SYNERGY LEARNING WITH SMALL MODELS PROMOTES LLM ZERO-SHOT TABULAR PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent development in large language models (LLMs) has demonstrated impressive zero-shot proficiency on unstructured textual or multi-modal tasks across various domains. However, despite with inherent world knowledge, their application on structured tabular data prediction still lags behind, primarily due to the numerical insensitivity and modality discrepancy that brings a gap between LLM reasoning and statistical machine learning. Unlike textual or vision data (e.g., electronic health records, medical images), tabular data is often presented in heterogeneous numerical values (e.g., blood test reports). This ubiquitous data format requires intensive expert annotation, and its numerical nature limits LLMs' ability to effectively transfer untapped domain expertise. In this paper, we propose `SERSAL`, a general loop of thought prompting method by synergy learning with small models to **unconditionally enhance zero-shot tabular prediction for LLMs**. Specifically, `SERSAL` utilizes the LLM's zero-shot outcomes as original soft noisy annotations, which are dynamically leveraged to teach a better small student model in a semi-supervised manner. Reversely, the outcomes from the trained small model are used to teach the LLM to further refine its real capability. Such mutual process can be repeatedly applied for continuous progress. Comprehensive experiments on widely used domain tabular datasets show that, without access to gold labels, applying `SERSAL` to OpenAI GPT reasoning process attains substantial improvement compared to linguistic prompting methods, which serves as an orthogonal direction for tabular LLM, and increasing prompting bonus is observed as more powerful LLMs appear.

1 INTRODUCTION

The advancement of large language models (LLMs) (Zhao et al., 2023) has made waves in both research and industry communities. Through friendly natural language interaction and powerful in-context reasoning ability, LLMs have shown their remarkable zero-shot generalization to various language processing (Wei et al., 2021; Wang et al., 2022), complex planning (Qin et al., 2023; Zan et al., 2023) and even vertical domain (e.g., healthcare (Cascella et al., 2023), law (Deroy et al., 2023), chemistry (Guo et al., 2023)) tasks compared to previous supervised small pre-trained language models (Kenton & Toutanova, 2019; Radford et al., 2019), all achieved with suitable prompting and no fine-tuning, yet they are still struggling for the numeric tabular prediction.

For example, GPT-4 achieves 81.7 % accuracy with zero-shot prompting on the United States Medical Licensing Examination (USMLE), which metric will be increased to 90.2 % when meticulous prompts are designed (Nori et al., 2023). In the left part of Fig. 1(a), we also exhibit performances of GPT-3.5, GPT-4 and the fully supervised BERT on top-5 ICD coding for MIMIC-III discharge summaries. Even with simple zero-shot prompting, GPT-3.5 has already surpassed the fine-tuned ClinicalBERT (Huang et al., 2019) and can obtain further improvement with linguistic prompting tricks (e.g., zero-shot CoT (Kojima et al., 2022)). However, when handling medical tables with numerical feature values, the trend becomes totally different. See right part of Fig. 1(a), such significant prompting bonus disappears, suggesting an undeniable void in the current LLM prompting taxonomy tailored for tabular prediction. There are two key points causing the gap:

(i) Existing competitions for general-purpose LLMs predominantly focus on their capabilities in processing unstructured data (Zhang et al., 2024a;b), which is naturally different from structured tabular data characterized by dense heterogeneous numerical features (Borisov et al., 2022; Yan et al., 2023), and the prevailing technical landscape of LLMs neglects nuanced sensitivity and understanding for numerical values (Qian et al., 2023; Yan et al., 2024).

(ii) Most LLM tasks of interest can be formulated as sample-level data understanding then reasoning by generation, the input semantics are unstructured and detailed, while the tabular prediction (e.g., disease diagnosis with numerical metrics from medical examinations and tests) requires overall statistical relation between numerical features and targets on the whole dataset or a specific task, which is hard to access through a single tabular instance in high-level and limited contexts.

Based on these observations, a straightforward question is, how to harness diverse knowledge of existing versatile LLMs, especially these commercial and blackbox (users cannot access the logit) ones (OpenAI, 2022; 2023), for the domain tabular prediction like disease diagnosis using medical examination results, which serves as a potential breakthrough to enable LLMs to handle statistical learning tasks.

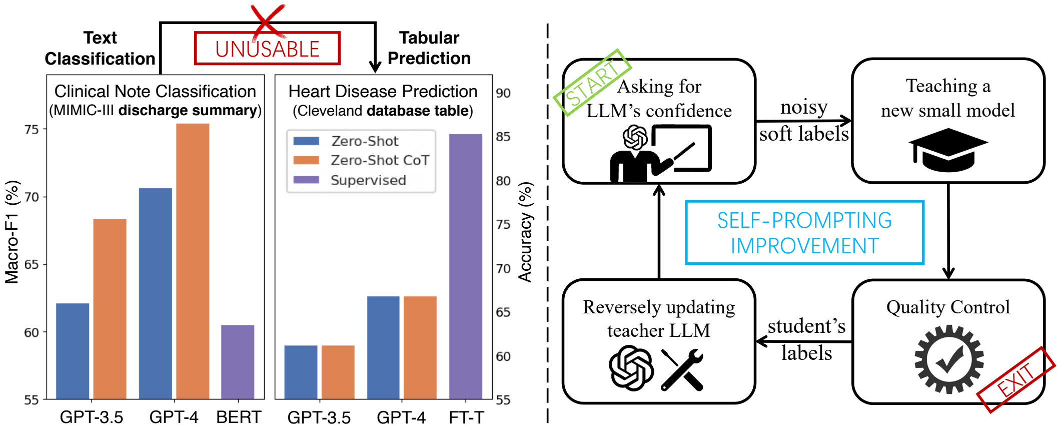
To fill the aforementioned technical gap and extend LLM's vertical capability to tabular prediction, we propose `SERSAL`, a **s**yn**er**gy learning pipeline between **s**mall models **a**nd **L**LMs, requiring no gold labels. Different from existing prompting techniques designing hard or soft prompts to augment inputs for unstructured data tasks, our `SERSAL` contributes from a brand new perspective that **improves zero-shot awareness of LLMs on numeric tabular data prediction**, providing an interface to adapt LLM untapped domain knowledge to such statistical learning featured tasks. `SERSAL` helps a blackbox LLM recognize and refine its statistical capability on vertical domain tabular data in the following steps: (1) Using simple zero-shot prompting to gather the LLM's zero-shot outcomes with confidence as initial coarse annotations on the whole dataset; (2) Teaching a better small tabular model (e.g., XGBoost (Chen & Guestrin, 2016), FT-Transformer (Gorishniy et al., 2021)) from scratch based on the soft confidence like semi-supervised learning with noisy labels; (3) Reversely fine-tuning the LLM using the outcomes of the trained small model to further update LLM annotations in the next loop; The process from step (1) to (3) can be repeatedly applied to the LLM for continuous progress on specific tabular dataset. Essentially, since tabular prediction relies on data statistical information, we use LLM predictions on the all samples as "indicators" and the small model as "probe" to represent and refine the well-expressed part to feedback for LLM self-improvement.

In this paper, the main experiment is based on the well-known online blackbox LLM OpenAI GPT-3.5 (OpenAI, 2022) & GPT-4 (OpenAI, 2023), and as a prompting counterpart, our `SERSAL` can be directly transferred to other general LLMs once the fine-tuning APIs are supported. In a nutshell, our main contributions are:

- For the first time, we bring the common challenge of existing general-purpose LLMs on numeric tabular prediction, a statistical learning featured task, to the spotlight that has not been covered by prevailing prompting techniques.
- We propose `SERSAL`, a novel prompting counterpart for LLM zero-shot improvement on tabular prediction, which leverages LLM outcomes and small models to refine statistical capability from untapped domain knowledge of the LLMs.
- Comprehensive experiments reveal that `SERSAL` is consistently more effective than common prompting methods on many medical tabular datasets, with general feasibility in other vertical domains discussed.

2 SERSAL: Loop of Thought for LLM Tabular Prediction

We propose `SERSAL`, a synergy learning pipeline using small models to adapt LLM's knowledge to tabular data, which is a fundamentally distinct prompting method and serves as a novel interface to extend current LLMs to tabular data prediction. Principally, `SERSAL` is inspired by the theory of semi-supervised noisy label learning and teacher-student model, while several key differences exist: (1) Noisy label learning setting requires a certain proportion of gold labels as the starting point, while `SERSAL` only access the LLM's **soft pseudo labels** to aware its statistical confidence on the whole dataset; (2) In teacher-student paradigm the student model is primarily considered to be comparable to the teacher, while `SERSAL` conservatively **teaches a better student model** by



(a) Linguistic prompting gap **(b) SERSAL loop prompting**

Figure 1: (a) Comparison of prompting effectiveness on unstructured textual data (Mullenbach et al., 2018) and structured tabular data (Detrano et al., 1989) from medical domain, it is clearly seen, even with surprising medical expertise (Nori et al., 2023), GPT-4 still struggles to catch up fully supervised small models (ClinicalBERT (Huang et al., 2019) for textual tasks and FT-Transformer (Gorishniy et al., 2021) for tabular ones) on tabular data, implying essential task discrepancy that makes it incompatible to rely on typical prompting techniques to unlock LLM knowledge on tabular prediction. (b) SERSAL pipeline to refine LLM knowledge for better zero-shot tabular prediction.

dynamically learning from the relatively clean pseudo labels and regularizing on noisy ones from the LLM to avoid misleading confirmation bias (Tarvainen & Valpola, 2017). The overall framework of `SERSAL` is outlined in Fig. 1(b) and formulated in Algorithm 1. Each part is detailed in the following subsections.

2.1 Soft LLM Pseudo Labeling

To access the statistical capability of the LLM on a specific tabular dataset, we first query its confidence on each sample using simple zero-shot prompt template. Specifically, the prompt consists of a task description and listed feature specifications, for example, "You are a professional doctor, here are some clinical metrics of a patient, please give a likelihood between 0 to 1 of suffering from a heart disease: [Age] 47 (years old); [Gender] Male; [Systolic Blood Pressure] 138 (mmHg); [Blood Lipid] 240 (mg/dL); . . . ". In this way the confidence of the LLM on the whole dataset is gathered, though the initial zero-shot performance is often far away from the one of a supervised small tabular model (see Fig. 1(a) and Table 2), we can dig into such fine-grained LLM confidence to separately leverage underlying clean and unclean supervision signal to teach a robust small model.

2.2 Small Model Teaching with Noisy Soft Labels

This step aims to teach a better small model with the collected soft outputs from the LLM. Intuitively, such LLM confidence is a kind of noisy labels, thereby a straightforward insight is to reformulate the teaching process as learning from noisy labels. To seamlessly and thoroughly utilize LLM's knowledge, we adopt a semi-supervised learning manner after dividing training samples into a more reliable labeled set and another unlabeled set, i.e., the small student model is fitted with the soft pseudo labels from the LLM in the labeled set and regularized on the ones of the unlabeled set, the data partition is based on per-sample loss since deep neural networks tend to fit samples with clean labels faster than one with wrong labels according to the noisy label learning (NLL) theory (Arpit et al., 2017) thus lower loss often indicates relatively cleaner labels (Chen et al., 2019).

In implementation, we use an adapted version of DivideMix (Li et al., 2019), a common semi-supervised NLL algorithm for image classification that dynamically fits a Gaussian Mixture Model on sample-wise loss to distinguish probably clean and noisy samples and trains a pair of neural networks simultaneously to keep them diverged to avoid confirmation bias in single-model self-training process (Tarvainen & Valpola, 2017). Apart from extending DivideMix to tabular data, the used soft noisy labels naturally apply label smoothing guided by the LLM. Besides, we leverage the pseudo labels in high LLM confidence for early stopping with underlying assumption that LLM prediction with higher confidence is tend to be more accurate, which is observed in Fig. 2 and discussed in Sec. 3.3. Specifically, we divide a training subset called early stopping set $D_{\text{es}} = \{(\mathbf{X}_i, \bar{\mathbf{y}}_i) | \max(\hat{\mathbf{y}}_i) \geq \tau\}$ to perform early stopping and hyper-parameter selection for the teaching process, where for the i-th sample, $\hat{\mathbf{y}}_i$ is its LLM's soft label confidence vector, and $\bar{\mathbf{y}}_i$ is the corresponding hard labels (i.e., $\bar{\mathbf{y}}_i = \text{argmax}(\hat{\mathbf{y}}_i)$), samples with maximum label confidence larger than threshold τ (we uniformly set $\tau = 0.9$ in the experiment) are considered to be accurate enough for early stopping. During the semi-supervised learning, samples in the early stopping set are also used for training since: (1) some domain (e.g., medicine) tabular datasets may suffer from data inadequacy, making the reduction of training subsets likely to distort data distribution; (2) This subset still contains noisy labels (not totally accurate) thus the small model will not overfit on the real training labels. We formulate this step in the line 3-5 of Algorithm 1 and conduct related ablations in Sec. 3.3.

In summary, this teaching step adapts semi-supervised noisy label learning process to LLM knowledge refinement and aggregation to amplify the LLM's real capabilities on tabular data through a small student model.

2.3 Quality Control

Since SERSAL operates iteratively, it requires a termination mechanism to manage the process exit. Here we provide three heuristic strategies, users can also define their own control flow in practice.

- **Metric-based Control.** In Sec. 2.2 we define the high-confidence training subset as early stopping set D_{es} which pseudo labels are relatively more accurate (see Fig. 2). Therefore, users can inspect metrics (e.g., AUC or accuracy scores for classification tasks) by treating these pseudo labels as the "ground truth" to control whether to end the loop.
- **External Validation Control.** If budget permits, human experts can collect and annotate appropriate external data as a validation set, e.g., in hospitals, regular medical data quality inspection needs to sample and label a small part of data, and learning quality can be assessed with such external set.
- **Rule-based Control.** For example, users can define a fixed iteration time.

For simplicity, in the main experiment we uniformly report one-loop SERSAL performance in medical and other domain tabular datasets (Table 2 & 5), which significantly surpassed the prevailing prompting techniques, and further discuss the effectiveness of multi-loop SERSAL in Sec. 3.4.

2.4 Reverse LLM Tuning

The final step is to reversely teach the LLM using the well-trained small student model to feedback the aggregated knowledge. Similar to using LLM soft confidence to teach the small model in Sec. 2.2, we also use soft confidence from the small model to fine-tune the LLM (fine-tune the online blackbox GPTs through their APIs in experiment). Specifically, the training samples are re-labeled by the trained small model with its guessed probabilities (line 7-8 in Algorithm 1), the same prompt templates in Sec. 2.1 are used to construct the training corpus for the LLM. To avoid the excessive memorization of the LLM on the small model outputs (Bordt et al., 2023), we employ a slightly tuning strategy that sets the maximum training epoch to 3 with proper early stopping (the fine-tuning APIs of GPT-3.5 & GPT-4 provide automatic early stopping in default), making the LLM slightly fitted on the guessed labels while keeping a non-zero minimum training loss. Then the updated LLM initiates the next SERSAL loop, forming an iterative process.

3 Experiments

In this section, we first compare SERSAL with prevailing prompting techniques (using GPT-3.5 & GPT-4) and the fully supervised small models on extensive medical domain tabular datasets (Sec. 3.2). Next, we conduct ablation on several key adaptations in semi-supervised noisy label

Algorithm 1 SERSAL Pipeline. Line 2: Soft pseudo labeling (Sec. 2.1); Line 3-5: Small model teaching (Sec. 2.2); Line 6: Quality control (Sec. 2.3); Line 7-9: Reverse LLM tuning (Sec. 2.4).

Input: Unlabeled training set $\mathbf{X}_{\text{train}}$ and test set $\mathbf{X}_{\text{test}}$, large language model $f_{\text{LLM}}^{(0)}$
Parameter: Confidence threshold τ, quality control function f_{ctr}
Output: Improved zero-shot tabular prediction $\mathbf{y}_{\text{test}}^{*}$

1: Let $t = 1$. // Initialize iteration number
2: Softly labeled dataset $D_{\text{train}}^{(t)} = (\mathbf{X}_{\text{train}}, \hat{\mathbf{y}}^{(t)})$ by current $f_{\text{LLM}}^{(t)}$.
3: Randomly initialize a small tabular model $\theta^{(t)}$.
4: Select early stopping set $D_{\text{es}}^{(t)} = \left\{(\mathbf{X}_i, \bar{\mathbf{y}}_i^{(t)}) | \max(\hat{\mathbf{y}}_i^{(t)}) \geq \tau\right\} \subseteq D_{\text{train}}^{(t)}$.
5: $\theta^{*(t)}$ = DivideMix$(D_{\text{train}}^{(t)}, D_{\text{es}}^{(t)}, \theta^{(t)}, \tau)$. // Adapted DivideMix (Li et al., 2019)
6: **while** $f_{\text{ctr}}(\theta^{*(t)}, \mathbf{X})$ **do**
7: $\quad \mathbf{y}_{\text{sm}}^{(t)} = \text{Predict}(\mathbf{X}_{\text{train}}; \theta^{*(t)})$. // Soft label guessing by the small model
8: $\quad \hat{\mathbf{y}}_{\text{sm}}^{(t)} = \text{Sharpen}(\mathbf{y}_{\text{sm}}^{(t)}, \text{temperature} = 0.1)$. // Simple temperature sharpening
9: $\quad f_{\text{LLM}}^{(t+1)} = \text{Finetune}(\mathbf{X}_{\text{train}}, \hat{\mathbf{y}}_{\text{sm}}^{(t)}, f_{\text{LLM}}^{(t)})$. // Reversely tune the LLM with guessed labels
10: $\quad t = t + 1$.
11: $\quad$ Repeat Line 2-5. // Self-prompting loop
12: **end while**
13: $\mathbf{y}_{\text{test}}^{*} = \text{Predict}(\mathbf{X}_{\text{test}}; \theta^{*(t)})$. // Final prediction with the taught small model
14: **return** $\mathbf{y}_{\text{test}}^{*}$

learning (Sec. 2.2) and inspect the effectiveness of multi-loop SERSAL (Sec. 3.4). Also, we explore the method interpretability by visualizing Shapely Value variation during SERSAL process (Sec. 3.5). Besides, we discuss the method feasibility on other domain tabular data (Sec. 3.6).

3.1 Experimental Setup

Datasets. We evaluate on ten widely recognized medical diagnosis tabular datasets on various diseases: Heart Failure Prediction (HF, Detrano et al. (1989)), Lung Cancer Prediction (LC, Ahmad & Mayya (2020)), Early Classification of Diabetes (ECD, Islam et al. (2020)), Indian Liver Patient Records (LI, Ramana et al. (2012)), Hepatitis C Prediction (HE, Hoffmann et al. (2018)), Pima Indians Diabetes Database (PID, Smith et al. (1988)), Framingham Heart Study (FH, O'Donnell & Elosua (2008)), Stroke Prediction (ST, Fedesoriano (2020)), COVID-19 Presence(CO, Hemanthhari (2020)) and Anemia Disease (AN, Kilicarslan et al. (2021)). Besides, datasets in clinical trail (Wang & Sun, 2022) and open domains (Gorishniy et al., 2021) are added to further inspect the effectiveness of SERSAL in difficult tasks and general data scenarios respectively. We split each tabular dataset (80 % for training and 20 % for testing), and keep the same label distribution in each split. Statistics of medical diagnosis datasets are given in Table 1.

Dataset	HF	LC	ECD	LI	HE	PID	FH	ST	CO	AN
# features	13	15	16	10	12	8	15	7	20	24
# samples	303	309	520	583	615	768	4238	5110	5434	15300
P/N	0.80	6.92	1.60	2.51	0.11	0.54	0.18	0.04	4.17	0.57
disease	Heart	Lung	Diabetes	Liver	Hepatitis C	Diabetes	Heart	Stroke	COVID-19	Anemia

Table 1: Dataset statistics of ten medical diagnosis datasets for binary classification on various diseases. "P/N" denotes the amount ratio of positive samples and negative ones.

Compared Methods. Since SERSAL serves as a tailored loop of thought prompting method for LLM zero-shot tabular prediction, we compare with existing linguistic prompting methods for LLM usage in general textual and tabular tasks, which focus on meticulously designed prompt texts: (1) **Zero-Shot Prompting** (0-shot) is the straightforward prompt that contains no examples; (2) **Zero-Shot CoT Prompting** (Kojima et al., 2022) (CoT) is a popular prompting method which asks the LLMs to answer with intermediate reasoning steps to enable complex reasoning capabilities; (3) **8-shot Prompting** (8-shot) is a common few-shot prompt setting in standard prompting studies (Wei et al., 2022; Kojima et al., 2022; Nori et al., 2023), it provides eight labeled samples (exemplars) to enrich prompt contexts and steer the LLM to the better outputs, in the experiment we randomly sample eight training examples and control the same positive-negative ratio (i.e., "P/N" in Table 1) with at least one example for each class; (4) **TabLLM** (Hegselmann et al., 2023) and (5) **LIFT** (Dinh et al., 2022) are two recent known linguistic prompt schemes for textualizing tabular data to fine-tune LLMs with gold labels, though TabLLM was additionally evaluated in zero-shot settings, **none of them are originally proposed for zero-shot tabular scenarios**, here we use their schemes for zero-shot comparison. Additionally, we provide a **fully supervised small tabular model** (FSSM) group for reference representing traditional supervised tabular prediction by fine-tuning dataset-specific small models.

Implementation Details. All experiments are conducted with PyTorch on Python 3.8 and run on NVIDIA RTX 3090. For the small model, we uniformly use FT-Transformer with the default model and training configurations provided in the original paper (Gorishniy et al., 2021). For `SERSAL`, the only adjustable hyper-parameter is the temperature of DivideMix (Li et al., 2019) with choices of 0.5, 5.0 and 10.0 in line 5 of Algorithm 1, which is selected by the metric of the early stopping set ($D_{\text{es}}^{(t)}$ in line 4 of Algorithm 1). The LLMs in the experiment includes OpenAI GPT-3.5 & GPT-4 to inspect the effectiveness of `SERSAL` across different LLM capabilities.

3.2 WHY WE NEED SERSAL?

	HF	LC	ECD	LI	HE	PID	FH	ST	CO	AN
Random guessing	37.22	40.18	46.25	50.28	62.73	63.24	50.39	41.76	71.55	51.28
FSSM*(supervised FT-T)	88.19	86.61	99.60	78.94	100.00	84.72	66.25	82.98	99.91	99.92
0-shot (GPT-3.5)	71.88	78.87	85.71	76.81	68.51	73.12	60.32	63.01	82.60	90.43
8-shot* (GPT-3.5)	73.65	78.87	**87.68**	76.81	68.51	73.12	58.27	60.85	77.63	87.19
CoT (GPT-3.5)	71.88	78.87	82.36	76.81	68.51	70.83	60.32	63.01	82.60	90.43
TabLLM (GPT-3.5)	76.37	78.87	87.06	78.24	74.39	75.69	61.78	68.48	85.78	89.11
LIFT (GPT-3.5)	78.23	80.69	83.92	73.60	72.57	73.12	60.32	70.92	87.93	90.43
`SERSAL` (GPT-3.5)	**91.39**	**85.42**	86.40	**79.39**	**85.14**	**78.97**	**63.97**	**76.36**	**96.85**	**98.37**
TabLLM+`SERSAL` (GPT-3.5)	93.82	85.42	88.39	80.71	89.27	82.54	65.02	81.74	97.51	98.16
`SERSAL` (GPT-4)	94.18	86.93	92.68	82.51	92.76	82.39	67.14	81.23	97.96	98.82

Table 2: The AUC scores (%) of different tabular prediction schemes on 10 medical diagnosis datasets. The top part is the traditional supervised small models, the middle one is compared LLM prompting methods (the top performances are marked in **bold**), the bottom part is additional combinations. Here the results of `SERSAL` are only based on a single loop. "$*$" denotes the groups use gold labels. "FSSM" is the fully supervised FT-Transformer. The results on more difficult clinical trial datasets are given in Table 7.

Main Results Analysis. The performances of different LLM prompting baselines are reported in the middle part of Table 2. An overall trend is that, when the GPT-3.5 meets medical domain tabular prediction tasks, the results using common prompting methods are consistently better than the ones of random guessing, demonstrating the general-purpose LLMs indeed contain medical domain expertise inherently, but they are still far from the traditional supervised small models (see group "FSSM"), and further performance enhancement can not be achieved through usual prompting tricks as in textual tasks (see Fig. 1(a)). Specifically, we observe 8-shot prompting slightly benefits the performances in small-scale datasets (e.g., HF and ECD) but hurts in the larger datasets (e.g., FH, ST, CO and AN) compared to the 0-shot prompting, which may be explained by the representativeness of the used examples, since the distribution of the smaller datasets are more likely to be covered by few examples, thus 8-shot performs better as data scale decreases, and vice versa. For 0-shot CoT prompting, it does not affect the overall results in most cases, but we find slight performance decline in two diabetes datasets (i.e., ECD and PID), this may be caused by the over-consideration of CoT on noisy features since diabetes can be diagnosed with several prominent features (e.g., blood sugar and lipid). Although carefully crafted prompt templates from recent LLM in-context tabular learning studies (i.e., TabLLM and LIFT) show modest improvement, **they still follow the linguistic nature to process numeric tabular data**, and are primarily designed for few-shot learning or supervised fine-tuning. Our `SERSAL` **explores a fundamentally novel prompting mechanism exploiting the information gain in the LLM's noisy outputs**, which breaks through the predicament from an orthogonal perspective and serves as an interface to effectively adapt the LLM's domain knowledge to numeric tabular data. After applying SERSAL, without access to gold labels, the GPT-3.5 is able to achieve significantly better reasoning on these medical domain tasks, with many cases competitive with the supervised small models.

Orthogonal Technical Contribution. Based on the above analysis, SERSAL works in a distinct underlying mechanism, and we can jointly adopt SERSAL and previous linguistic prompting methods for better combined performances (see group "TabLLM+SERSAL").

Continuous Performance Growth. We additionally apply SERSAL to OpenAI GPT-4 on medical diagnosis datasets (the bottom part of Table 2) and more difficult clinical trial datasets (see Table 7). It can be seen SERSAL can further realize substantial performance gains as the capability of used LLMs becomes more powerful, which can even surpass the traditional supervised paradigm (N00041119 and N00312208 datasets in Table 7), indicating ample room for continuous prompting bonus in SERSAL alongside the emergence of more advanced LLMs.

3.3 Several Key Adaptations

	HF	LC	ECD	LI	HE	PID	FH	ST	CO	AN
SERSAL	91.39	85.42	86.40	79.39	85.14	78.97	63.97	76.36	96.85	98.37
w/o soft pseudo	84.58	76.58	87.24	78.25	75.79	75.93	62.58	75.05	93.97	97.53
w/o ES	84.03	74.11	75.92	59.39	47.41	68.43	57.08	74.70	90.57	97.57

Table 3: The AUC scores of ablation on two key adaptations. "w/o soft pseudo" means replacing the LLM's soft outputs with hard ones during teaching the student model, "w/o ES" denotes no early stopping during DivideMix (line 5 in Algorithm 1).

In Sec. 2.2, to adapt the LLM's outputs to a semi-supervised learning process to teach a small student model, we gather per-sample confidence from the LLM as soft noisy annotations and heuristically select high-confidence samples for early stopping. In this section, we will analyze the effect of the two designs which distinguish our SERSAL from traditional noisy label learning settings.

The Effect of using Soft Labels. We query soft confidence from the LLM (see Sec. 2.1) rather than directly using hard outputs for small model teaching. The prediction probabilities inherently reflect the LLM's prior knowledge on the domain tabular data and can be naturally treated as a kind of label smoothing. Besides, the probability values can be used to select relatively reliable labels to early stop the teaching process and avoid overfitting. In Table 3 we compare the effect of using soft labels by replacing it with hard ones during SERSAL reasoning (group "w/o soft pseudo"). We find that using hard ones is usually suboptimal since it loses both prediction uncertainty and label smoothing, which is unable to exploit fine-grained LLM knowledge.

The Effect of Early Stopping. In addition to using LLM soft outputs, a relatively clean training subset is selected by threshold clipping on the per-sample confidence (line 4 in Algorithm 1) for early stopping. Table 3 report the ablation results by directly training 100 epochs (group "w/o ES"). It can be clearly seen, simply following the original DivideMix is far from the desired results, since tabular features are heterogeneous and high-level compared to the well-patterned pixels of images (Chen et al., 2023; Yan et al., 2024), and in medical tabular domain the typically limited available data further makes it prone to overfit without early stopping, for example, except large AN dataset, all other tabular datasets appear to be significantly impacted by removing the early stopping mechanism. The heuristic design of selecting high-confidence sample is inspired from the empiri-

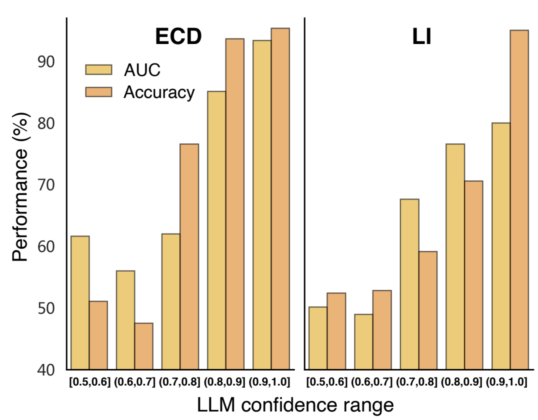


Figure 2: Performances in different LLM confidence ranges on ECD and LI datasets. High-confidence samples are relatively more reliable.

cal assumption that confident predictions from the LLM are more likely to be accurate, which is supported by the performance variation of different confidence ranges in Fig. 2.

3.4 Effectiveness of Multi-loop SERSAL

Since SERSAL is essentially a self-prompting loop (see Fig. 1(b)), we further inspect the effectiveness of multi-loop SERSAL for GPT-3.5 reasoning process. Specifically, we repeatedly apply the pipeline three times on ECD and LI datasets, the result variations are reported in Table 4.

During three loops, progressive improvement on both the small model (SERSAL outputs are from the well-trained student model of each loop) and the GPT-3.5 is observed. Surprisingly, even inferior to the 8-shot prompting baseline on ECD dataset after the first loop (see Table 2), we find SERSAL can reduce the gap and even outperform few-shot baselines after several loops. Such continuous progress probably comes from the synergy learning between the small model and the LLM that **shares a similar underlying mechanism of co-teaching** (Han et al., 2018), i.e., both sides purposely learn from a part of reliable pesudo labels from each other and it makes them diverged to avoid confirmation bias, forming a mutual improvement manner to aggregate and refine untapped domain knowledge for LLM tabular prediction.

# Loop	ECD		LI	
	SERSAL	LLM 0-shot	SERSAL	LLM 0-shot
1	86.40	85.71	79.39	76.81
2	87.00	86.42	82.47	80.26
3	89.00	87.81	84.07	82.91

Table 4: The AUC score variation of SERSAL outputs and zero-shot prompting of the tuned GPT-3.5 (LLM 0-shot) on LI and ECD datasets during three loops. "# Loop" is the same as the variable t in line 1 of of Algorithm 1.

3.5 Interpretability of SERSAL

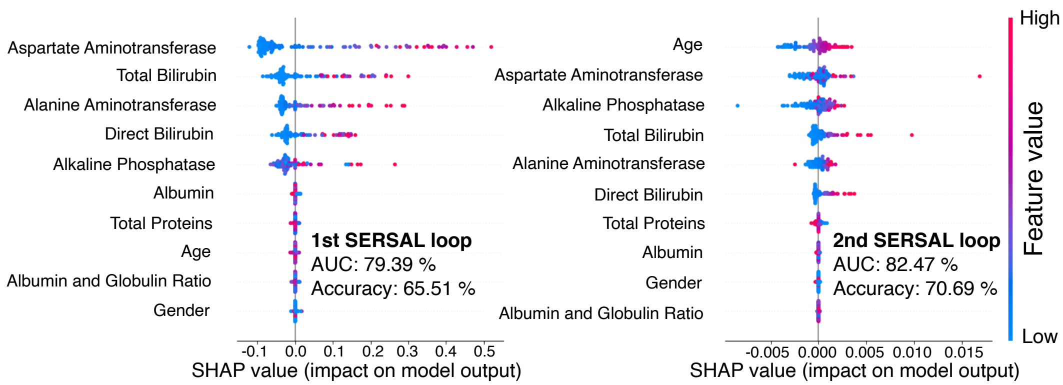


Figure 3: Interpretability visualization from feature importance perspective: the variation of the Shapley Values (treat SERSAL outputs as the targets) and performances on Indian Liver Patient Records (LI dataset) after one and two SERSAL loops using GPT-3.5.

In Fig. 3 we visualize the variation of Shapely Values on Indian Liver Patient Records (LI) dataset after one (left) and two (right) SERSAL loops by treating the predictions (i.e., Algorithm 1) as targets. It can be clearly seen the feature "Age" is adequately considered after one loop self-prompting, which highlights a strong and reasonable positive correlation between age and the incidence of liver diseases that aligns with the medical expertise. Besides, a negative correlation with "Total Proteins", a guiding clinical metric reflecting the liver's synthetic function, is captured in the right figure to contribute the prediction, since a lower total protein level indicates a risk of liver cirrhosis. These two reasonable changes of feature importance interpret the SERSAL prompting is able to iteratively refine the domain expertise in the LLM, calibrating the statistical feature-target relationship for better reasoning results during the process.

3.6 GNERALIZED DATA ADAPTABILITY ON OTHER DOMAINS

In this section, we further explore the feasibility of SERSAL on other domain tabular data. We use three classic binary classification datasets: Churn Modeling (Iyyer, 2019), Credit (Credit Fusion, 2011) and Adult (Kohavi et al., 1996), which are widely included in general tabular prediction studies (Gorishniy et al., 2021; Yan et al., 2023; Grinsztajn et al., 2022). Additionally, we build a dataset "Fake" by randomly generating features and binary labels to emulate the extreme case where the LLM has no related knowledge at all. The data information and the results are given in Table 5. As in the medical domains, the GPT-3.5 indeed holds the world knowledge and can directly achieve the considerable results with simple zero-shot prompting, and SERSAL further enhances the zero-shot performance significantly. However, when facing the tabular data from an unknown domain (i.e., the Fake dataset), the LLM outputs high confidence on wrong labels, SERSAL is unable to recognize such misleading confidence. Therefore, our SERSAL shares the same basic limitation as other linguistic prompting methods that the applied LLMs require a certain level of knowledge in the target domain.

	Churn	Credit	Adult	Fake
domain	Business	Finance	Sociology	N/A
# features	10	10	14	6
# samples	10000	16714	48842	1000
Random guessing	66.35	43.80	58.73	53.85
FSSM*	86.27	84.88	91.39	55.31
0-shot (GPT-3.5)	77.81	69.05	75.10	46.28
SERSAL (GPT-3.5)	83.29	79.36	88.72	38.72

Table 5: The data statistics and AUC scores on other general data domains. "Fake" is a generated dataset with random labels and features. The denotations follow the ones in Table 1 and Table 2.

4 RELATED WORK

Prompt Engineering for In-Context Learning Prompt engineering is a flourishing discipline for better LLM reasoning through meticulously designed linguistic input contexts or interaction process. The most common and straightforward prompting is the single-round instruction that directly asks with zero (zero-shot) or several (few-shot) demonstrations (Brown et al., 2020; Wei et al., 2021), but such prompt style fails to work in more complex reasoning tasks (Wei et al., 2022). To tackle this deficiency and improve the LLM's capacity on a wide range of tasks, recently, studies on more advanced prompting methods are emerging, such as chain-of-thought (CoT) (Wei et al., 2022; Kojima et al., 2022; Zhang et al., 2023a), tree-of-thought (ToT) (Yao et al., 2023) and self-consistency (Wang et al., 2023). However, current prompting methods are mostly designed to serve unstructured data tasks (Zhang et al., 2023b). Although recent studies on LLM in-context learning for tabular data (e.g., TabLLM (Hegselmann et al., 2023), LIFT (Dinh et al., 2022)) propose table-friendly prompting strategies, their linguistic nature still hinders the numeric table understanding (Yan et al., 2024).

Semi-supervised Noisy Label Learning Semi-supervised learning treats the unlabeled samples as regularization for better model generalization (Lee et al., 2013; Tarvainen & Valpola, 2017; Miyato et al., 2019; Berthelot et al., 2019). Recently, the related theory has been introduced to noisy label learning scenarios (Song et al., 2022) that actively divide samples into clean labeled group and noisy unlabeled one (Li et al., 2019) to achieve robust training even with noisy labels.

5 CONCLUSIONS

This paper revealed the common challenge of existing general-purpose LLMs on numeric tabular prediction and proposed SERSAL, a novel loop of thought prompting method in non-linguistic mechanism that unlocks the LLM's domain knowledge for better zero-shot tabular prediction. This is achieved through a co-teaching process between the LLM and a small student tabular model which learn from the other's noisy outputs to refine the untapped capability of the LLM. Extensive experiments on various domain tabular datasets demonstrate that, SERSAL is consistently more suitable to trigger the LLM's latent power on numeric tabular prediction, serving as an orthogonal prompting landscape to extend the LLMs to the domain tabular data.

REFERENCES

Ahmad S Ahmad and Ali M Mayya. A new tool to predict lung cancer based on risk factors. *Heliyon*, 2020.

Devansh Arpit, Stanislaw Jastrzkebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *ICML*, 2017.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *NeurIPS*, 2019.

Sebastian Bordt, Harsha Nori, and Rich Caruana. Elephants never forget: Testing language models for memorization of tabular data. In *NeurIPS 2023 Second Table Representation Learning Workshop*, 2023.

Vadim Borisov, Tobias Leemann, Kathrin Sessler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. *TNNLS*, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *NeurIPS*, 2020.

Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios. *Journal of Medical Systems*, 2023.

Jintai Chen, KuanLun Liao, Yanwen Fang, Danny Z. Chen, and Jian Wu. TabCaps: A capsule neural network for tabular data classification with BoW routing. In *ICLR*, 2023.

Pengfei Chen, Ben Ben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. In *ICML*, 2019.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *KDD*, 2016.

Will Cukierski Credit Fusion. Give me some credit, 2011. URL `https://kaggle.com/competitions/GiveMeSomeCredit`.

Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. How ready are pre-trained abstractive models and llms for legal case judgement summarization? *arXiv preprint arXiv:2306.01248*, 2023.

Robert Detrano, Andras Janosi, et al. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 1989.

Tuan Dinh, Yuchen Zeng, Ruisu Zhang, Ziqian Lin, Michael Gira, Shashank Rajput, Jy-yong Sohn, Dimitris Papailiopoulos, and Kangwook Lee. Lift: Language-interfaced fine-tuning for non-language machine learning tasks. In *NeurIPS*, 2022.

Fedesoriano. Stroke prediction dataset. `https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset`, 2020.

Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Revisiting deep learning models for tabular data. *NeurIPS*, 2021.

Leo Grinsztajn, Edouard Oyallon, and Gael Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *NeurIPS*, 2022.

Taicheng Guo, Kehan Guo, et al. What indeed can gpt models do in chemistry? a comprehensive benchmark on eight tasks. *arXiv preprint arXiv:2305.18365*, 2023.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018.

Stefan Hegselmann, Alejandro Buendia, Hunter Lang, Monica Agrawal, Xiaoyi Jiang, and David Sontag. Tabllm: Few-shot classification of tabular data with large language models. In *AISTATS*, 2023.

Hemanthhari. Symptoms and covid presence (may 2020 data). `https://www.kaggle.com/datasets/hemanthhari/symptoms-and-covid-presence`, 2020.

Georg Hoffmann, Andreas Bietenbeck, Ralf Lichtinghagen, and Frank Klawonn. Using machine learning techniques to generate laboratory diagnostic pathways—a case study. *J Lab Precis Med*, 2018.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

MM Islam, Rahatara Ferdousi, Sadikur Rahman, and Humayra Yasmin Bushra. Likelihood prediction of diabetes at early stage using data mining techniques. In *Computer Vision and Machine Intelligence in Medical Image Analysis*, 2020.

Shruti Iyyer. Churn modelling, 2019. URL `https://www.kaggle.com/shrutimechlearn/churn-modelling`.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 2016.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

Serhat Kilicarslan, Mete Celik, and Şafak Sahin. Hybrid models based on genetic algorithm and deep learning algorithms for nutritional anemia disease classification. *Biomedical Signal Processing and Control*, 2021.

Ron Kohavi et al. Scaling up the accuracy of Naive-Bayes classifiers: A decision-tree hybrid. In *KDD*, 1996.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. 2022.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 2013.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2019.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *TPAMI*, 2019.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *NAACL*, 2018.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

Christopher J O'Donnell and Roberto Elosua. Cardiovascular risk factors. insights from framingham heart study. *Revista Española de Cardiología (English Edition)*, 2008.

OpenAI. Openai: Introducing chatgpt. `https://openai.com/blog/chatgpt`, 2022.

OpenAI. Gpt-4 technical report. `https://cdn.openai.com/papers/gpt-4.pdf`, 2023.

Jing Qian, Hong Wang, Zekun Li, Shiyang Li, and Xifeng Yan. Limitations of language models in arithmetic and symbolic induction. In *ACL*, 2023.

Yujia Qin, Shihao Liang, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*, 2023.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Bendi Venkata Ramana, M Surendra Prasad Babu, and NB Venkateswarlu. A critical comparative study of liver patients from usa and india: an exploratory analysis. *IJCSI*, 2012.

Jack W Smith, James E Everhart, et al. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, 1988.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*, 2017.

Thomas Wang, Adam Roberts, Daniel Hesslow, Teven Le Scao, Hyung Won Chung, Iz Beltagy, Julien Launay, and Colin Raffel. What language model architecture and pretraining objective works best for zero-shot generalization? In *ICML*, 2022.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

Zifeng Wang and Jimeng Sun. Transtab: Learning transferable tabular transformers across tables. In *NeurIPS*, 2022.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *ICLR*, 2021.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022.

Jiahuan Yan, Jintai Chen, Yixuan Wu, Danny Z Chen, and Jian Wu. T2G-Former: Organizing tabular features into relation graphs promotes heterogeneous feature interaction. In *AAAI*, 2023.

Jiahuan Yan, Bo Zheng, Hongxia Xu, Yiheng Zhu, Danny Chen, Jimeng Sun, Jian Wu, and Jintai Chen. Making pre-trained language models great on tabular prediction. In *ICLR*, 2024.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *NeurIPS*, 2023.

Daoguang Zan, Bei Chen, Fengji Zhang, Dianjie Lu, Bingchao Wu, Bei Guan, Wang Yongji, and Jian-Guang Lou. Large language models meet nl2code: A survey. In *ACL*, 2023.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*, 2024a.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024b.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *ICLR*, 2023a.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023b.

Wayne Xin Zhao, Kun Zhou, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

A LIMITATIONS & IMPACTS

As discussed in Sec. 3.6, though our SERSAL is distinguished from traditional prompting methods by its non-linguistic mechanism, it still requires the LLMs with latent knowledge in the target domain to be effective. Therefore, in practice the user should have prior understanding of the used LLM's capability or advantageous application fields. SERSAL contributes to the progress in both LLM prompting and tabular data community through providing a novel interface to adapt untapped knowledge in LLMs to the tabular prediction tasks in a zero-shot manner, which is particularly useful in the regime where limited data or annotation is available.

B DATASETS AND EXPERIMENT DETAILS

We provide detailed data information of the experiment tabular datasets in Table 6. We drop the samples with missing features and adopt the same preprocessing as Gorishniy et al. (2021) before training. For MIMIC-III discharge summary dataset (Johnson et al., 2016; Mullenbach et al., 2018) used in Fig. 1(a), we retain the most frequent 5 labels (medical codes) since our goal is just to demonstrate the prompting effectiveness on medical textual tasks and conducting validation on the full label version (several thousands labels) is inconvenient. During conducting zero-shot prompting for GPT-3.5v and GPT-4v on the MIMIC-III dataset, we follow the PhysioNet Credentialed Data Use Agreement [1] and enroll in the Azure OpenAI service without human review of the data to protect the data from third-party access.

Dataset	Abbr.	# Sample	# Feature	P/N	Source Link
Indian Liver Patient Records	LI	583	10	2.51	https://www.kaggle.com/datasets/uciml/indian-liver-patient-records
Pima Indians Diabetes Database	PID	768	8	0.54	https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
Framingham Heart Study	FH	4238	15	0.18	https://www.kaggle.com/datasets/mohannapd/ramingham-heart-study
Stroke Prediction	ST	5110	7	0.04	https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset
Hepatitis C Prediction	HE	615	12	0.11	https://www.kaggle.com/datasets/fedesoriano/hepatitis-c-dataset
COVID-19	CO	5434	20	4.17	https://www.kaggle.com/datasets/hemanthhari/symptoms-and-covid-presence
Lung Cancer Prediction	LC	309	15	6.92	https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer
Heart Failure Prediction	HF	303	13	0.80	https://archive.ics.uci.edu/dataset/45/heart+disease
Early Classification of Diabetes	ECD	520	16	1.60	https://www.kaggle.com/datasets/andrewmvd/early-diabetes-classification
Anemia Disease	AN	15300	24	0.57	https://www.kaggle.com/datasets/serhathoca/anemia-disease
Churn Modeling	-	10000	10	0.26	-
Give Me Some Credit	-	16714	10	1.00	https://www.kaggle.com/c/GiveMeSomeCredit
US Adult Income	-	48842	14	0.31	https://www.kaggle.com/datasets/johnolafenwa/us-census-data

Table 6: Detailed data information of used tabular datasets (10 from the medical domain and 3 from others). "P/N" denotes the amount ratio of positive samples and negative ones.

C RESULTS ON CLINICAL TRIAL DATASETS

We evaluate SERSAL on clinical trail mortality datasets, which require specialized scientific knowledge for clinical trials. Although SERSAL prompting with GPT-3.5 cannot directly achieve good performance on such vertical tasks, further performance gains are still observed once we use more powerful GPT-4, indicating room for continuous improvement as more advanced LLMs appear.

	N00041119	N00174655	N00312208	N00079274	N00694382
FSSM*(supervised FT-T)	62.38	89.20	77.83	71.78	73.89
0-shot (GPT-3.5)	56.79	73.08	63.49	59.85	62.70
CoT (GPT-3.5)	56.79	73.08	60.73	59.85	62.70
SERSAL (GPT-3.5)	**58.31**	**82.64**	**71.92**	**64.17**	**66.31**
SERSAL (GPT-4)	65.08	88.62	78.39	67.94	71.47

Table 7: The AUC scores (%) of different tabular prediction schemes on clinical trail mortality datasets used in Wang & Sun (2022) (see ClinicalTrials.gov). The similar denotations are used as Table 2. No gold labels are used for prompting methods here. It can be seen SERSAL can achieve continuous improvement and even perform comparably with the traditional supervised paradigm once more powerful base LLMs are applied.

[1] https://physionet.org/news/post/415

D Mechanism Explanation of DivideMix in SERSAL

To make the paper friendly to the audiences from different background, in this section we provide detailed mechanism explanation of noisy label learning and how to learn a better small (neural network) model from LLM noisy annotations using DivideMix.

In the traditional noisy data learning field, it was theoretically proved and empirically observed that the "memorization" behavior of neural networks leads to different optimization behavior on real data and noisy ones that neural networks tend to learn simple patterns first before fitting label noise (Arpit et al., 2017). Based on this theoretical foundation, a typical group of noisy label learning methods (Berthelot et al., 2019; Li et al., 2019) exploit per-sample training loss to judge the noisy labels, for example, in our paper we adopt DivideMix (Li et al., 2019) to learn the small model from LLM noisy annotations, which models the noise probabilities of each sample by dynamically fitting a Gaussian Mixture Model (GMM) on per-sample losses, all training samples are divided into a clean set and a noisy set based on a probability threshold. During the DivideMix training process, samples in the clean set are used for supervised learning (using their LLM annotations), while ones in the noisy set is used in an unsupervised manner (only using their features), e.g., regularization loss or reconstruction task. The process will be ended until the average loss of heuristically selected early stopping subset (high-LLM-confidence samples) is converged. Notably, clean sample is not equal to high-LLM-confidence samples, but the sample which is easier to fit by the model. Since the small model is only supervised by clean data and regularized on noisy data, all data is fully and reasonably exploited to acquire a better pattern.