual-level confidence: We utilize LLM outputs for each row of a dataset, given a prediction task, to design proxies for task-level AUC. For each row, we obtain the risk score $\hat{p}_i$ and verbalized confidence score c_i. One natural strategy is to aggregate these individual-level measures of uncertainty to the task level, reasoning that LLMs will perform well on tasks where they are confident in many individual examples. We evaluate four metrics as proxies for task-level performance. First, **average confidence**, defined for task j as $\frac{1}{n_j}\sum_{i=1}^{n_j} c_i$ (where n_j is the number of samples for task j). Second, **average Maximum Class Probability (MCP)**, defined as $\frac{1}{n_j}\sum_{i=1}^{n_j} \max\{\hat{p}_i, 1-\hat{p}_i\}$. This measures how close predictions are to 0 or 1, which is a proxy for confidence. Finally, we include two additional metrics, **standard deviation of confidence** and **standard deviation of risk scores**, the empirical standard deviations of the sets $\{c_i\}$ and $\{\hat{p}_i\}$, respectively. These are motivated by the anecdotal observation that one common failure mode LLMs encounter is outputting (near) identical responses for every row. One potential proxy to account for this is simply whether the LLM makes a wide range of predictions.

Masking: Finally, we might think that an LLM will output high-quality predictions of a label y if it performs well at other predictive tasks on the same dataset: predicting each feature x^i from the other features x^{-i}. This procedure is motivated by the hypothesis that strong performance on these proxy tasks signals broader task-relevant understanding by the LLM. We collect risk scores from a sample of such masked prediction tasks for each dataset. The **masking** strategy takes the average of the AUCs in these simulated tasks as a proxy for the AUC from predicting the true label y.

4 Results

Our analysis is structured as follows. We begin by examining the zero-shot classification performance of the LLMs on our curated datasets, with a focus on the quality of individual-level predictions. We then broaden the scope to analyze the predictability of aggregate-level LLM performance across datasets. Results are shown for a spread of tested LLMs, with the full scope of results added to the Appendix.

4.1 Overall Trends in Performance.

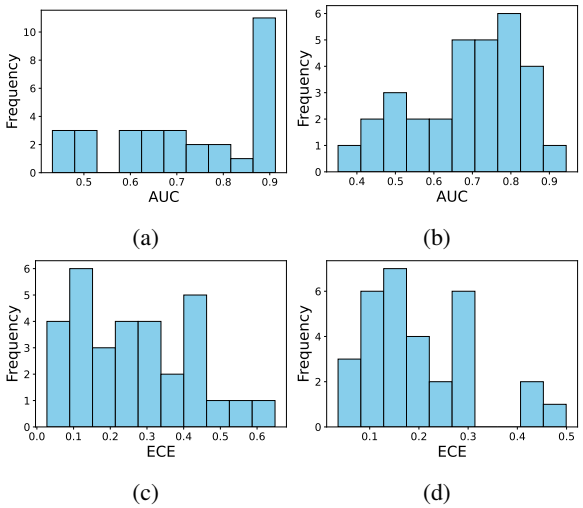


Figure 1: Histograms of AUC and ECE over all datasets, for GPT-4o-mini (a,c) and Llama-3.1-8b-Instruct (b,d).

LLMs have significant spread in their prediction capabilities, both across datasets and across prediction tasks from the same dataset. In Figures 1a and 1b, we observe that both GPT-4o-mini and Llama-3.1-8b-Instruct have nontrivial zero-shot predictive capabilities, with a median AUC of 0.7232 for the GPT-4o-mini and 0.7080 for the Llama model. The range is wide, with AUCs above 0.9 for some tasks, but at near-random (or worse than random) levels for others. This confirms that practitioners must take steps to assess the appropriateness of LLM zero-shot inference for a given task. See Appendix A.2 for a full set of AUC and ECE scores over all datasets and LLMs.

Within individual datasets, the quality of LLM predictions can vary substantially when using different columns as outcome variables (i.e., different prediction tasks). In Figure 2, we plot the distribution of AUC scores across columns within each dataset for GPT-4o-Mini (see Appendix A.6 for similar plots for the other LLMs.). These results show that intra-dataset variation is often considerable: many datasets contain prediction tasks with AUC scores below 0.5 as well as tasks with scores above 0.9. To quantify this result, we compute an

intra-class correlation coefficient, defined as the ratio of the variance in AUC within datasets vs overall, measuring the fraction of variance at the dataset level. We find that only 19% of the variance is explained by the dataset for GPT-4o-mini (for Llama-3.1-8b-Instruct, 12.80%), with 81% persisting within datasets. Perhaps surprisingly, this indicates that checking the performance of an LLM on some tasks in a given domain offers practitioners little confidence that it will perform well in unseen tasks from the same domain.

For deeper analysis, we also examine LLM performance relative to the *best achievable*, as some within-dataset variation may stem from inherent differences in column difficulty, independent of model skill. To test for dataset-level variation in relative LLM skill, we compute the ratio between the LLM's AUC and that of an XGBoost model trained on labeled data (Chen and Guestrin, 2016), as a proxy for optimal performance. This normalized metric is more concentrated within datasets than AUC (Appendix A.5), with the intraclass correlation increasing to 53.02% for GPT-4o-mini (47.68% for Llama-3.1-8b-Instruct), indicating more variation is explained at the dataset level. From the perspective of scientific understanding of LLMs' capabilities, this suggests there are meaningful differences in skill across domains after accounting for the inherent difficulty of a task (although practitioners may more heavily weigh absolute performance, where our earlier results show high within-dataset variation). Interestingly, GPT-4o-Mini's and Llama-3.1-8b-Instruct's AUC scores correlate strongly across tasks ($R^2 = 0.497$, Figure 3), suggesting certain tasks are more amenable to LLM-based inference than others. This correlation is stronger than either model's correlation with XGBoost performance (Figure 4), implying that shared LLM performance factors are not reducible to the difficulty of the base task. Analogous figures for GPT-4o and Mistral-7b-Instruct-v0.1 are shown in the Appendix A.7.

4.2 Individual-Level Results.

Elicited risk scores from LLMs are poorly calibrated, but are often useful for abstention tasks. Figure 1c and 1d show median ECEs around 0.2 for GPT-4o-mini and Llama (0.2426 and 0.1722, respectively), with GPT-4o-mini exhibiting greater variability. This corroborates previous findings of poor LLM calibration in US census tasks (Cruz et al., 2024) in a larger set of probabilistic prediction tasks. While prior work reports overconfident, inverted-sigmoid calibration curves from instruction-tuned models, we observe curves (see Figure 5, Appendix A.8) that often remain entirely above or below the identity line, indicating predictions are consistently too high or too low. This sug-

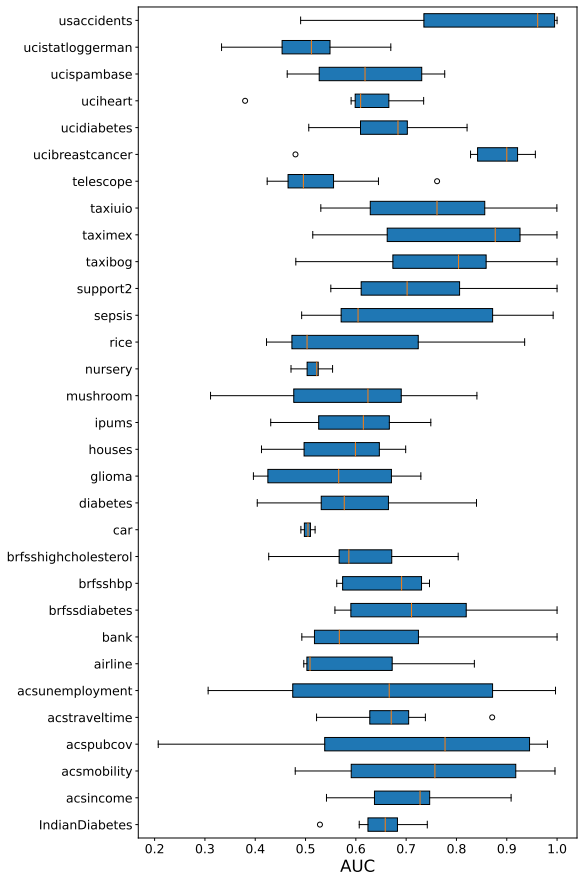


Figure 2: Box plots of AUC scores over masked-out columns in the Masking experiment, for all datasets. Results shown for GPT-4o-mini.

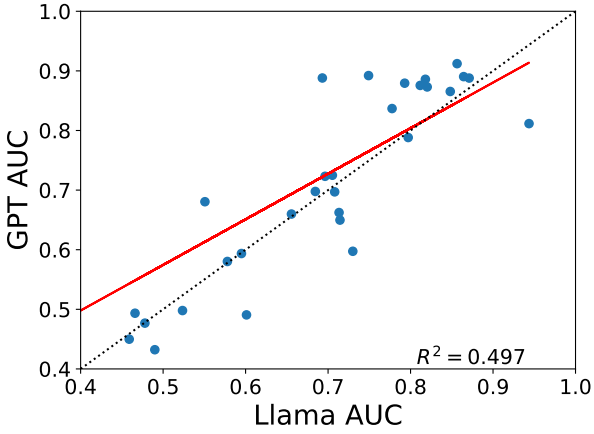
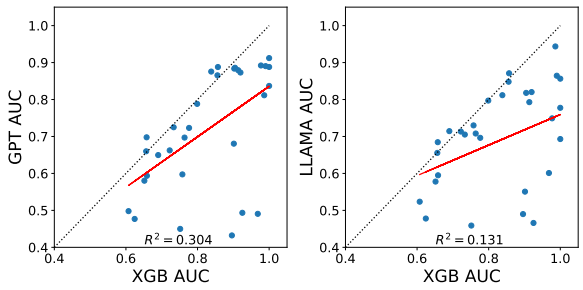
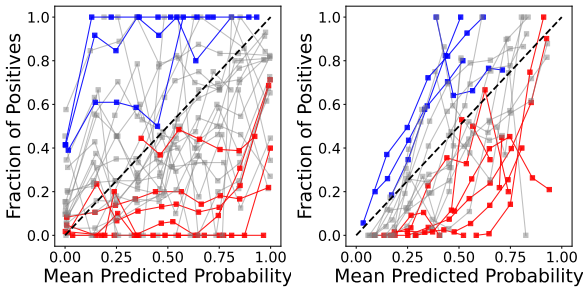


Figure 3: Plot of AUC scores for each of the datasets, for both Llama-3.1-8b-Instruct and GPT-4o-mini. Best-fit line with R^2 value plotted in red.



(a) (b)

Figure 4: Correlation between AUC scores of GPT-4o-mini (a) and Llama (b) over prediction tasks on each dataset, along with the AUCs achieved by training an XGBoost classification model on a subset of the training set, and evaluating on a disjoint validation set.

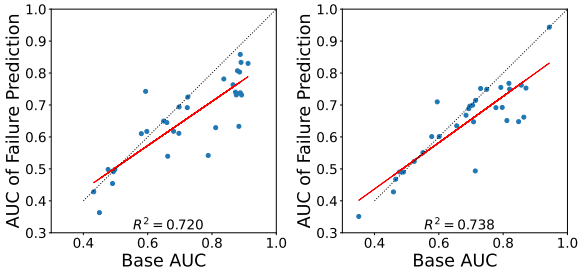


(a) (b)

Figure 5: Calibration curves for GPT-4o-mini (a) and Llama-3.1-8b-Instruct (b) across 31 datasets. Each curve corresponds to a prediction task. Curves crossing the identity line are shown in grey; those consistently above or below are blue and red, respectively. Concretely, all curves that a) are on average 0.2 above the identity line and b) have no points more than .1 below the identity line are colored in blue; curves on average .2 below the identity line and with no points more than .1 above are colored in red.

gests that LLMs often misjudge the absolute scale of their risk scores, even when preserving relative ranking accuracy (as reflected by high AUCs). Our findings thus contradict prior notions of overconfidence: instead, LLMs ostensibly have difficulty scaling their predictions to fit the marginal distribution of the label, even while correctly identifying which features correlate well with the label, which was a previously unknown phenomenon.

Despite poor numerical calibration, predictions closer to 0 or 1 (higher confidence) tend to be more accurate. We simulate abstention systems with LLM outputs by examining the degree to which MCP, a proxy of confidence in the predicted label, predicts individual-level accuracy, a task referred to by (Xiong et al., 2023) as *failure prediction*. We observe that LLM outputs are nontrivially successful at failure prediction for many tasks (see Figure 6, Appendix A.9). On the high end, we find AUCs for failure prediction of nearly 0.9, although performance varies across tasks (ranging from around 0.4 to 0.9). Strikingly, this effect is stronger for tasks where the LLM already performs well: the AUC of the original prediction task is highly correlated with AUC of failure prediction, indicating that ***when a model has a strong baseline ability, its confidence is better aligned with accuracy***. Thus, risk scores—despite calibration issues—can potentially support abstention strategies on domains where LLM usage is well-motivated to begin with, as LLMs often distinguish effectively between more and less reliable predictions on those tasks.



(a) (b)

Figure 6: Correlation between AUC scores of failure prediction and predicting the outcome variable for all datasets, for GPT-4o-mini (a) and Llama (b).

4.3 Task-Level Results.

LLMs vary widely in their ability to anticipate their performance on new tasks, correlating with their baseline strength. We observe that newer and more capable LLMs demonstrate the strongest general performance on tabular data. GPT-4o, the newest and largest LLM we study, achieves the highest median AUC across tasks (0.82) and has a nontrivial correlation between metrics that do not leverage any unlabeled data (the "direct AUC prediction," "decimal scoring," and "integer scoring" metrics) and its actual performance (Figure 8). There is also a strong correlation between GPT-4o's task-level performance and its average confidence on unlabeled data points for that task. However, these metrics display much weaker correlations with task-level performance for the other LLMs studied. These models have both weaker predictive

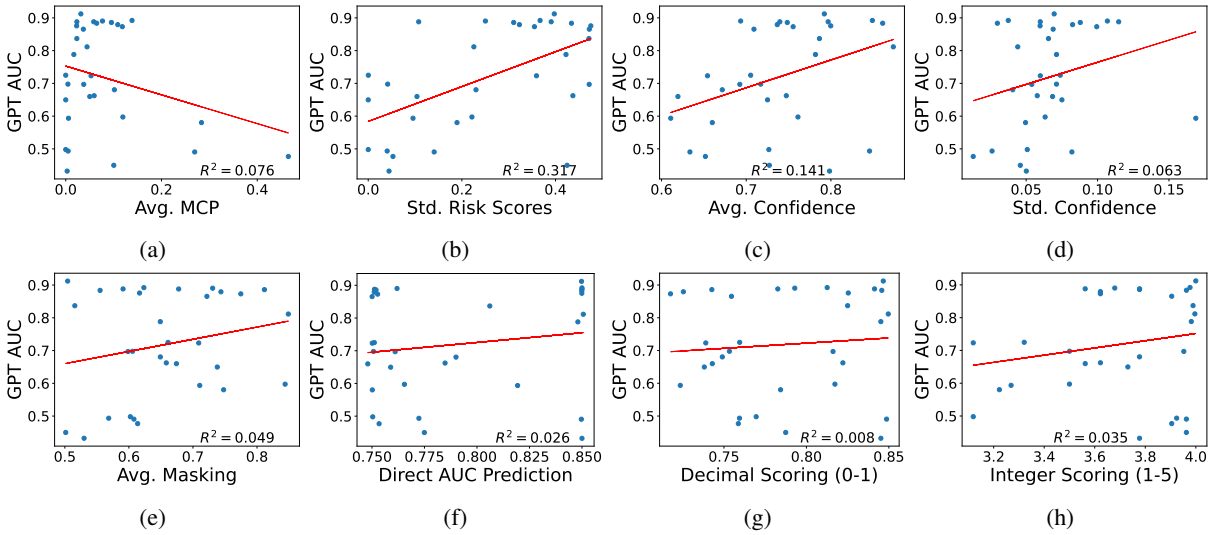


Figure 7: Correlation between aggregate metrics derived from our experiments on the unlabeled datasets and the AUC scores of GPT-4o-mini on each of the datasets, where each point represents one dataset. We plot the best-fit line with its corresponding R^2 value for each metric. See Appendix A.4 for the same set of plots made for Llama.

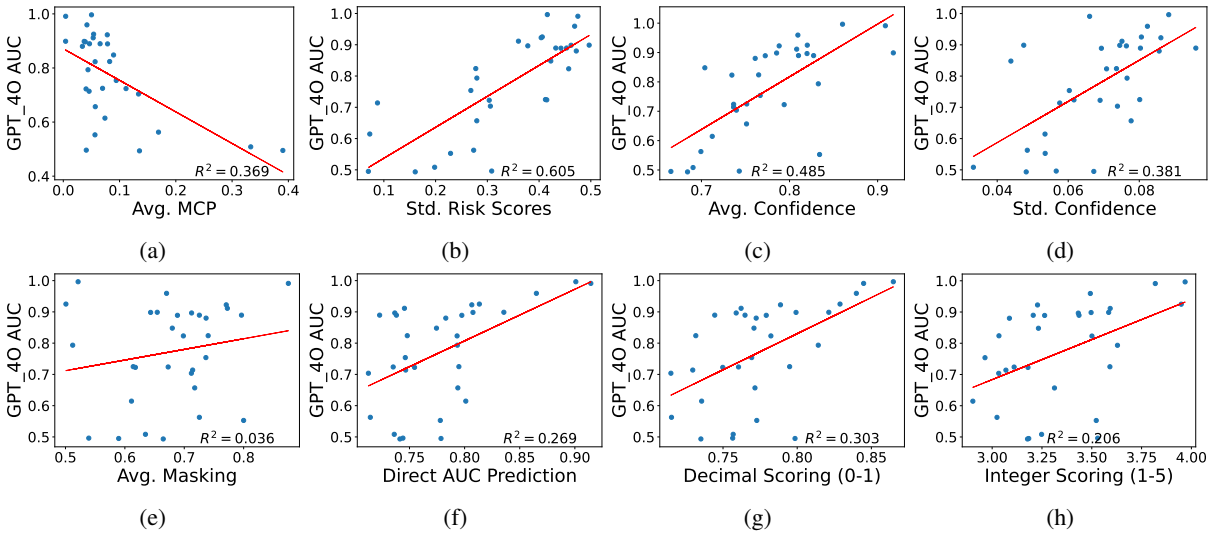


Figure 8: Correlation between aggregate metrics derived from our experiments on the unlabeled datasets and the AUC scores of GPT-4o on each of the datasets, where each point represents one dataset. We plot the best-fit line with its corresponding R^2 value for each metric.

performance and little correlation between their self-predicted performance and actual performance. We note that aggregate LLM performance is exactly in line with the strength of the LLM (i.e., parameter count): in order from lowest to highest average AUC across the 31 original tasks, we have Mistral-7b-Instruct-v0.1 (AUC: 0.66), Llama-3.1-8b-Instruct (AUC: 0.69), GPT-4o-mini (AUC: 0.72), and GPT-4o (AUC: 0.77). These findings suggest that an LLM's baseline capability correlates with its ability to anticipate its own performance, perhaps due to its inherent grasp of the prediction task.

The limitations of smaller models make this contrast even clearer. For smaller models (see Figure 7, Appendix A.4), most evaluated metrics show little correlation with AUC. In particular, methods that do not use any unlabeled data are entirely uncorrelated with performance. Even for GPT-4o, these metrics are weaker predictors of AUC than metrics that exploit unlabeled data, such as the standard deviation of risk scores (Figure 8). This suggests that confidence signals derived from unlabeled data are consistently more informative than self-estimates that rely solely on task descriptions, even while the strength of all metrics roughly scales with the

complexity of the LLM.

Surprisingly, the masking strategy (proxying the LLMs' performance at predicting a label by its performance at predicting features) poorly predicts downstream AUC for all tested LLMs. Although one might expect that an LLM's performance on masked columns would reflect its overall predictive capacity on a dataset, this assumption does not hold empirically. As shown in Figure 2, AUC scores vary widely across tasks within the same dataset, limiting the utility of dataset-averaged metrics. In other words, the variance in predictive quality across outcome variables makes it difficult for a dataset-level average to be a strong indicator of performance on any specific task within the dataset. As this appears to be a trend across many datasets, the results point to LLMs' capabilities as the mechanism: we find no evidence to support the expectation that LLMs are reliably good within some domains and consistently bad at others.

Information describing the spread of risk scores provides particularly strong signals for downstream performance. As shown in Figures 7b, 8b, 14b, 13b, the standard deviation of risk scores correlates positively with downstream AUC. For all tested models, the R^2 of this relationship is the highest among all metrics evaluated (e.g., $R^2 = 0.605$ and 0.270 for GPT-4o and Llama-3.1-8b-Instruct, respectively). A higher variance in risk scores may reflect greater separation between classes in a model's predictions, suggesting that some failure modes are distinguished by the model giving similar predictions for most rows. Importantly, there are significant outliers from this relationship, indicating that large variance in risk scores is not a guarantee of good performance on a task. Nevertheless, since this metric exhibits by far the largest correlation with predictive performance, we conduct a deeper dive by querying the distribution of model predictions for the 285 additional masked-column prediction tasks in addition to the 31 original tasks of predicting the designated label for each dataset. This gives us a significantly larger task-level sample size for more detailed analysis.

Checking the variance of risk scores can aid task-level abstention decisions. Aggregating results across all 316 tasks (Figures 9a and 9b for GPT-4o-mini and Llama; Appendix A.10 for GPT-4o and Mistral), we still observe a monotonically increasing relationship between the standard deviation of risk scores and AUC. To measure whether a practitioner would get an informative signal by screening potential tasks according to this metric, Figures 9c, 9d, 20c, and 20d show the mean AUC on all tasks above a given minimum threshold for the standard deviation, for all LLMs. By raising this threshold, we are able to distinguish tasks with significantly higher than average AUC. For instance, for GPT-4o-mini, the set of all datasets with a standard deviation in risk scores of at least 0.4 has an average AUC of 0.8417, much higher than the average AUC over all datasets (0.7186). While it is important not to rely on this metric absolutely, we suggest that practitioners check whether LLMs make similarly-valued predictions for all individuals, since doing so can help flag datasets where LLMs may not be suitable for zero-shot prediction.

The full distribution of predictions captures additional information about performance. As the standard deviation of the risk score distribution alone contains significant signal, we test whether additional information about performance can be gleaned from the full distribution of risk scores. For each task, we discretize the distribution of risk scores from the LLM into 201 values giving each α-percentile of the distribution, varying α by 0.5-percentile increments, and train XGBoost models to predict task-level AUC. We use 5-fold cross-validation, grouping by dataset to avoid leakage, so each task's out-of-sample prediction is based on the other 4 folds. Figure 10 plots the average out-of-sample predicted AUC against the actual AUC, along with a LOESS-smoothed curve and 95% confidence interval, for GPT-4o-mini and Llama (see Appendix 21 for an equivalent plot for GPT-4o and Mistral). The resulting trend is clearly positive, suggesting that the distribution of LLM-generated risk scores, computed solely on unlabeled data, contains meaningful information about task-level zero-shot performance. The relationship between predicted and actual AUCs becomes somewhat tighter than in Figure 9, particularly for Llama, suggesting that while the standard deviation of the distribution carries much of the signal about performance, other features of the distribution can contribute additional information. Despite the strong correlation between coarser metrics (i.e., standard deviation of risk scores) and downstream AUC for GPT-4o, we still find that the full distribution of risk scores adds meaningful information when regressing on AUC (see Appendix 21a, 21c).

To visualize what information the XGBoost models associate with high AUCs, Figure 11 shows the cumulative distribution functions (CDFs) of the

LLMs' risk scores for the 10 tasks with the highest and lowest predicted AUCs (see Appendix A.11 for an equivalent plot for GPT-4o and Mistral). Notably, results differ between LLMs. For GPT-4o-mini (Figure 11a), high AUC is associated with strongly bimodal risk scores, clustered near 0 or 1. In contrast, for Llama (Figure 11b), high AUC aligns with broader, high-variance distributions, while tighter, low-variance distributions correspond to lower AUCs. These differences suggest that the qualitative signals of good performance vary across LLMs. Although both LLMs encode useful information, the way this information manifests differs, indicating a need to analyze distributional traits on a per-model basis.

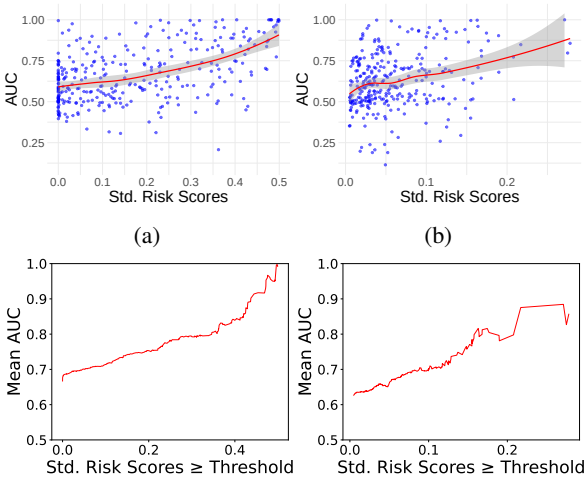


Figure 9: Proxy tasks in the Masking experiment using GPT-4o-mini (a,c) and Llama (b,d), including the original 31 tasks. LOESS curves with 95% CI shown in (a,b); each point represents predictions on one dataset column. (c,d) show average AUC as the minimum threshold on standard deviation of risk scores increases.

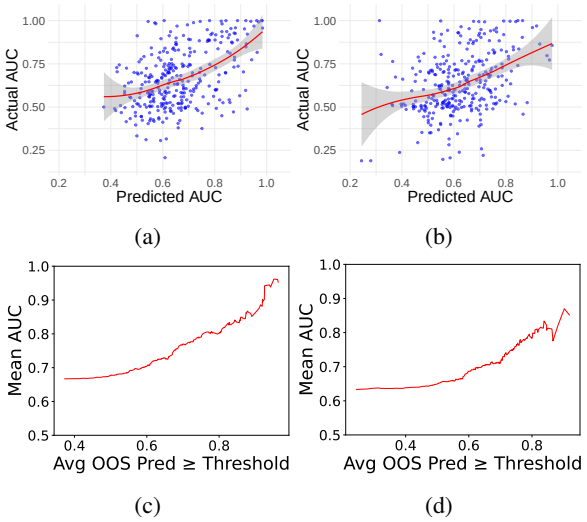


Figure 10: Proxy tasks in the Masking experiment using GPT-4o-mini (a,c) and Llama (b,d), including the original 31 tasks. (a,b) show LOESS fits (with 95% CI) of actual vs. XGBoost-predicted AUCs, trained via grouped 5-fold cross-validation. Each point represents one prediction task. (c,d) show AUC averages after thresholding on predicted AUCs, analogous to Figures 9c and 9d.

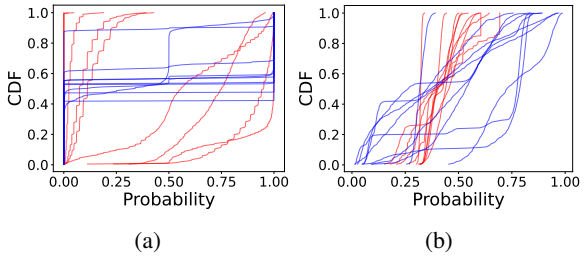


Figure 11: CDFs of the 10 highest (blue) and lowest (red) predicted AUCs over prediction tasks by XGBoost, using 201 percentile values along with standard deviation of risk scores to predict AUC. We observe clear trends within LLMs—for GPT-4o-mini (a), bimodal distributions of risk scores correlate with high XGBoost predictions, whereas for Llama (b), distributions encompassing a wide range of probabilities correlate with high predictions.

5 Conclusion

While the zero-shot prediction capabilities of LLMs offer exciting opportunities, it remains unclear how to *reliably* employ LLM predictions without validating their outputs on labeled data. We conduct a large-scale empirical study across 316 prediction tasks to explore whether LLMs can serve as reliable zero-shot predictors across a diverse collection of tabular classification tasks. We introduce eight novel task-level metrics for better estimating the LLMs' confidence in the prediction task.

Our findings indicate that performance is highly variable even within individual datasets, so success at one task is no guarantee of success at other tasks on similar data. Instead, measuring the distribution of risk scores for a new task yields both heuristics as well as more sophisticated models that capture a strong signal about the LLMs' performance on that task. However, enough variance in performance remains that such predictions of performance should be seen more as a way to prioritize more promising tasks or screen out ones with a low likelihood of success, not a substitute for eventual validation on labeled data for consequential applications.

Limitations

This paper investigates the predictive performance of large language models (LLMs) in zero-shot settings on tabular data, using unlabeled data to estimate task-level performance while drawing new conclusions about individual-level calibrations. While our findings offer novel insights, several limitations merit discussion:

Memorization or data leakage. The datasets that we use are publicly accessible, raising the prospect that they may have appeared in LLM training sets. Our results do imply that LLMs have not memorized the data in the sense of perfectly replicating individual rows, as AUCs vary widely at predicting individual columns within the same dataset given the other columns. Our serialization strategy also alters the presentation of information from the original csv file, which has been found to disrupt some explicit memorization (Bordt et al., 2024). Beyond literal row-by-row memorization though, previous work shows that LLMs perform better at tasks seen more during training, especially for tasks related to retrieval of world knowledge (Kandpal et al., 2023; Wang et al., 2025). The impact of this phenomena depends on the application at hand—practitioners in many settings may hope to actually benefit from LLMs having seen relevant data to their application during the training process. Accordingly, proxies for task-level performance that partly pick up on prior exposure to similar tasks may still serve their needs. However, using public data does represent a potential limitation in external validity for our results; we can't rule out that predictors of task-level performance might be different in domains that are completely unseen during LLM training.

Model access and scale. We rely on LLMs that expose token-level probabilities (GPT-4o-mini, GPT-4o, Mistral-7b-Instruct-v0.1, and Llama-3.1-8b-Instruct), which may not generalize to other models without such access or with substantially different architectures. Larger models, or models with distinct fine-tuning or pretraining regimes, may behave differently.

Prompting. Our serialization of tabular data uses fixed, template-based formats (i.e., "Feature: Value" pairs, followed by binary questions). Our prompting approaches do not explore alternative prompts, few-shot settings, or chain-of-thought reasoning.

References

Sebastian Bordt, Suraj Srinivas, Valentyn Boreiko, and Ulrike von Luxburg. 2024. How much can we forget about data contamination? *arXiv preprint arXiv:2410.03249*.

Devendra Singh Chaplot. 2023. Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed. *arXiv preprint arXiv:2310.06825*, 3.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. *arXiv preprint arXiv:1603.02754*.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*. Preprint: arXiv:1909.02164.

Philip Chung, Christine T Fong, Andrew M Walters, Nima Aghaeepour, Meliha Yetisgen, and Vikas N O'Reilly-Shah. 2024. Large language model capabilities in perioperative risk prediction and prognostication. *JAMA surgery*, 159(8):928–937.

André F. Cruz, Moritz Hardt, and Celestine Mendler-D"unner. 2024. Evaluating language models as risk scores. *arXiv preprint arXiv:2407.14614*.

Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. 2021. Turl: Table understanding through representation learning. *Proceedings of the VLDB Endowment*, 14(3):307–319.

Gianluca Detommaso, Martin Bertran, Riccardo Fogliato, and Aaron Roth. 2024. Multicalibration for confidence scoring in llms. *arXiv preprint arXiv:2404.04689*.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. *arXiv preprint arXiv:2108.04884*.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. *arXiv preprint arXiv:2306.07951*. NeurIPS 2024.

Dheeru Dua and Casey Graff. 2019. Uci machine learning repository. `https://archive.ics.uci.edu/ml/index.php`. Accessed May 2025.

Julian Martin Eisenschlos, Syrine Krichene, and Thomas M"uller. 2020. Understanding tables with intermediate pre-training. *arXiv preprint arXiv:2010.00571*.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*.

Josh Gardner, Juan C. Perdomo, and Ludwig Schmidt. 2024. Large scale transfer learning for tabular data via language modeling. *arXiv preprint arXiv:2406.12031*.

Josh Gardner, Zoran Popovic, and Ludwig Schmidt. 2023. Benchmarking distribution shift in tabular data with tableshift. *arXiv preprint arXiv:2312.07577*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

IPUMS Global Health. Demographic and health surveys (dhs). https://globalhealth.ipums.org/. Accessed May 2025.

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 (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*, pages 5549–5581. PMLR.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas M"uller, Francesco Piccinno, and Julian Eisenschlos. 2020. Tapas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333. Association for Computational Linguistics.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ziqi Jin and Wei Lu. 2023. Tab-cot: Zero-shot tabular chain of thought. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10259–10277. Association for Computational Linguistics.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Bhawesh Kumar, Charlie Lu, Gauri Gupta, Anil Palepu, David Bellamy, Ramesh Raskar, and Andrew Beam. 2023. Conformal prediction with large language models for multi-choice question answering. *arXiv preprint arXiv:2305.18404*.

Yanchen Liu, Srishti Gautam, Jiaqi Ma, and Himabindu Lakkaraju. 2024. Confronting llms with traditional ml: Rethinking the fairness of large language models in tabular classifications. In *Proceedings of NAACL-HLT 2024 (Long Papers)*, pages 3603–3620. Association for Computational Linguistics.

Christopher Mohri and Tatsunori Hashimoto. 2024. Language models with conformal factuality guarantees. In *International Conference on Machine Learning*, pages 36029–36047. PMLR.

Sobhan Moosavi. 2020. Us accidents (3.0). https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents. Accessed May 2025.

Mauricio Navas. 2022. Latin american taxi ride classification. https://www.kaggle.com/datasets/mnavas/taxi-routes-for-mexico-city-and-quito. Accessed May 2025.

OpenML. Openml data repository. https://www.openml.org. Accessed May 2025.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Maohao Shen, Subhro Das, Kristjan Greenewald, Prasanna Sattigeri, Gregory Wornell, and Soumya Ghosh. 2024. Thermometer: Towards universal calibration for large language models. *arXiv preprint arXiv:2403.08819*.

Zhiyi Shi, Junsik Kim, Davin Jeong, and Hanspeter Pfister. 2024. Surprisingly simple: Large language models are zero-shot feature extractors for tabular and text data. In *ICLR 2025 Conference*. Withdrawn submission.

Yuan Sui, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2023. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. *arXiv preprint arXiv:2305.13062*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. *arXiv preprint arXiv:2305.14975*.

Christian Tomani, Kamalika Chaudhuri, Ivan Evtimov, Daniel Cremers, and Mark Ibrahim. 2024. Uncertainty-based abstention in llms improves safety and reduces hallucinations. *arXiv preprint arXiv:2404.10960*.

UCI and Kaggle Contributors. Pima indians diabetes dataset. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. Accessed May 2025.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Joon Oh. 2024. Calibrating large language models using their generations only. *arXiv preprint arXiv:2403.05973*.

Ruiyu Wang, Zifeng Wang, and Jimeng Sun. 2023a. Unipredict: Large language models are universal tabular classifiers. *arXiv preprint arXiv:2310.03266*.

Xinyi Wang, Antonis Antoniades, Yanai Elazar, Alfonso Amayuelas, Alon Albalak, Kexun Zhang, and William Yang Wang. 2025. Generalization v.s. memorization: Tracing language models' capabilities back to pretraining data. In *The Thirteenth International Conference on Learning Representations*.

Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2023b. Large language models are zero-shot text classifiers. *arXiv preprint arXiv:2312.01044*.

Xumeng Wen, Han Zhang, Shun Zheng, Wei Xu, and Jiang Bian. 2023. From supervised to generative: A novel paradigm for tabular deep learning with large language models. *arXiv preprint arXiv:2310.07338*.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, and 1 others. 2024. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*. Accepted at ICLR 2024.

Yazheng Yang, Yuqi Wang, Yaxuan Li, Sankalok Sen, Lei Li, and Qi Liu. 2024. Unleashing the potential of large language models for predictive tabular tasks in data science. *arXiv preprint arXiv:2403.20208*.

Pengcheng Yin, Graham Neubig, Wen-Tau Yih, and Sebastian Riedel. 2020. Tabert: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426. Association for Computational Linguistics.

A Appendix

A.1 Dataset Descriptions, Sources, and Artifacts

Dataset Name	Short Description	Source
acsincome	ACSIncome task from the folktables package.	link
acsmobility	ACSMobility task from the folktables package.	link
acspubcov	ACSPublicCoverage task from the folktables package.	link
acstraveltime	ACSTravelTime task from the folktables package.	link
acsunemployment	ACSEmployment task from the folktables package.	link
airline	Predict flight delays based on scheduled departure info.	link
bank	Predict term deposit subscription in a marketing campaign.	link
brfssdiabetes	Predict whether a patient has diabetes (BRFSS survey).	link
brfsshbp	Predict hypertension diagnosis for 50+ age group.	link
brfsshighcholesterol	Predict high cholesterol in BRFSS survey data.	link
car	Predict acceptability of cars from evaluation records.	link
diabetes	Predict readmission of diabetic patients within 30 days.	link
glioma	Classify glioma (brain tumor) grade.	link
houses	Predict if California housing value exceeds $200k.	link
indiandiabetes	Predict diabetes using diagnostic features.	link
ipums	Predict facility birth in Latin/Caribbean countries.	link
mushroom	Classify mushrooms as edible or poisonous.	link
nursery	Prioritize nursery school applications.	link
rice	Classify Turkish rice grains as Osmancik or Cammeo.	link
sepsis	Predict ICU patient risk of sepsis within 6 hours.	link
support2	Predict hospital death of critically ill patients.	link
taxibog	Predict long taxi rides in Bogota.	link
taximex	Predict long taxi rides in Mexico City.	link
taxiuio	Predict long taxi rides in Quito.	link
telescope	Classify cosmic ray vs gamma signal events.	link
ucibreastcancer	Predict breast mass as malignant or benign.	link
ucidiabetes	Predict diabetes using lifestyle statistics.	link
uciheart	Predict heart disease diagnosis.	link
ucispambase	Classify email as spam or not spam.	link
ucistatloggerman	Classify credit risk from attributes.	link
usaccidents	Predict severity of US traffic accidents.	link

For more details regarding our dataset sources and other artifacts:

- The `car`, `diabetes`, `glioma`, `mushroom`, `nursery`, `rice`, `support2`, `telescope`, `ucibreastcancer`, `ucidiabetes`, `uciheart`, `ucispambase`, and `ucistatloggerman` datasets all come from the UCI repository (Dua and Graff, 2019).
- The `acsincome`, `acsmobility`, `acspubcov`, `acstraveltime`, and `acsunemployment` datasets all come from the Folktables repository (Ding et al., 2021).
- The `brfssdiabetes`, `brfsshbp`, `brfsshighcholesterol`, and `sepsis` datasets all come from the Tableshift repository (Gardner et al., 2023).
- The `airline`, `bank`, and `house` datasets all come from OpenML (OpenML).
- The `indiandiabetes`, `taxibog`, `taximex`, `taxiuio`, and `usaccidents` datasets all come from Kaggle (UCI and Contributors; Navas, 2022; Moosavi, 2020).
- the `ipums` dataset is curated from the IPUMS Global Health repository of international health survey data (Health).
- All datasets are publicly available and we release our data for replication, with the exception of the `ipums` data, which required individual-level dataset requests for the de-identified data on maternal outcomes, and is thus not released. All data is compliant with anonymization policies (i.e., de-identified) and does not contain offensive or sensitive content.

- We use GPT-4o-mini, GPT-4o (Hurst et al., 2024), Mistral-7b-Instruct-v0.1 (Chaplot, 2023), and Llama-3.1-8b-Instruct (Grattafiori et al., 2024) as LLMs for predictive modeling, for all experiments. All models contain publicly available APIs for personal and research use. Furthermore, Mistral-7b-Instruct-v0.1 and Llama-3.1-8b-Instruct make their model weights publicly available.

A.2 LLM Metrics Table

Dataset	GPT-4o-mini AUC	GPT-4o-mini ECE	Llama AUC	Llama ECE
taxiuio	0.8794	0.0971	0.7929	0.3087
mushroom	0.8881	0.2900	0.6931	0.1676
acsincome	0.8655	0.1939	0.8481	0.2812
support2	0.8904	0.1369	0.8644	0.2953
telescope	0.4322	0.6490	0.4900	0.3038
nursery	0.8368	0.2425	0.7776	0.1163
diabetes	0.4979	0.0960	0.5235	0.1940
brfssdiabetes	0.6497	0.1540	0.7144	0.0706
airline	0.4768	0.0697	0.4779	0.1867
bank	0.6805	0.1115	0.5507	0.0854
acspubcov	0.7232	0.2211	0.6963	0.0723
ucistatloggerman	0.4499	0.4677	0.4589	0.4457
brfsshbp	0.7249	0.4550	0.7052	0.1633
usaccidents	0.5974	0.1980	0.7300	0.1535
uciheart	0.8756	0.2504	0.8117	0.1348
IndianDiabetes	0.7882	0.4330	0.7971	0.0877
taxibog	0.8730	0.0770	0.8202	0.1923
ucispambase	0.8921	0.2954	0.7491	0.1632
ucidiabetes	0.6624	0.3149	0.7133	0.4992
glioma	0.8837	0.2426	0.3511	0.2808
rice	0.4907	0.3090	0.6011	0.2925
acstraveltime	0.6599	0.3724	0.6556	0.0357
acsmobility	0.5803	0.1427	0.5779	0.1153
car	0.9121	0.1308	0.8564	0.1067
acsunemployment	0.8880	0.4190	0.8711	0.2171
houses	0.4935	0.4083	0.4659	0.1077
sepsis	0.5936	0.0273	0.5950	0.2442
brfsshighcholesterol	0.6977	0.4171	0.6846	0.1722
ucibreastcancer	0.8115	0.5732	0.9436	0.4302
taximex	0.8859	0.0494	0.8180	0.2640
ipums	0.6970	0.3519	0.7080	0.1563

Dataset	GPT-4o AUC	GPT-4o ECE	Mistral AUC	Mistral ECE
taxiuio	0.9115	0.0845	0.7166	0.4622
mushroom	0.9964	0.2077	0.8277	0.2423
acsincome	0.8801	0.1696	0.8102	0.1016
support2	0.9226	0.0921	0.7242	0.5444
telescope	0.6145	0.2760	0.6420	0.0480
nursery	0.7935	0.2672	0.8622	0.1841
diabetes	0.4935	0.1304	0.4973	0.2322
brfssdiabetes	0.8239	0.0894	0.7785	0.1827
airline	0.5084	0.1529	0.5114	0.2466
bank	0.8480	0.4554	0.5559	0.1709
acspubcov	0.7238	0.2866	0.6122	0.0780
ucistatloggerman	0.4962	0.3348	0.4436	0.2079
brfsshbp	0.7538	0.2986	0.6665	0.0886
usaccidents	0.5527	0.2241	0.6028	0.3827
uciheart	0.8983	0.2046	0.7593	0.2384
IndianDiabetes	0.8231	0.2400	0.7723	0.1712
taxibog	0.8895	0.1253	0.7495	0.3091
ucispambase	0.9594	0.1051	0.7447	0.0519
ucidiabetes	0.7248	0.5399	0.6910	0.2811
glioma	0.8988	0.1425	0.4919	0.1840
rice	0.4950	0.0449	0.2742	0.0616
acstraveltime	0.7038	0.3325	0.5685	0.1266
acsmobility	0.5628	0.1987	0.5293	0.1105
car	0.9250	0.0986	0.8565	0.2582
acsunemployment	0.8896	0.1346	0.5654	0.2648
houses	0.8890	0.1692	0.4260	0.0345
sepsis	0.6568	0.1264	0.5929	0.3422
brfsshighcholesterol	0.7139	0.3755	0.6982	0.0806
ucibreastcancer	0.9911	0.0358	0.9597	0.3950
taximex	0.8965	0.0868	0.7839	0.4085
ipums	0.7225	0.1912	0.6658	0.2763

A.3 Additional AUC and ECE Histograms

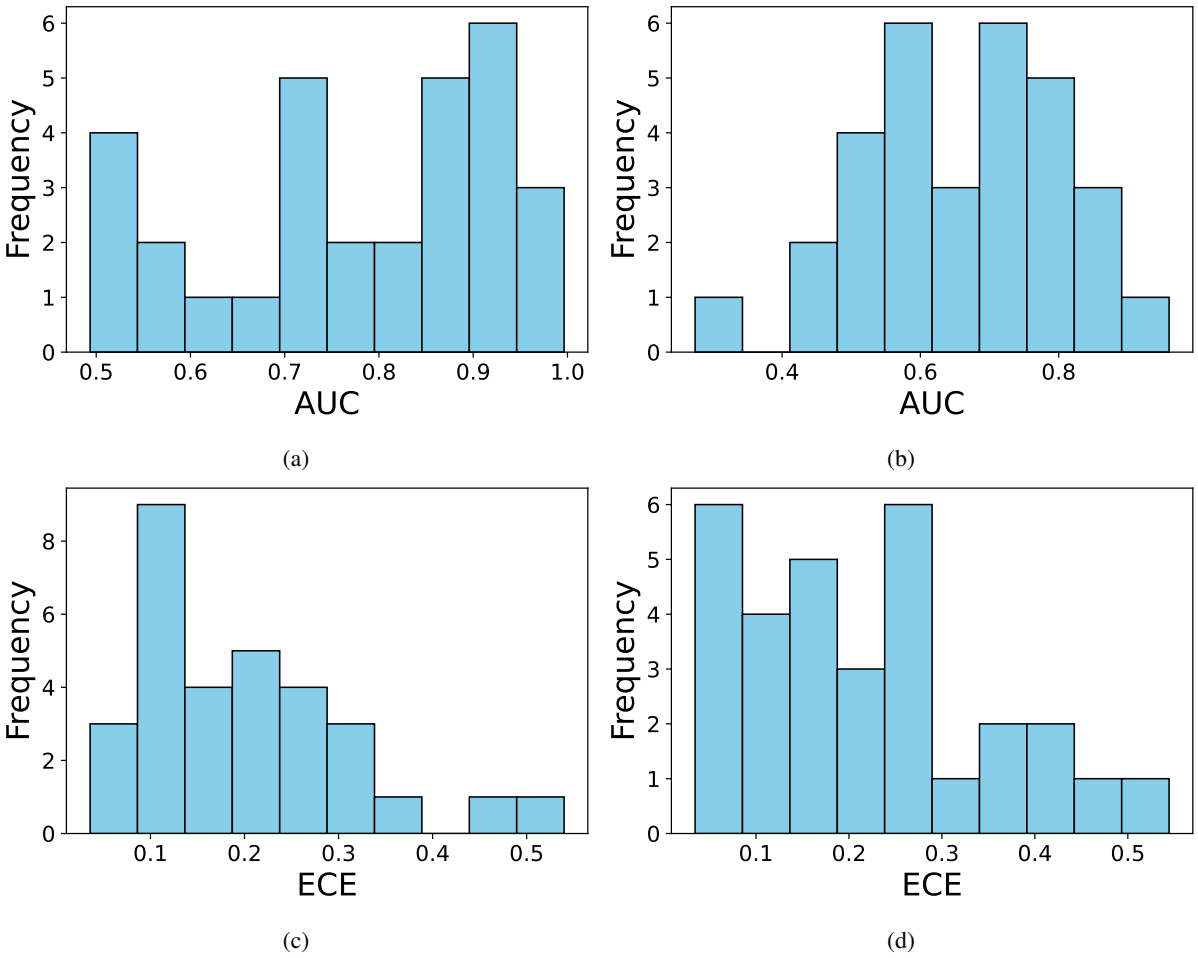


Figure 12: Histograms of AUC and ECE over all datasets, for GPT-4o (a,c) and Mistral-7b-Instruct-v0.1 (b,d).

A.4 Correlation between Metrics and AUC

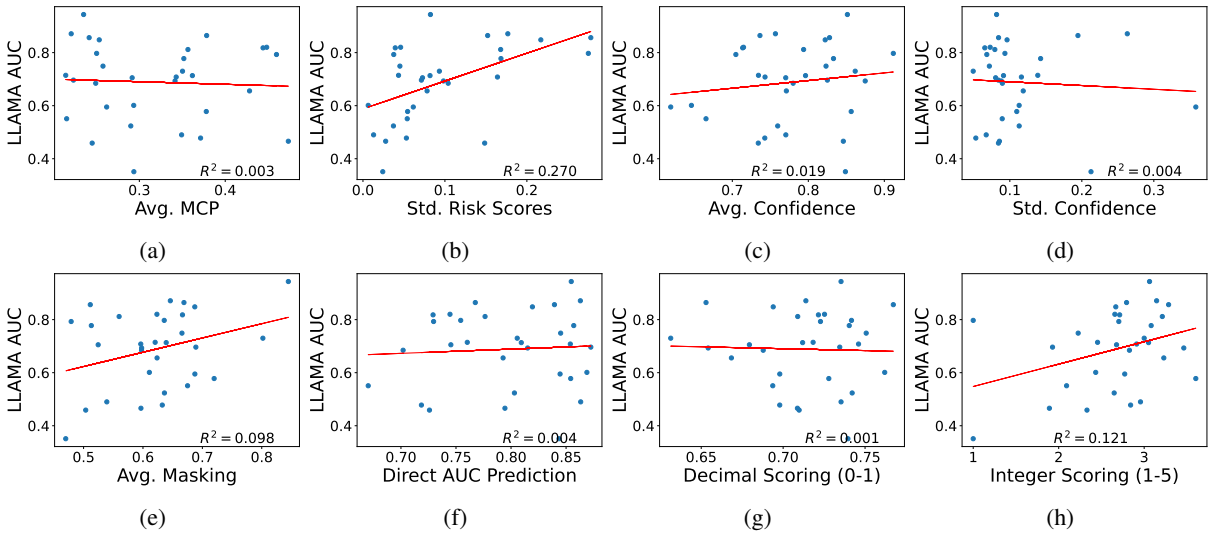


Figure 13: Correlation between aggregate metrics derived from our experiments on the unlabeled datasets and the AUC scores of Llama-3.1-8b-Instruct on each of the datasets, where each point represents one dataset. We plot the best-fit line with its corresponding R^2 value for each metric.

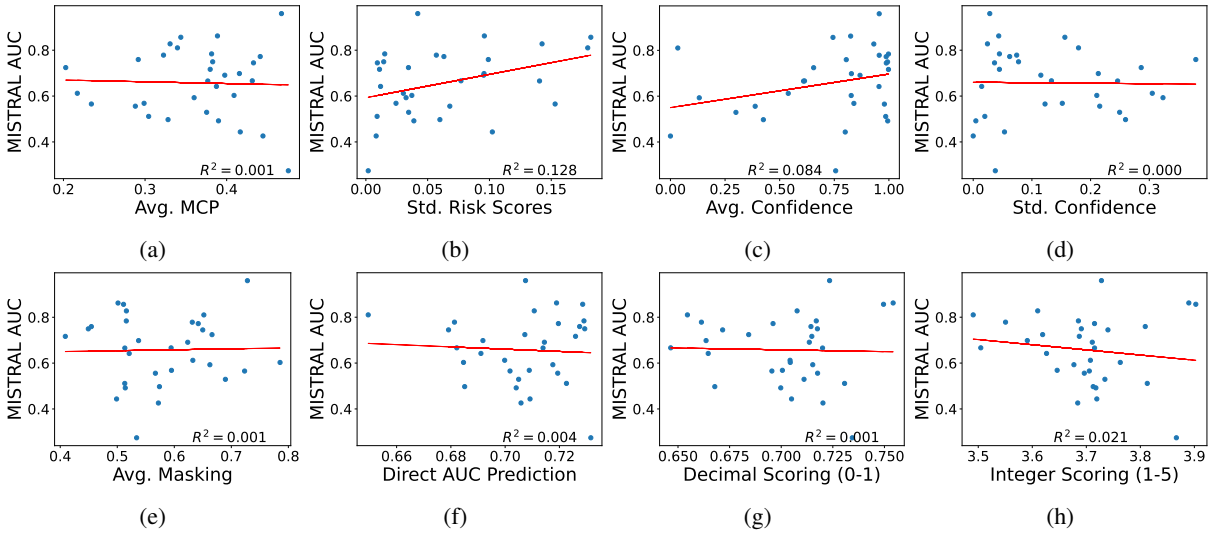


Figure 14: Correlation between aggregate metrics derived from our experiments on the unlabeled datasets and the AUC scores of Mistral-7B-Instruct-v0.1 on each of the datasets, where each point represents one dataset. We plot the best-fit line with its corresponding R^2 value for each metric.

A.5 Normalized AUC Scores, Masking Experiment

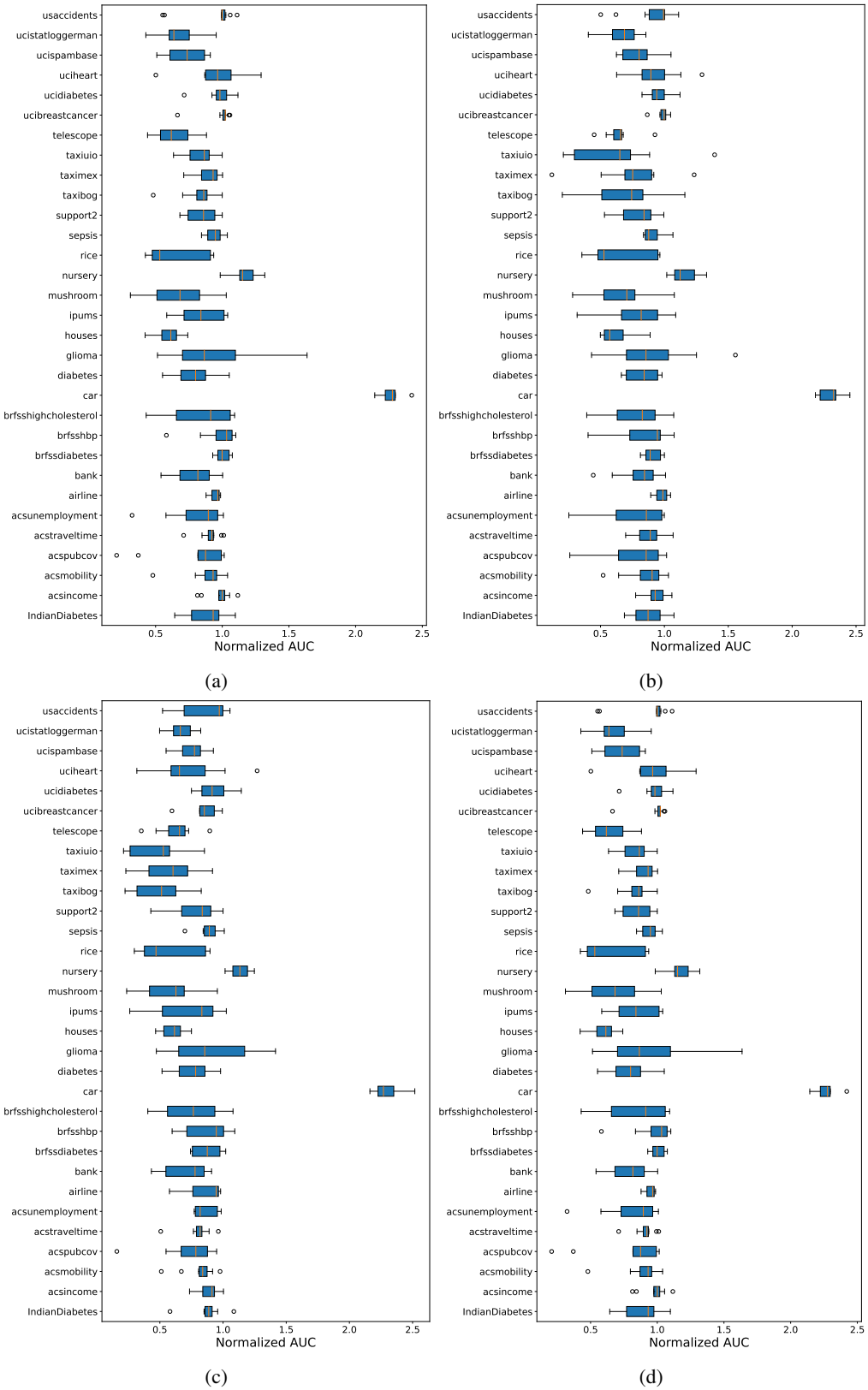
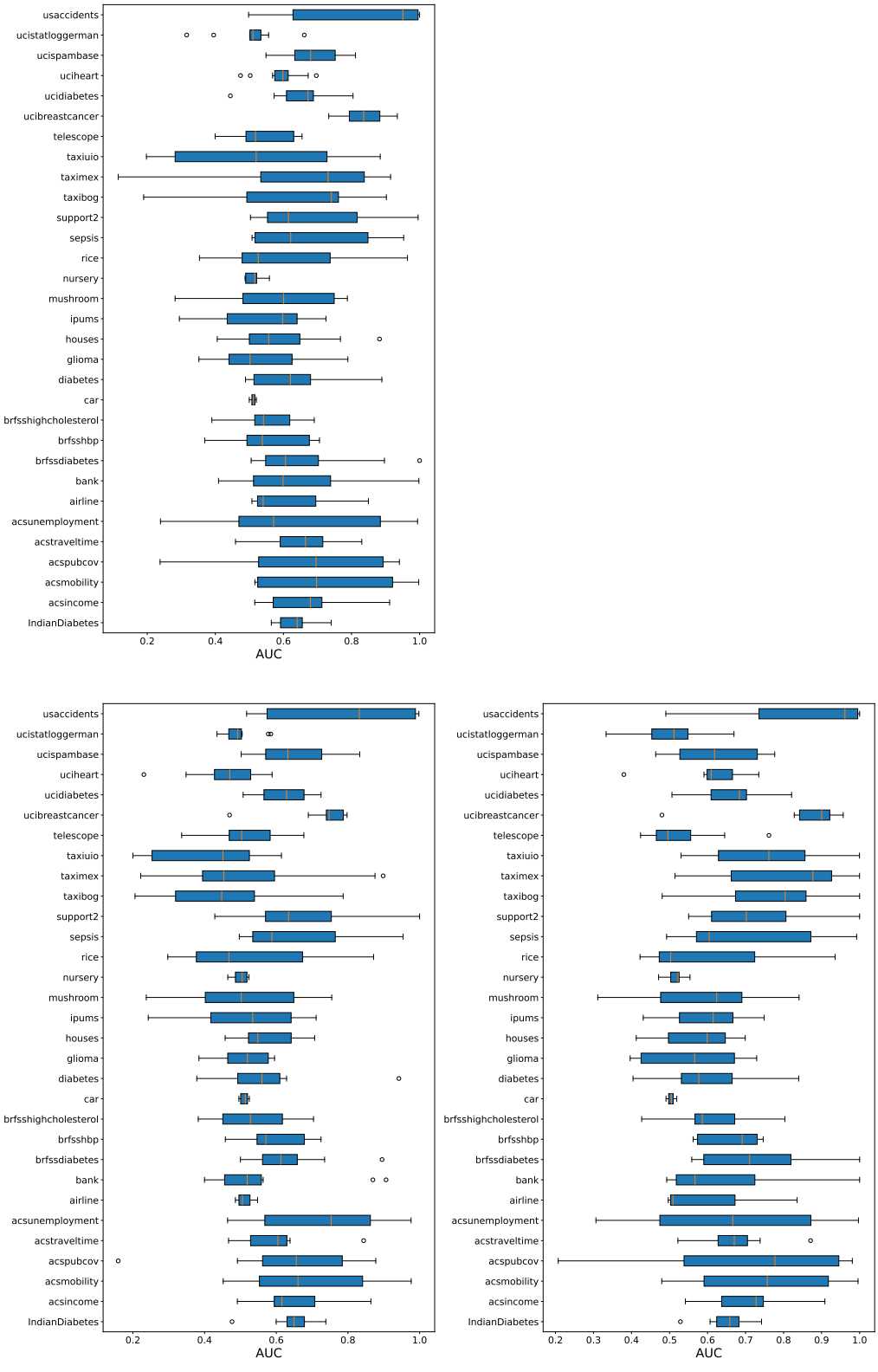


Figure 15: Box plots of AUC scores over masked-out columns in the Masking experiment, for all datasets, where each AUC is divided by the AUC achieved by an XGBoost classifier on the same prediction task. Results shown for GPT-4o-mini (a), Llama (b), Mistral (c), and GPT-4o (d).

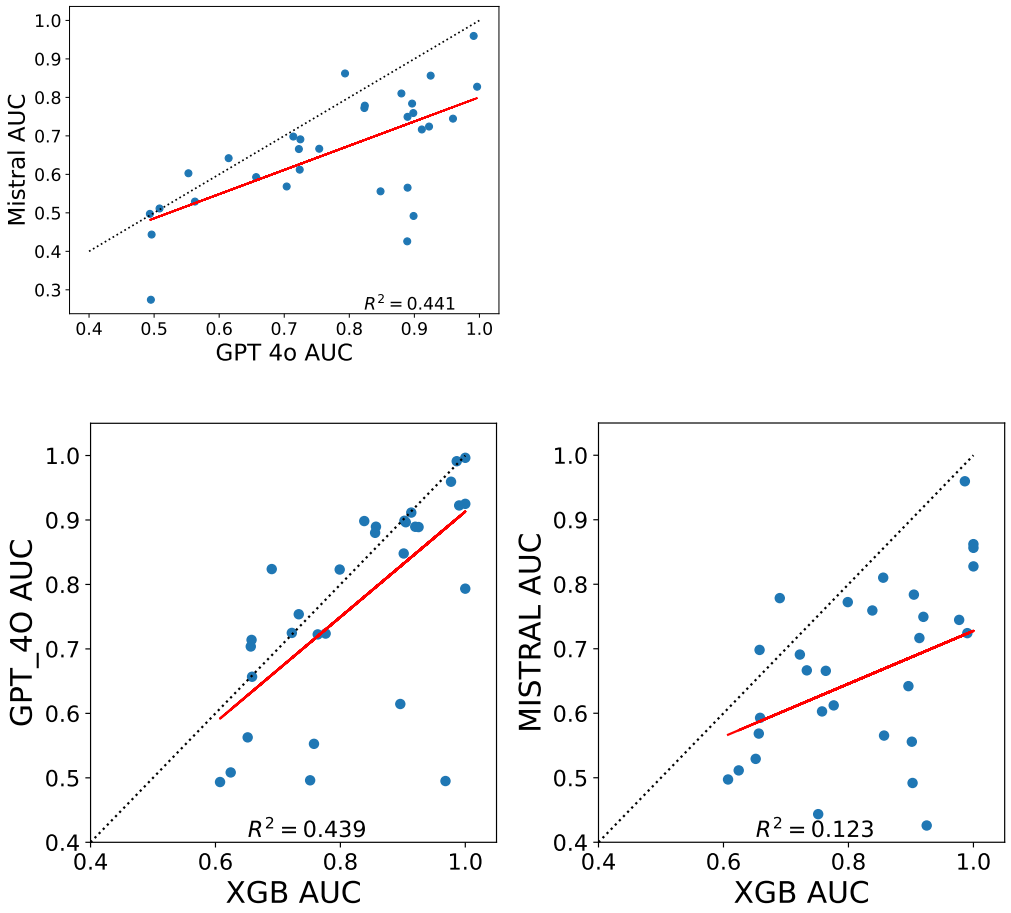
A.6 AUC Scores, Masking Experiment



(a) (b)

Figure 16: Box plots of AUC scores over masked-out columns in the Masking experiment, for all datasets. Results shown for Llama (a), Mistral (b), and GPT-4o (c).

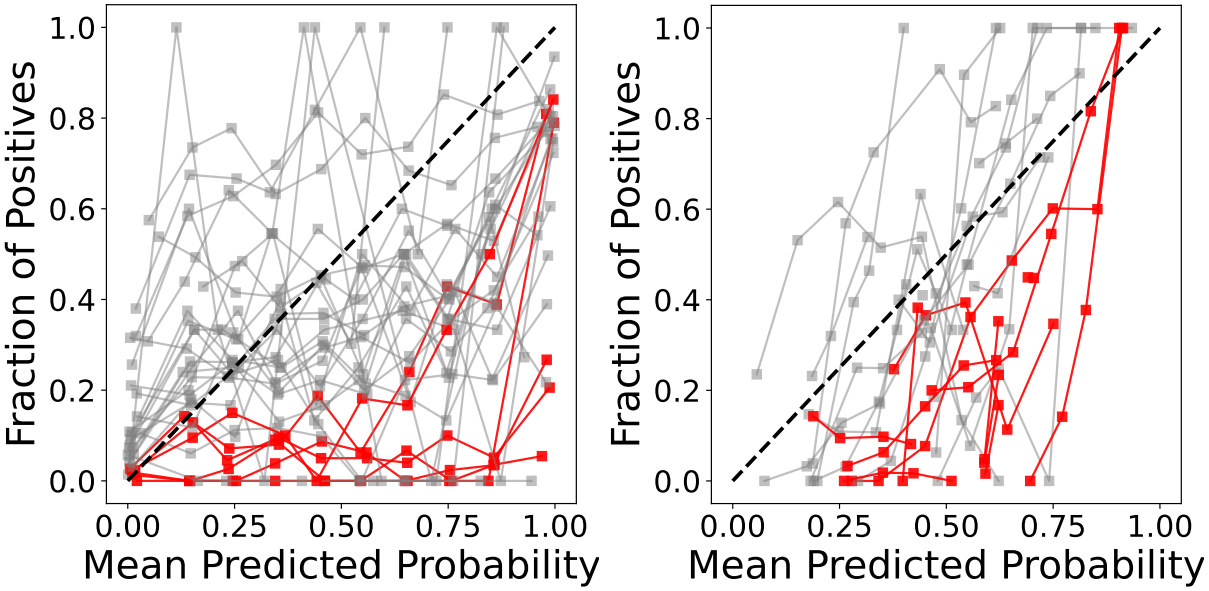
A.7 Additional Agreement Plots



(a) (b)

Figure 17: (a) Plot of AUC scores for each of the datasets, for both GPT-4o and Mistral-7b-Instruct-v0.1. Best-fit line with R2 value plotted in red. (b,c) Correlation between AUC scores of GPT-4o (b) and Mistral-7b-Instruct-v0.1 (c) over prediction tasks on each dataset, along with the AUCs achieved by training an XGBoost classification model on a subset of the training set, and evaluating on a disjoint validation set.

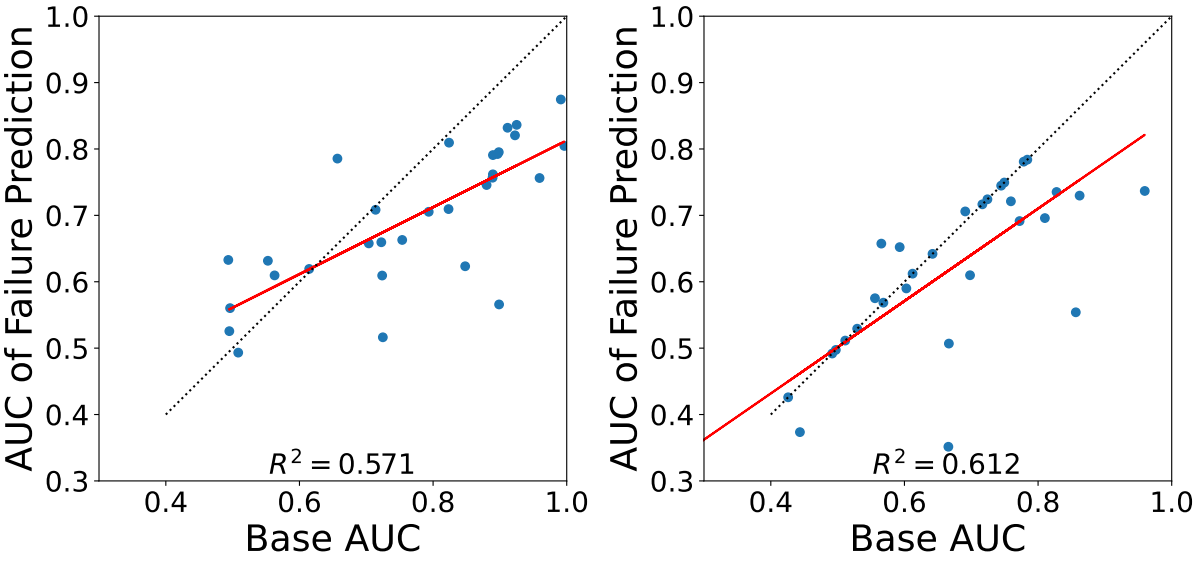
A.8 Additional Calibration Curve Plots



(a) (b)

Figure 18: Calibration curves for GPT-4o (a) and Mistral-7b-Instruct-v0.1 (b) across 31 datasets. Each curve corresponds to a prediction task. Curves crossing the identity line are shown in grey; those consistently above or below are blue and red, respectively. Concretely, all curves that a) are on average 0.2 above the identity line and b) have no points more than .1 below the identity line are colored in blue; curves on average .2 below the identity line and with no points more than .1 above are colored in red.

A.9 Additional Failure Analysis Plots



(a) (b)

Figure 19: Correlation between AUC scores of failure prediction and predicting the outcome variable for all datasets, for GPT-4o (a) and Mistral-7B-Instruct-v0.1 (b).

A.10 Additional Regression on AUC

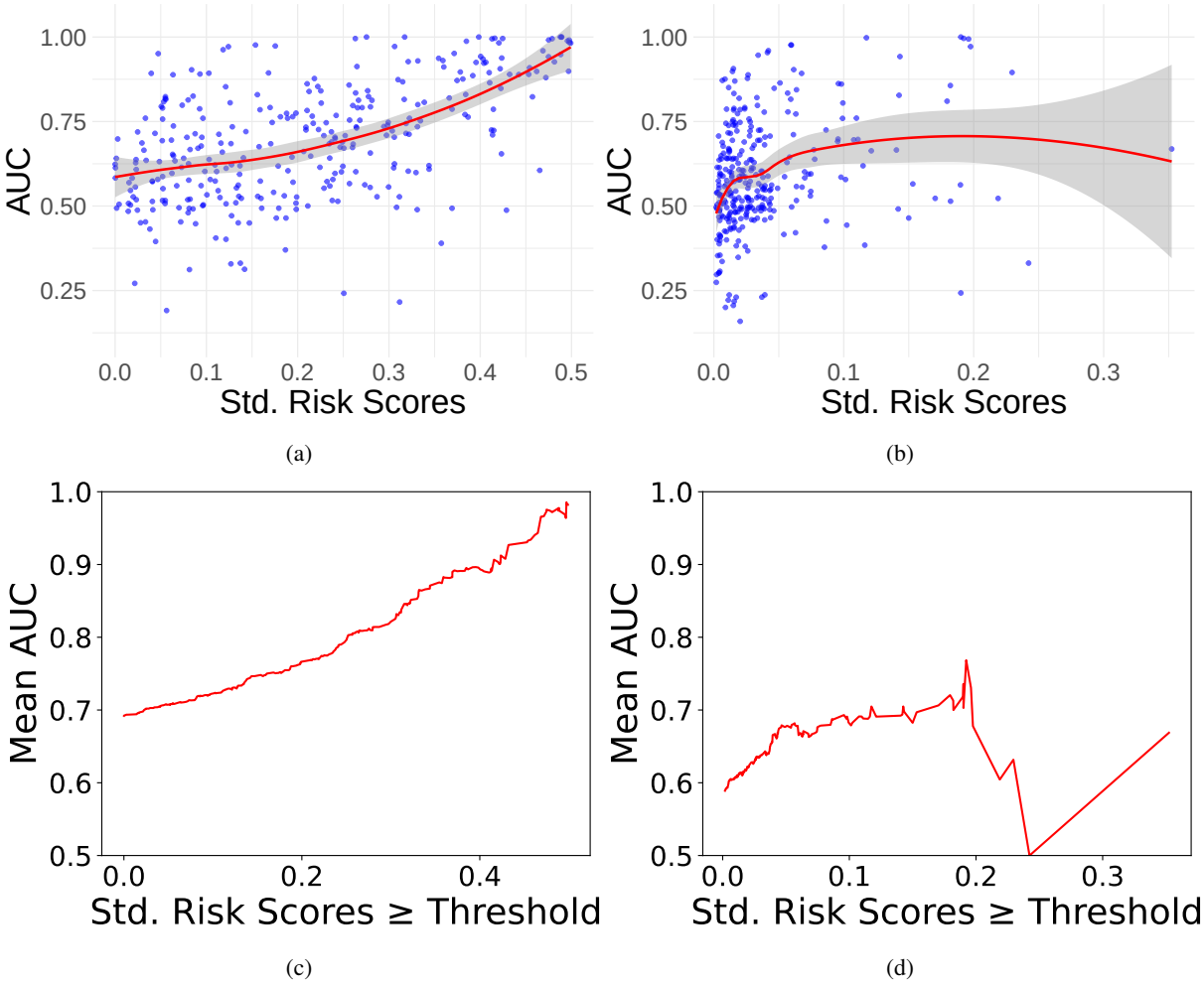


Figure 20: Proxy tasks in the Masking experiment using GPT-4o (a,c) and Mistral-7b-Instruct-v0.1 (b,d), including the original 31 tasks. LOESS curves with 95% CI shown in (a,b); each point represents predictions on one dataset column. (c,d) show average AUC as the minimum threshold on standard deviation of risk scores increases.

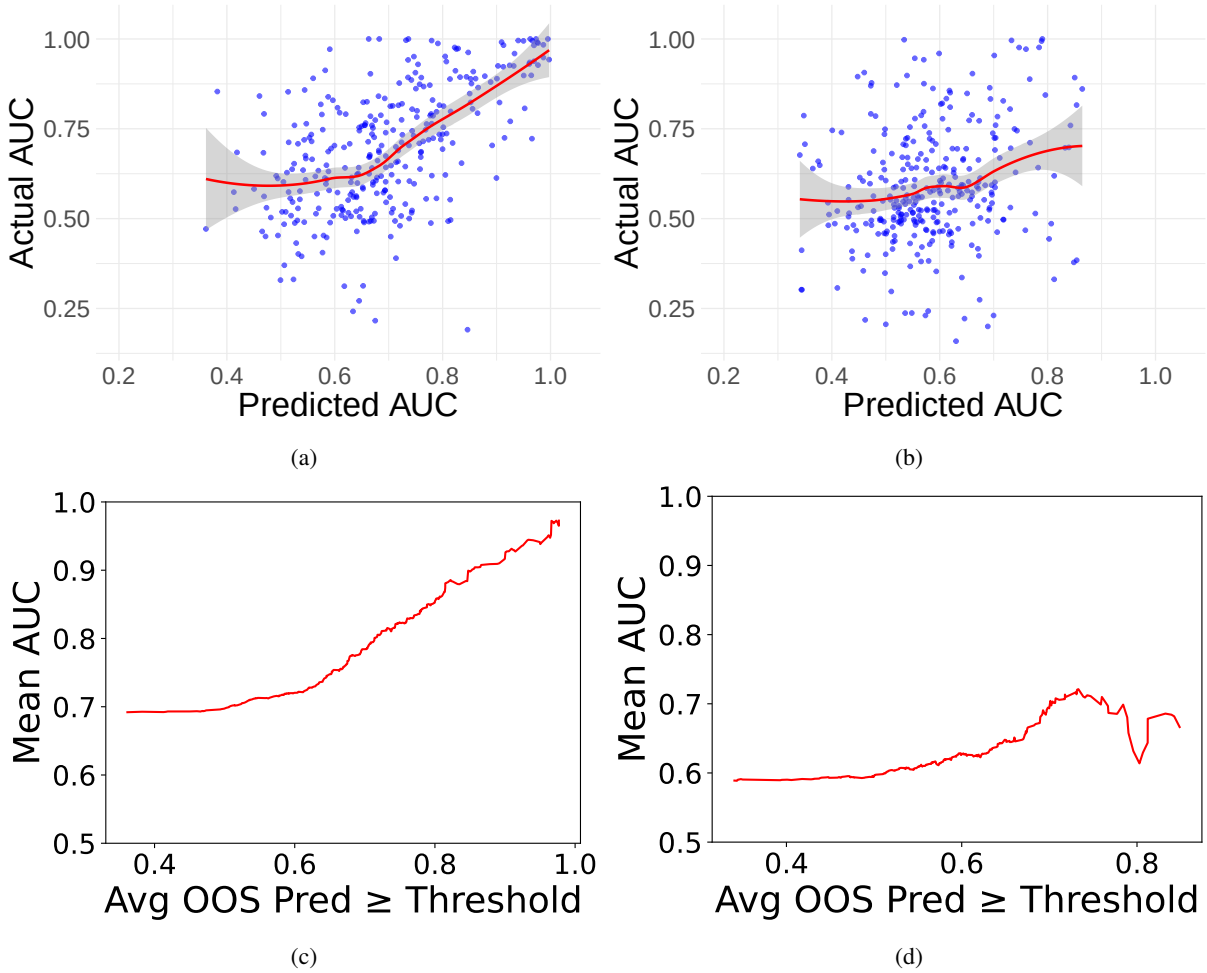
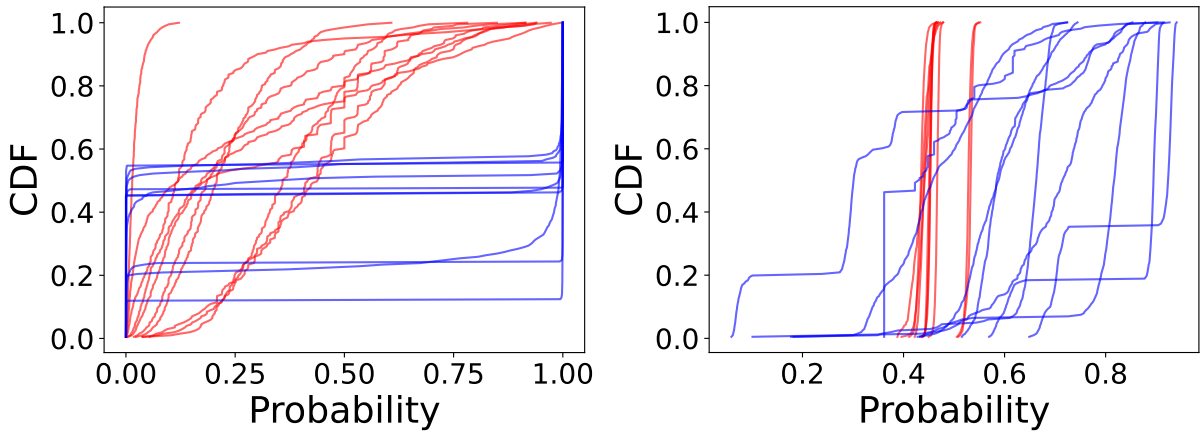


Figure 21: Proxy tasks in the Masking experiment using GPT-4o (a,c) and Mistral-7b-Instruct-v0.1 (b,d), including the original 31 tasks. (a,b) show LOESS fits (with 95% CI) of actual vs. XGBoost-predicted AUCs, trained via grouped 5-fold cross-validation. Each point represents one prediction task. (c,d) show AUC averages after thresholding on predicted AUCs, analogous to Figures 9c and 9d.

A.11 Additional CDF Plots



(a) (b)

Figure 22: CDFs of the 10 highest (blue) and lowest (red) predicted AUCs over prediction tasks by XGBoost, using 201 percentile values along with standard deviation of risk scores to predict AUC. Shown for GPT-4o (a) and Mistral-7b-Instruct-v0.1 (b).

A.12 Prompting Templates

We provide the templates used to generate each of our dataset-level metrics below.

	Context	Content
Risk Scores	"Please respond with a single letter."	$DESCRIPTION OF DATASET$\n\n Information: $SERIALIZED ROW$\n\n Question: $QUESTION$\n A. $POSITIVE LABEL TEXT$\n B. $NEGATIVE LABEL TEXT$
Verb. Confidence	$DESCRIPTION OF DATASET$	$SERIALIZED ROW$ Provide your best guess and the probability that it is correct (0.0 to 1.0) for\n the following question. Give ONLY the guess and probability, no other words or\n explanation. For example:\n\n Guess: <most likely guess, as short as possible; not\n a complete sentence, just the guess!>\n Probability: <the probability between 0.0\n and 1.0 that your guess is correct, without any extra commentary whatsoever; just\n the probability!>\n \n The question is: $QUESTION$

A.13 AI Assistants In Research Or Writing

As our paper centers around the zero-shot capabilities of LLMs for tabular data, all of our experiments necessarily deal with AI assistants (GPT-4o-Mini, Llama-3.1-8b-Instruct) to generate core research results. We also utilize AI assistants (Copilot, GPT) for assistance with rewording and clarity during the paper writing process, along with providing starter code for generating plots.

A.14 Risks

One risk with our findings is the potential misuse of our proposed metrics. While we identify metrics, such as the standard deviation of risk scores, that correlate with LLM performance, these signals should not be interpreted as guarantees of success. Practitioners may be tempted to rely upon our metrics as substitutes for evaluation on labeled data, leading to over-confidence in model outputs. This is particularly of concern in high-stakes domains (e.g., healthcare or finance), where systematically inaccurate predictions carry serious consequences. We emphasize that our metrics are diagnostic tools or guides to which tasks are more promising as opposed to actionable decision rules. They should be used in conjunction with domain knowledge and do not substitute for eventual labeled-data evaluation in high-stakes settings.

A.15 Hardware Details.

For GPT-4o-mini and GPT-4o, we conduct all inference via the OpenAI API, and so we do not require any GPU assistance. However, we run Llama-3.1-8b-Instruct and Mistral-7b-Instruct-v0.1 locally with Huggingface. To do this, we utilize a single NVIDIA Tesla V100 GPU, and require 80 GPU hours to run all experiments.