Open Access Library Journal
2025, Volume 12, e14603
ISSN Online: 2333-9721
ISSN Print: 2333-9705

In-Document Table Data Inference Based on LLM-ICL Model

Xinrui Dou

School of Communication and Electronic Engineering, Jishou University, Jishou, China
Email: xinrui.dou@jsu.edu.cn

How to cite this paper: Dou, X.R. (2025) In-Document Table Data Inference Based on LLM-ICL Model. *Open Access Library Journal*, **12**: e14603.
https://doi.org/10.4236/oalib.1114603

Received: November 13, 2025
Accepted: November 30, 2025
Published: December 3, 2025

Abstract

This paper proposes a structured data prediction method based on Large Language Models with In-Context Learning (LLM-ICL). The method designs sample selection strategies to choose samples closely related to the prediction task and converts structured data into text sequences, which are then provided as input to large language models for prediction through in-context learning. To validate the effectiveness of the method, experiments were conducted using the IPUMS dataset. In the few-shot setting with only 10 demonstration samples, Results demonstrate that with extremely limited samples (only 10), the best-performing model Qwen-plus achieves a prediction accuracy of 79.4%, significantly outperforming traditional supervised machine learning algorithms trained on the same sample size (XGBoost at 73.5% and KNN at 71.1%). Further analysis reveals that KNN and XGBoost require approximately 500 and 16,000 samples respectively to achieve comparable accuracy levels to LLM-ICL using just 10 samples. Additionally, sample selection strategy significantly impacts performance—employing nearest neighbor sampling further enhances accuracy compared to random selection. This research demonstrates the substantial potential and application value of LLM-ICL in few-shot structured data prediction tasks.

Subject Areas

Aerospace Engineering, Agricultural Engineering

Keywords

Large Language Models, In-Context Learning, Table Reasoning, Few-Shot Learning

1. Introduction

Since 2022, generative artificial intelligence represented by ChatGPT and GPT4

[1] has garnered widespread attention globally. These advanced systems, known as Large Language Models (LLMs), are built on the Transformer architecture [2] and typically contain hundreds of billions of parameters. Trained on massive text corpora, LLMs can effectively understand natural language instructions and generate human-like responses. Large language models have demonstrated state-of-the-art performance across various natural language processing tasks. For instance, on the question-answering dataset TriviaQA [3], GPT3 improved accuracy by 3.2% over previous best models; on the Multimodal Multitask Language Understanding (MMLU) dataset [4], GPT4 achieved more than 10% accuracy improvement compared to contemporary best models.

Although large language models excel at processing unstructured text data, there is still limited research on their application in prediction tasks primarily involving structured data. Unlike text and image data, structured data is typically presented in tabular form [5]. In such tables, each row represents a prediction instance, while each column represents a feature of that instance, describing its attributes, which can be continuous or discrete values. The fundamental challenge in prediction tasks with structured data is to predict the target variable of an instance based on its attribute data. This target variable may be a continuous numerical value, such as predicting housing prices, or a discrete category label, such as determining whether an email is spam. **Table 1** shows an example of structured data and its target variable to be predicted. The left portion of the table displays various attributes including age, education level, gender, ethnicity, and marital status. Among these, ethnicity and marital status are discrete variables. The right portion shows the prediction target: based on the attributes in the left table, the goal is to predict whether annual salary exceeds $50,000. It is worth noting that age is a continuous variable, while education level, gender, ethnicity, and marital status are discrete variables.

Table 1. Structured data and prediction target.

Age	Education	Gender	Ethnicity	Marital Status	Annual Salary > $50K
49	Bachelor's	Male	USA	Divorced	No
...	...	...	...	...	...
25	Master's	Female	USA	Unmarried	Yes

Table Reasoning (also known as structured data reasoning), particularly applications like table fact verification and question answering, holds significant value in the information age. Due to its advantages of being information-dense and highly structured, tables serve as an indispensable core data format in numerous fields such as finance and public health [1]. Unlike typical data processing tasks, table fact verification and question answering require complex reasoning that integrates unstructured information such as statement/question text or textual content within tables. For instance, tasks include judging the accuracy of a statement based on table data, determining whether a data-

supported diagnosis is reliable, or answering various questions according to the table data [2]. Traditionally, these complex reasoning tasks had to be performed manually, which is not only time-consuming but also prone to errors. Consequently, major international conferences in the field of artificial intelligence, such as ACL [6], ICLR [7], and NIPS [8], have been actively discussing the latest technologies and methods for automated table reasoning in recent years.

Generally speaking, current table reasoning primarily employs two types of methods: 1) Training machine learning models for table reasoning by using large amounts of joint data comprising both tables and text [9]; 2) Leveraging the advantages of large language models (such as the GPT and LLaMA series) in natural language understanding and logical reasoning, either through fine-tuning or utilizing In-Context Learning (ICL), to employ LLMs for table reasoning [10]. However, both approaches have limitations. On one hand, methods based on machine learning models involve complexity and require substantial computational resources during model training and selection, and they struggle with understanding irregular table formats and performing commonsense reasoning. On the other hand, table reasoning methods based on LLM-ICL, while largely addressing the above issues, are still in the preliminary exploration stage and face challenges such as token limitations, uncertainty, and difficulties in debugging, necessitating validation and exploratory improvements based on real-world data.

Based on the context learning capabilities of large language models, this paper proposes a prediction algorithm based on the LLM-ICL model for structured data prediction problems. This method designs a sample selection strategy to select samples closely related to the problem and transforms these samples, which consist of structured data, into text sequences, which are then used as input for context learning and provided to the large language model for prediction. This approach not only enables the large language model to analyze patterns based on the input samples but also integrates the common sense learned by the large language model during the pre-training stage into the analysis process.

2. Related Research

2.1. Large Language Model

LLMs are advanced artificial intelligence models built on deep learning architectures. These models undergo pre-training on massive text corpora, allowing them to learn complex linguistic structures and patterns. Typically comprising hundreds of millions to hundreds of billions of parameters, they exhibit exceptional language comprehension and generation capabilities, allowing them to perform complex natural language processing tasks.

In 2018, OpenAI introduced GPT (Generative Pre-trained Transformer) [11], a language model with 150 million parameters that pioneered the era of large-scale pre-trained language models. Subsequently, models such as GPT2 [12] and GPT3 [13] were released, continuously pushing the boundaries of training dataset sizes,

parameter scales, and performance, thereby advancing the field of NLP. By 2023, multiple companies including Meta, Baidu, Alibaba, and DeepSeek had launched LLMs with parameter counts exceeding tens of billions. Among them, several globally recognized large language models were trained on datasets ranging from hundreds of billions to trillions of tokens.

Compared to traditional deep learning models, LLMs exhibit two distinctive characteristics: first, their parameter complexity is extremely high, with counts ranging from billions to hundreds of billions; second, the scale of their training datasets is immense. These two core features endow LLMs with remarkable comprehension and generation capabilities. Taking the GPT series as an example, GPT1 had 117 million parameters [11], GPT2 increased to 1.5 billion parameters [12], and GPT3 reached a staggering 175 billion parameters [13]. Alongside the significant growth in parameter scale, the training datasets also expanded accordingly. GPT1 was trained on BookCorpus, a dataset containing approximately 7,000 unpublished books with a data volume of about 4.6 GB [11]. GPT2 used WebText, a dataset constructed from 8 million Reddit articles that received at three upvotes, comprising 40 GB of text data [12]. GPT3 utilized an even more extensive training dataset, including two undisclosed book datasets, an expanded version of WebText, and a filtered Common Crawl dataset, totaling over 570 GB of text data [13].

The training corpora for LLMs encompass diverse resources such as books, internet text, news articles, and forum posts. During training, the models learn linguistic rules and knowledge from these corpora and store this information internally as parameters. Consequently, when a model receives a text sequence as input, it can generate corresponding outputs based on these parameters.

2.2. In-Context Learning

In-Context Learning (ICL) enables large language models to perform tasks directly through natural language prompts without requiring parameter updates or additional training. When provided with appropriate task descriptions and representative demonstrations, LLMs can effectively interpret task requirements and generate appropriate responses. The carefully designed input sequence used to guide the model, known as a prompt, typically comprises three essential components: the task specification, demonstration examples, and the target query.

For effective ICL, the task specification must clearly define the objective and expectations, while the demonstration set consists of input-output pairs illustrating similar problems and their solutions. For example, in machine translation, demonstrations would show source texts with their corresponding translations. These illustrative examples help the model recognize task-specific patterns and expected output formats.

Research has shown that model performance strongly depends on prompt construction strategies, including both the formulation of task instructions and the selection of demonstrations. Recent studies indicate that curating examples

based on semantic similarity to the target problem significantly enhances performance in reasoning tasks. The LENS methodology further advances this approach by employing support vector selection for demonstration assembly, demonstrating consistent accuracy improvements across various text classification benchmarks.

3. Prediction Methods Based on Large Language Models

In this section, we formalize our proposed approach for structured data prediction using LLMs with in-context learning.

Let $X \in \mathbb{R}^d$ be a d -dimensionalrandom vector representing the feature vector of the known structured data, and $Y \in \mathbb{R}$ denote the target value to be predicted. Assume that N_1 samples of X and their corresponding target values have been observed, and that $Y \in \mathbb{R}$ follows some distribution D .The goal of this structured data prediction task is to predict the corresponding Y value for any given X .

Supervised machine learning algorithms transform this prediction problem into a parameter optimization problem: define $f(X,\theta)$ as a function with parameters θ , and $L(f(X),Y)$ as a loss function. Based on the observed sample set $D_1 = \{(X_i, Y_i), i = 1,2,\cdots,N_1\}$, solve the optimization problem:

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N_1} \sum_{i=1}^{N_1} L\left(f\left(\vec{X}_i, \theta\right), Y_i\right), \tag{1}$$

To obtain $f(X,\theta)$, which is then used to predict Y for any X . Common loss functions include Mean Squared Error (MSE) and Cross-Entropy. MSE measures the average squared difference between predicted and true values, typically used for regression problems:

$$L_{\text{MSE}} = \frac{1}{N_1} \sum_{i=1}^{N_1} \left(f\left(X_i, \theta\right) - Y_i\right)^2. \tag{2}$$

Cross-Entropy measures the discrepancy between predicted and true probability distributions, commonly applied to classification problems:

$$L_{\text{CE}} = -\frac{1}{N_1} \sum_{i=1}^{N_1} \sum_{c=1}^{C} y_{i,c} \log\left(f_c\left(X_i, \theta\right)\right), \tag{3}$$

where C is the number of classes and $y_{i,c}$ is the binary indicator for class c .

After estimating parameters $\hat{\theta}$, performance is validated on test dataset $D_2 = \{(X_i, Y_i), i = 1,2,\cdots,N_2\}$:

$$\mathbb{E}_{(X,Y)\sim D}\left[L\left(f\left(X,\hat{\theta}\right),Y\right)\right] \approx \frac{1}{N_2} \sum_{i=1}^{N_2} L\left(f\left(X_i,\hat{\theta}\right),Y_i\right). \tag{4}$$

Unlike supervised methods, LLMs require no parameter training but rely on prompt design. Let $p(X_i, Y_i)$ and $p(X_t)$ denote text sequences for the demonstration examples and target problem respectively. Let $\mathcal{C}$ represent the mapping from structured data to text sequences. The demonstration set C is a subset of training dataset D . Define $F : X_t \to C$ as the strategy for selecting

demonstrations given target feature vector X_t. The prediction using LLMs is:

$$P_{\text{LLM}}\left(Y_t \mid \left\{p\left(X_i, Y_i\right) : \left(X_i, Y_i\right) \in F\left(X_t\right)\right\}, p\left(X_t\right)\right). \tag{5}$$

The optimization problem becomes:

$$\hat{p}, \hat{F} = \arg\min_{p,F} \frac{1}{N_1} \sum_{i=1}^{N_1} L\left(P_{\text{LLM}}\left(Y_i \mid \left\{p\left(X_j, Y_j\right) : \left(X_j, Y_j\right) \in F\left(X_i\right)\right\}, p\left(X_i\right)\right), Y_i\right). \tag{6}$$

Two key elements are: determining F (demonstration selection strategy) and defining mapping p (structured data to text).

The demonstration set C significantly impacts performance. This paper designs two selection strategies:

- **Random Selection**: Randomly select K samples from D_1
- ***K*-Nearest Neighbors**: Select K samples minimizing:

$$\text{dist}\left(X_t, X_i\right) = \sum_{j=1}^{d} \left(x_{t,j} - x_{i,j}\right)^2, \tag{7}$$

where $x_{t,j}$ and $x_{i,j}$ are the j-th components of X_t and X_i.

For textual representation, samples are converted as:

$$p\left(X_i, Y_i\right) = \text{“Attribute}_1 = x_{i,1}, \text{Attribute}_2 = x_{i,2}, \cdots, \text{Attribute}_d = x_{i,d}.\text{Target} = Y_i\text{”}$$

$$p\left(X_t\right) = \text{“Attribute}_1 = x_{t,1}, \text{Attribute}_2 = x_{t,2}, \cdots, \text{Attribute}_d = x_{t,d}.\text{Target} = \_\_\_\text{”}$$

where Attribute_j is the j-th attribute name, $x_{i,j}$ and $x_{t,j}$ are attribute values, and Y_i is the target value.

4. Technical Approach and Experimental Methods

4.1. Technical Roadmap

This paper aims to build an automated fact verification and question answering system based on LLM-ICL technology for tabular data analysis. Our research initiative serves two primary purposes: firstly, it provides the basic functional framework for subsequent research; secondly, it explores the potential of current LLM-ICL tabular reasoning methods in practical applications and identifies potential problems. A key technical challenge we address is the contradiction between token restrictions and reasoning with large tabular data. To systematically tackle this challenge, our research can be divided into eight sub-tasks, unfolding in three phases, with the specific technical roadmap shown in **Figure 1**.

4.2. System Setup

4.2.1. Data Preparation

This paper uses the IPUMS dataset as an example to validate the effectiveness of the structured data prediction algorithm based on large language models. The dataset originates from the Integrated Public Use Microdata Series (IPUMS) database and includes various features describing individual social attributes such as age, occupation category, education level, and profession. The prediction target is to determine whether an individual's annual income exceeds $50,000.

Relying on social common sense, one can make a reasonable prediction of an individual's income level based solely on social attributes provided by the dataset, such as "age" and "occupation category", without the need for complex data analysis. This characteristic provides an experimental basis for investigating whether the prior knowledge embedded in the parameters of large language models can positively influence the prediction task.

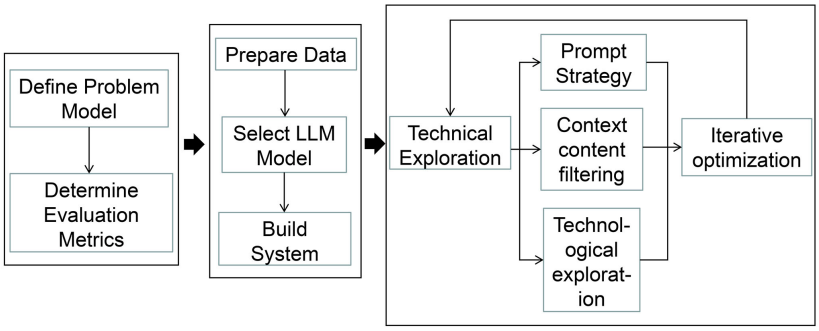

Figure 1. Technical roadmap for LLM-ICL-Based table reasoning system development.

Specifically, the dataset includes 14 attribute variables, *i.e.*, $d=14$, with the feature vector $\vec{X} \in \mathbb{R}^{14}$, where each component corresponds to an attribute in Table 2 (one row). Among these, eight are discrete variables, including work unit nature, education level, marital status, occupation, social relationship, race, gender, and nationality; six are continuous variables, including age, census enumerator serial number, years of education, capital gains, capital losses, and hours worked per week. Table 2 provides the variable types and specific meanings. The prediction target is whether an individual's annual income exceeds $50,000, *i.e.*, $Y \in \mathbb{R}$, where $Y=1$ indicates that the individual's annual income exceeds $50,000, and $Y=0$ indicates that it does not.

Table 2. Variable types and descriptions of the IPUMS dataset.

Variable Name	Variable Type	Description
age	Continuous	Age
workclass	Discrete	Nature of work unit
fnlwgt	Continuous	Census enumerator serial number
education	Discrete	Education level
education-num	Continuous	Years of education
marital-status	Discrete	Marital status
occupation	Discrete	Occupation
relationship	Discrete	Social relationship
race	Discrete	Race
sex	Discrete	Gender
capital-gain	Continuous	Capital gains
capital-loss	Continuous	Capital losses
hours-per-week	Continuous	Hours worked per week
native-country	Discrete	Native country

4.2.2. Selection of LLM Models

Given the diversity of Large Language Models (LLMs), several different LLM models-including GPT-3.5, Ernie-lite,Qwen-plus, Deepseek-V3.1, and Llama-4-will be used for systematic evaluation. A performance comparison will be conducted to determine the performance of each model across various specific tasks, including processing speed, accuracy, and adaptability to the economic and social development data of Xiangxi.

4.2.3. System Setup

Based on the performance evaluation results, the selected models will be integrated and optimized to improve the input-output process and adapt to the economic and social development data of western Hunan. Furthermore, multi-model fusion strategies will be explored so that the system can automatically select the most suitable model or model combination based on the nature of the question. Through model selection and strategy optimization, an efficient and accurate question-answering system prototype will be built, and system construction and preliminary testing will be conducted to ensure its stability and response speed, laying the foundation for the next stage of technological exploration and iterative improvement.

4.3. Technological Exploration

To improve the performance of table reasoning, this study will explore technologies from several aspects, including prompting strategies, selection of contextual content, and table serialization methods, to address challenges such as token quantity limitations. The accuracy of the system's fact verification and question answering based on the evaluation indicators determined in the previous stage will be iteratively optimized.

Prompting Strategies: By designing effective prompting strategies, the model can be guided to reason along the correct thought path, thereby improving the quality and accuracy of the answers. A prompting strategy can be selected at runtime based on the specific statement/question. Various prompting strategies, including direct prediction, self-consistent decoding, thought chain, thought tree, and thought graph, will be tested.

Contextual Content Selection: Since LLMs typically have input token length limitations and computational costs are positively correlated with the number of tokens, how to fit as much accurate and relevant contextual content as possible into a limited input space is a key issue. Solving this problem is of great significance for complex data and large table data application scenarios. The plan is to draw on the large table processing methods in research and the vector database-similarity search method commonly used in current research.

Table serialization refers to transforming complex tabular data into a format more suitable for language models to process, thereby improving the LLM's ability to understand and process table content. Currently testable and improveable table serialization methods include table-to-text description conversion, key-value pair

conversion, entity relation extraction, and cell coordinate encoding.

5. Computer Experiments

5.1. Overall Evaluation Results

To rigorously evaluate the predictive performance of LLM-ICL in few-shot settings, this study employs a dual baseline approach, using XGBoost as the representative parametric machine learning algorithm and KNN for non-parametric approaches.

XGBoost: Developed by Chen Tianqi and colleagues, XGBoost [14] is an open-source machine learning tool based on the second-order tree algorithm GBDT (Gradient Boosting Decision Tree). The tool incorporates significant improvements to the original GBDT method at both algorithmic and engineering levels. In recent years, XGBoost has emerged as one of the most stable and effective machine learning tools available.

KNN: The K-Nearest Neighbors algorithm, introduced by Cover *et al.* [15], represents a non-parametric machine learning approach. This method identifies the K closest samples to the target instance in the feature space and determines the classification based on the majority class among these neighbors. KNN's strength lies in its simplicity and intuitive nature, as it operates without relying on specific assumptions or model structures.

To assess LLM-ICLs' predictive capability with limited samples, this research randomly selects 10 samples from the IPUMS dataset as the training set for both XGBoost and KNN algorithms, denoted as $D_1 = \{(X_i, Y_i), i = 1, 2, \cdots, 10\}$, while using 100 randomly chosen samples as the test set, represented as $D_2 = \{(X_i, Y_i), i = 1, 2, \cdots, 100\}$. A grid search strategy is implemented to optimize hyperparameters for both XGBoost and KNN algorithms. The optimal hyper-parameter values are determined through five-fold cross-validation on the training dataset.

When applying the proposed LLM-ICL-based prediction framework, the "demonstration dataset" in the prompt consists of exactly 10 samples, meaning the contextual learning for LLM-ICLs is constrained to this limited sample size. For comprehensive evaluation of LLM-ICL performance, this study examines several prominent models: OpenAI's GPT-3.5-turbo [16], Baidu's Ernie-lite [17], Alibaba's Qwen-plus [18], DeepSeek's Deepseek-V3.1 [19], and Meta's Llama-4 [20].

The evaluation employs accuracy, the most commonly used metric for binary classification tasks, defined as the proportion of correctly predicted samples. LLM-ICL outputs of "yes" correspond to predictions of "annual income exceeding $50,000", while "no" indicates "annual income below $50,000". If an LLM-ICL generates "uncertain" as output, it is considered an incorrect prediction regardless of the actual income level. To mitigate random variability, the study adopts a multiple random sampling strategy, with the final accuracy metric representing the average across all test iterations.

Table 3 presents the accuracy performance of XGBoost and KNN on the test dataset with equivalent training sample sizes, compared against LLM-ICLs utilizing the same 10 samples as demonstration examples. With a training sample size of N = 10, The XGBoost and KNN models achieve accuracy rates of 0.735 and 0.711, respectively. Although performance fluctuates across different sample sets, this value consistently falls below the accuracy levels attained by LLM-ICLs using equivalent sample quantities. These results demonstrate that LLM-ICLs substantially outperform supervised machine learning algorithms in scenarios with limited training data.

Further analysis reveals that when employing only 10 demonstration samples, the top-performing Qwen-plus model achieves an accuracy of 0.794. Remarkably, KNN and XGBoost require approximately 500 and 16,000 samples respectively to reach comparable accuracy levels. This finding underscores LLM-ICLs' exceptional capability in handling structured data prediction tasks. Even with minimal samples, LLM-ICLs leverage their powerful learning capacities to achieve accuracy comparable to supervised machine learning models trained on substantially larger datasets, highlighting their significant potential for structured data prediction applications.

Table 3. Accuracy comparison of different algorithms on test dataset (presented as mean ± standard deviation).

Model Category	Model Name	Accuracy
2*Supervised Machine Learning	KNN	0.711 ± 0.018
	XGBoost	0.735 ± 0.022
5*LLM-ICL Models	GPT-3.5-turbo	0.749 ± 0.010
	Ernie-lite	0.768 ± 0.007
	Qwen-plus	0.794 ± 0.009
	Deepseek-V3.1	0.776 ± 0.014
	Llama-4	0.759 ± 0.011

5.2. In-Depth Analysis of Sample Strategies and Scaling Effects

Table 4 presents a systematic comparative analysis of how "random sampling" and "nearest neighbor selection" strategies affect large language model prediction performance across varying sample sizes. In this carefully controlled experimental series, all data features were converted using natural language descriptions to ensure fair comparison. The experimental results unequivocally demonstrate that when maintaining identical sample quantities, models utilizing the nearest neighbor strategy for constructing demonstration sets significantly outperform those using random sampling. This consistent pattern confirms that selecting demonstration samples with high similarity to target samples effectively enhances model performance in structured data prediction tasks, thereby aligning with and extending findings observed in natural language processing applications.

As the number of demonstration samples increased from 3 to 10, all models exhibited steady improvement in prediction accuracy, regardless of the selection strategy employed. This indicates that providing more examples initially helps the model better understand the task pattern.

However, when sample size was further expanded to 15, certain models showed performance degradation. This phenomenon reveals a crucial characteristic of LLMs' in-context learning mechanism: model performance does not continuously improve with increasing sample size.

Potential explanations for this pattern include: first, the "lost-in-the-middle" effect, where excessively long input sequences may impair models' ability to identify critical information; second, an overabundance of demonstration examples could complicate pattern recognition processes, as the model struggles to identify the most relevant reasoning path among too many examples.

Additionally, the fixed context window of Transformer architectures forces a trade-off: more demonstration samples consume limited context resources, potentially leaving insufficient space for the model to effectively process the target query.

Table 4. Prediction accuracy under different sample selection strategies and sample sizes.

LLM-ICL Model	Selection Strategy	$N=3$	$N=10$	$N=15$
2*GPT-3.5-turbo	Random Strategy	0.736	0.749	0.745
	Nearest Neighbor Strategy	0.748	0.757	0.754
2*Ernie-lite	Random Strategy	0.752	0.768	0.758
	Nearest Neighbor Strategy	0.764	0.781	0.788
2*Qwen-plus	Random Strategy	0.775	0.794	0.788
	Nearest Neighbor Strategy	0.788	0.803	0.805
2*Deepseek-V3.1	Random Strategy	0.761	0.776	0.783
	Nearest Neighbor Strategy	0.774	0.788	0.786
2*Llama-4	Random Strategy	0.742	0.759	0.774
	Nearest Neighbor Strategy	0.754	0.771	0.788

6. Conclusions

This paper presents a novel LLM-ICL framework for structured data prediction, demonstrating significant advantages in few-shot learning scenarios. Experimental results on the IPUMS dataset show that our method achieves 79.4% accuracy with only 10 demonstration samples, substantially outperforming traditional machine learning approaches. The nearest neighbor sample selection strategy proves particularly effective, while an optimal sample size (10 - 15) balances information richness and model capacity.

Despite token limitations and model dependency challenges, this work validates LLM-ICL's potential for tabular data reasoning, especially in data-scarce

environments. Future research will focus on optimizing context selection strategies, exploring efficient table serialization methods, and extending the framework to multi-modal and cross-domain tabular reasoning tasks to enhance practical applicability.

Fund Project

This work was supported in part by the Natural Science Foundation of China under Grants 62466018, and by the Jishou University Scientific Research Project Jd24007.

Conflicts of Interest

The author declares no conflict of interest.

References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Ale-man, F.L., *et al.* (2023) GPT-4 Technical Report. arXiv: 2303.08774.

[2] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., *et al.* (2023) A Survey on Vision Transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 87-110. https://doi.org/10.1109/tpami.2022.3152247

[3] Joshi, M., Choi, E., Weld, D. and Zettlemoyer, L. (2017) TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Vancouver, 30 July-4 August 2017, 1601-1611. https://doi.org/10.18653/v1/p17-1147

[4] Arulraj, A., Chandra, A., Chen, W., Fan, R., Guo, S., He, X., *et al.* (2024) MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark. *Advances in Neural Information Processing Systems* 37, Vancouver, 10-15 December 2024, 95266-95290. https://doi.org/10.52202/079017-3018

[5] Cafarella, M.J., Halevy, A. and Madhavan, J. (2011) Structured Data on the Web. *Communications of the ACM*, **54**, 72-79. https://doi.org/10.1145/1897816.1897839

[6] Markatos, K., Kaseta, M.K., Lallos, S.N., Korres, D.S. and Efstathopoulos, N. (2012) The Anatomy of the ACL and Its Importance in ACL Reconstruction. *European Journal of Orthopaedic Surgery & Traumatology*, **23**, 747-752. https://doi.org/10.1007/s00590-012-1079-8

[7] Park, C.F., Lee, A., Lubana, E.S., Yang, Y.Y., Okawa, M., Nishi, K., Wattenberg, M. and Tanaka, H. (2024) ICLR: In-Context Learning of Representations. arXiv: 2501.00070. https://arxiv.org/abs/2501.00070

[8] Goodfellow, I. (2016) NIPS 2016 Tutorial: Generative Adversarial Networks. arXiv: 1701.00160. https://arxiv.org/abs/1701.00160

[9] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D. and Steinhardt, J. (2020) Measuring Massive Multitask Language Understanding. arXiv: 2009.03300. https://arxiv.org/abs/2009.03300

[10] Vacareanu, R., Negru, V.A., Suciu, V. and Surdeanu, M. (2024) From Words to Numbers: Your Large Language Model Is Secretly a Capable Regressor When Given In-Context Examples. arXiv: 2404.07544. https://arxiv.org/abs/2404.07544

[11] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., *et al.* (2018) Improving Lan-

guage Understanding by Generative Pretraining.
https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

[12] Radford, A., Child, R., Luan, D., Amodei, D., Sutskever, I., *et al.* (2019) Language Models Are Unsupervised Multitask Learners.
https://storage.prod.researchhub.com/uploads/papers/2020/06/01/language-models.pdf

[13] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., *et al.* (2020) Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems*, **33**, 1877-1901.

[14] Friedman, J., Tibshirani, R. and Hastie, T. (2000) Additive Logistic Regression: A Statistical View of Boosting (with Discussion and a Rejoinder by the Authors). *The Annals of Statistics*, **28**, 337-407. https://doi.org/10.1214/aos/1016120463

[15] Larose, D.T. and Larose, C.D. (2014) k-Nearest Neighbor Algorithm.
https://doi.org/10.1002/9781118874059.ch7

[16] Wang, W.H., Wang, S.Y., Huang, J.Y., Liu, X.D., Yang, J., Liao, M., *et al.* (2023) An Investigation Study on the Interpretation of Ultrasonic Medical Reports Using OpenAI's GPT-3.5-Turbo Model. *Journal of Clinical Ultrasound*, **52**, 105-111.
https://doi.org/10.1002/jcu.23590

[17] Shang, J., Yu, W. and Chen, J. (2024) Crowdsourcing Canada Goldenrod Identification from Multimodal Weibo Data. *Proceedings of the* 24*th ACM/IEEE Joint Conference on Digital Libraries*, Hong Kong, 16-20 December 2024, 1-5.
https://doi.org/10.1145/3677389.3702581

[18] Pan, S., Liu, K., Chen, W. and He, B. (2024) Performance Analysis of Chinese Large Language Models in Solving Math Word Problems. 2024 *International Conference on Intelligent Education and Intelligent Research* (*IEIR*), Macau, 6-8 November 2024, 1-8. https://doi.org/10.1109/ieir62538.2024.10960109

[19] Taşyürek, M., Adıgüzel, Ö., Gündoğar, M., Goncharuk-Khomyn, M. and Ortaç, H. (2025) Comparative Evaluation of the Responses from ChatGPT-5, Gemini 2.5 Flash, and Deepseek-V3.1 Chatbots to Patient Inquiries about Endodontic Treatment in Terms of Accuracy, Understandability, and Readability. *International Dental Research*, **15**, 91-95. https://doi.org/10.5577/intdentres.662

[20] Cui, J.J., Wang, P.L., Holmes, J., Sun, L.S., Hinni, M.L., Pockaj, B.A., Vora, S.A., Sio, T.T., Wong, W.W., Yu, N.Y., *et al.* (2025) An Automated Retrieval-Augmented Generation LLaMA-4 109B-Based System for Evaluating Radiotherapy Treatment Plans. arXiv: 2509.20707. https://arxiv.org/abs/2509.20707