# COMPARING TRADITIONAL MACHINE LEARNING AND LLM ICL ON STRUCTURED AND TEXT TASKS

**Long Jiyuan (120090850), Zeng Yaqi (223040194), Zhang Zhitao (224040254)**
The Chinese University of Hong Kong(Shenzhen)
Shenzhen, China
[120090850,223040194,224040254]@link.cuhk.edu.cn

## ABSTRACT

**This study presents a comprehensive empirical evaluation comparing the performance of traditional machine learning methods with in-context learning (ICL) approaches using large language models (LLMs) across diverse classification tasks. We systematically investigate the effectiveness of these methods on both structured datasets (Iris, Digits, Wine) and text classification benchmarks (AG News, SST2, TREC), while also examining the impact of model size and prompt design on ICL performance. Our results demonstrate that traditional machine learning methods maintain superior performance on structured data tasks (achieving up to 100 accuracy), while ICL shows particular strength in semantic understanding tasks like sentiment analysis (95% accuracy on SST2). Furthermore, we reveal that increasing LLM size improves ICL accuracy, but architecture and prompt engineering are equally critical factors. The study provides practical insights for method selection in different application scenarios and establishes guidelines for optimizing ICL implementations.**

## 1 INTRODUCTION

The rapid advancement of large language models has introduced in-context learning (ICL) as a promising alternative to traditional machine learning approaches for various classification tasks. This paradigm shift raises fundamental questions about when and how to employ these distinct methodologies effectively.

**Research Topic** Our research focuses on conducting a systematic comparison between traditional machine learning methods and ICL approaches across different types of classification tasks. We examine their relative strengths, limitations, and optimal application scenarios through rigorous experimentation.

**Importance of the Task** Understanding the comparative performance of these approaches is crucial for several reasons. First, it guides practitioners in selecting appropriate methods for specific problems, potentially saving computational resources and development time. Second, it reveals fundamental differences in how these methods process information and make decisions. Third, as ICL becomes more prevalent, establishing its boundaries and optimal use cases relative to established machine learning techniques has significant practical implications for real-world applications.

**Achievement of the Task** We designed controlled experiments across six benchmark datasets spanning structured data and text classification tasks. Our evaluation included seven traditional machine learning algorithms and multiple LLMs of varying sizes (1.5B to 7B parameters). Through quantitative analysis, we identified clear patterns in method effectiveness: traditional approaches excel at structured data processing while ICL shows advantages in semantic understanding tasks. We further demonstrated that ICL performance depends non-linearly on prompt length and that model architecture affects results as much as size. These findings provide actionable insights for method selection and implementation in practical applications.

## 2 Experiment Design

The aim of this experiment is to compare the performance of traditional machine learning methods and In Context Learning (ICL) on different types of datasets, and to identify directions for improving the classification performance of ICL. The experiment covered structured and textual datasets, and compared the accuracy of multiple algorithms through a systematic evaluation process, generating visual results.

### 2.1 Definition of the task

The core task focuses on classification problems within a supervised learning framework, covering two distinct types of datasets: classic machine learning datasets (Iris, Digits, Wine) and natural language processing benchmark datasets (AG News, SST2, TREC). The former involves multi-class classification tasks based on numerical features, such as predicting iris species based on floral morphological characteristics or identifying wine origins using chemical composition. The latter includes text classification tasks such as news categorization, sentiment analysis, and question-type recognition.

### 2.2 Dataset Selection

This study employs two categories of benchmark datasets to evaluate model performance: **structured datasets** (numerical/tabular data) and **text classification datasets**. All datasets are sourced from public standard libraries.

**Structured Datasets** The structured dataset consists of three datasets: Iris dataset, Digits dataset, and Wine dataset.

Iris Dataset This dataset contains 150 samples with four botanical measurements (sepal length, sepal width, petal length, petal width) for three-class flower species (Iris setosa, Iris versicolor, Iris virginica) classification, representing a classic small-scale multivariate analysis challenge.

Digits Dataset This dataset comprises 1,797 grayscale 8×8 pixel images of handwritten numerals (0-9), used for 10-class of digits classification, testing high-dimensional pattern recognition capabilities.

Wine Dataset This dataset provides 178 samples with 13 chemical attributes across three imbalanced Italian wine varieties, evaluating performance on limited, non-uniform data.

**Text Classification Datasets** The Text Classification dataset consists of three standard NLP benchmarks : AG news dataset, Stanford sentiment treebank dataset, and TREC question dataset. These datasets serve as important benchmarks for text classification research, addressing different aspects such as long-text comprehension, sentiment analysis, and question classification.

AG News Dataset This dataset is a large-scale long-text classification benchmark, consisting of 120,000 training samples and 7,600 test samples. Each data entry includes a news headline and summary, requiring classification into one of four topic categories: World, Sports, Business, or Sci-Tech. This dataset effectively evaluates a model's ability to understand semantic hierarchies.

Stanford Sentiment Treebank Dataset This dataset provides 67,349 movie review sentences and 872 test samples, requiring sentence-level binary classification into positive or negative sentiment. Its distinguishing feature lies in its rich syntactic structure at the sentence level, making it suitable for fine-grained sentiment analysis research.

TREC Question Dataset This dataset focuses on short-text classification, containing 5,452 training questions and 500 test questions. It requires categorizing open-domain questions into six imbalanced semantic classes, such as ”Description.” This dataset is particularly useful for testing a model's ability to capture diverse semantic intents.

### 2.3 Data Splitting and Prepossessing

All datasets were partitioned via **stratified sampling** into training (80%) and test sets (20%), with a fixed random seed (42) for reproducibility. Text data were vectorized (TF-IDF) for traditional ML models, while retained in raw form for in-context learning (ICL) with large language models.

### 2.4 Experiment Methods

In terms of methodological comparison, the experiment selects seven representative traditional machine learning algorithms, including Logistic Regression, Decision Tree, Random Forests, Support Vector Machines, Gradient Boosting Trees, Naive Bayes and K-Nearest Neighbors, all implemented with default parameter settings to ensure fairness. The ICL approach, as a contrast, is implemented using the DeepSeek-chat large language model. By designing few-shot learning prompt templates, test samples are combined with carefully selected training examples to form contextual demonstrations, from which the model generates predictions. This approach leverages the implicit knowledge transfer capabilities of large language models, eliminating the need for explicit model training.

## 3 Experiments

### 3.1 Comparison of Classification Performance between Traditional ML Methods and ICL

This part aims to delve into two core questions. First, in which task—structured data classification or text classification—does ICL demonstrate unique performance advantages? Second, what are the key factors and underlying mechanisms responsible for the performance differences observed in these two tasks? Addressing these questions will provide theoretical guidance for algorithm selection in practical applications.

#### 3.1.1 Experiment Procedure

This study employs a controlled variable approach to design a systematic comparative experiment, selecting seven classic traditional machine learning algorithms (including SVM, Random Forest, Decision Tree, K-Nearest Neighbors, Logistic Regression, Naive Bayes, and Gradient Boosting) along with an ICL implementation based on the DeepSeek model for performance comparison. In the ICL experiments, a 20-shot prompting strategy was adopted to ensure fair comparison with traditional machine learning methods in terms of sample utilization efficiency. The evaluation primarily utilizes classification accuracy as the metric to intuitively reflect the models' overall performance on the classification task, facilitating horizontal comparison across different algorithms.

#### 3.1.2 Quantitative Results

**Performance in Structured Datasets** Experimental results on structured datasets show that traditional machine learning methods exhibit outstanding classification performance. Specifically, the KNN algorithm achieves 100% accuracy on the Iris dataset, SVM attains 99.17% accuracy on the Digits dataset, and random forest achieves perfect classification on the Wine dataset. In contrast, ICL only reaches an average accuracy of 70% on these datasets, with particularly poor performance on the Digits dataset (40%).

| | **Iris** | **Digits** | **Wine** |
|---|---|---|---|
| Best Traditional ML Method | KNN | SVM | RF |
| Best Traditional ML Method Accuracy | 100% | 99.17% | 100% |
| Average Traditional ML Method Accuracy | 97.15% | 92.78% | 90.47% |
| ICL Accuracy | 90% | 40% | 80% |

Table 1: Accuracies of Different Methods (Vertical Format)

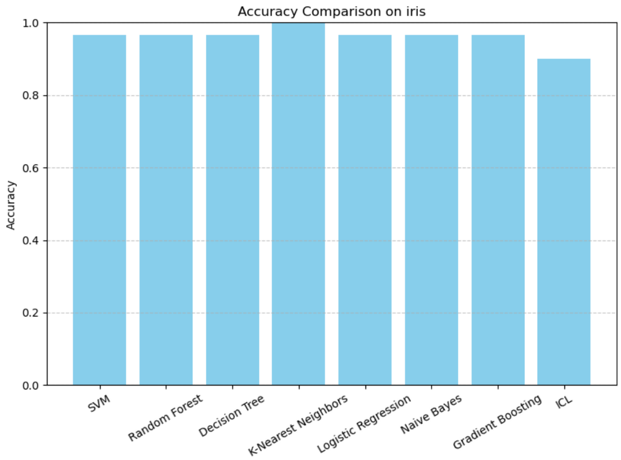

Figure 1: Accuracy Comparison on Iris Dataset

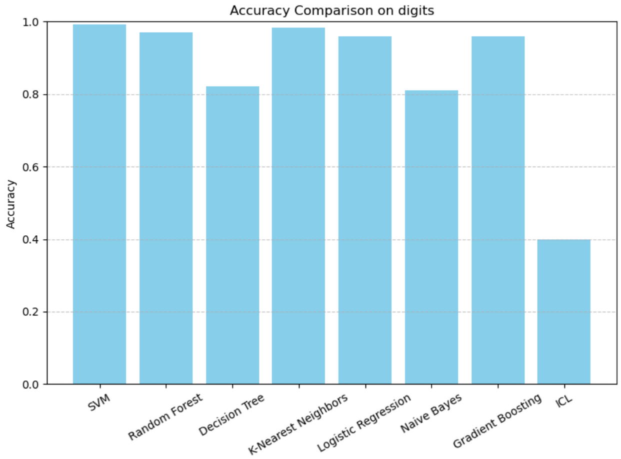

Figure 2: Accuracy Comparison on Digits Dataset

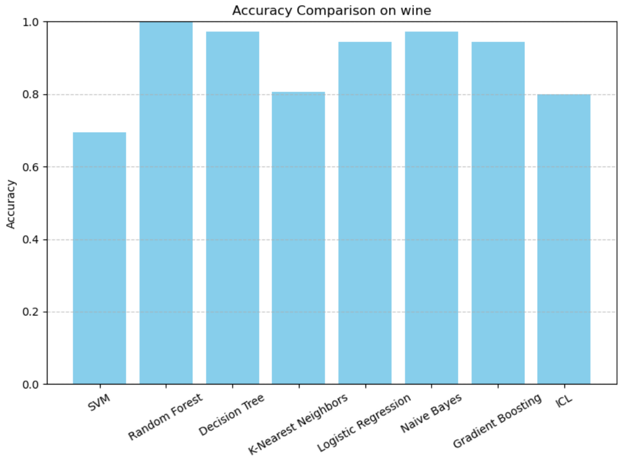

Figure 3: Accuracy Comparison on Wine Dataset

**Performance in Text Classification Datasets** The experimental results on text datasets reveal a different pattern. On the AG News and TREC datasets, traditional methods (KNN and GBDT) achieve accuracies of 75% and 62%, respectively, slightly outperforming ICL's 60% and 40%. However, in the SST2 sentiment analysis task, ICL significantly surpasses traditional methods with 95% accuracy compared to their 65% performance.

| | **AG News** | **SST2** | **TREC** |
|---|---|---|---|
| Best Traditional ML Method | KNN | LR | GBDT |
| Best Traditional ML Method Accuracy | 75% | 65% | 62% |
| Average Traditional ML Method Accuracy | 64.86% | 57% | 54.29% |
| ICL Accuracy | 60% | 95% | 40% |

Table 2: Accuracies of Different Methods (Vertical Format)

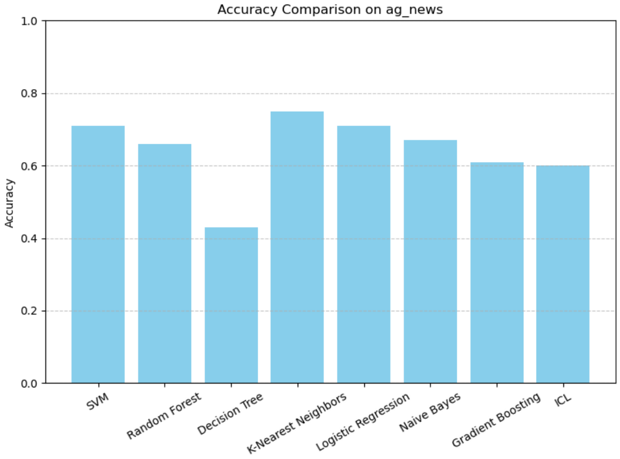

Figure 4: Accuracy Comparison on AG News Dataset

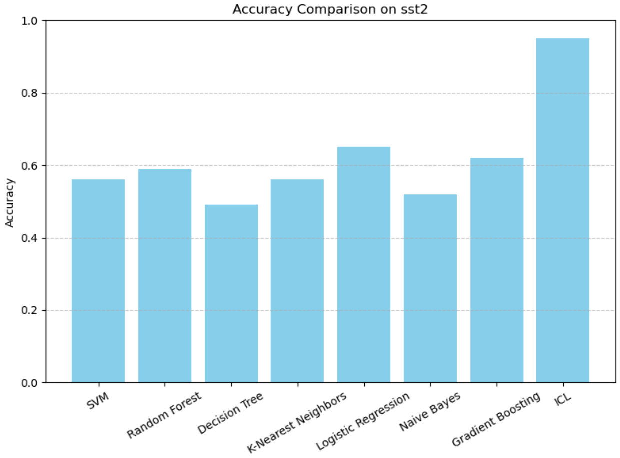

Figure 5: Accuracy Comparison on SST2 Dataset

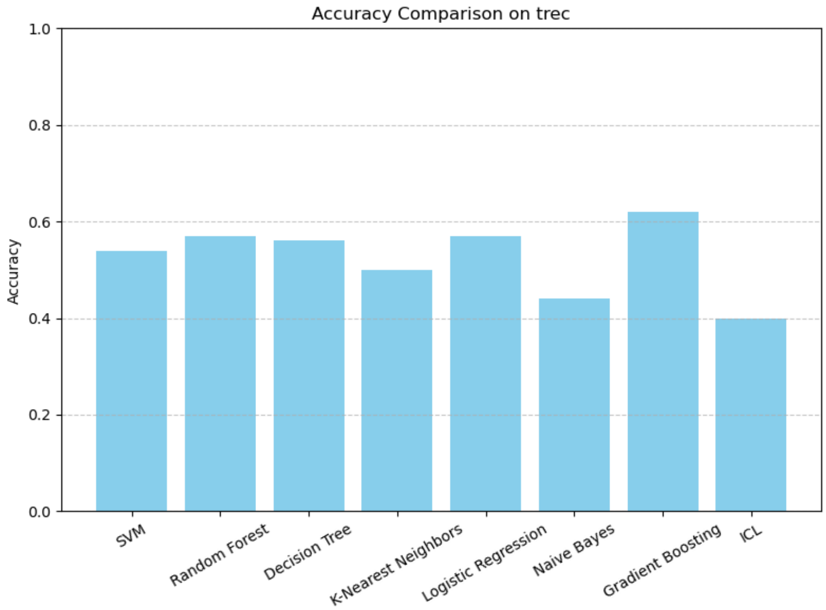

Figure 6: Accuracy Comparison on TREC Dataset

### 3.1.3 Results Analysis

**Performance in Structured Datasets** The performance advantages of traditional machine learning in structured dataset may stem from multiple aspects. Firstly, numerical features can be directly used for distance calculations and decision boundary partitioning, avoiding information loss during feature transformation. Secondly, the inductive bias of traditional algorithms aligns better with the characteristics of structured data. For example, the recursive partitioning of feature space by decision trees highly matches the distribution properties of numerical data. Additionally, ICL requires converting numerical features into textual descriptions, a process that inevitably leads to information loss and distortion of the original data, thereby affecting classification performance.

**Performance in Text Classification Datasets** The different performance of ICL in text classification tasks warrants in-depth exploration.

Task Nature and ICL Adaptability In terms of task nature and ICL adaptability, the performance differences across various text classification tasks exhibit notable patterns. Sentiment analysis tasks achieve 95% accuracy, primarily due to the inherent advantages of large language models in semantic understanding. Sentiment polarity judgment, as one of the most frequently encountered task types during pre-training, has relatively simple patterns and strong subjectivity, aligning well with human annotation standards. In contrast, the moderate performance of news classification tasks at 60% reflects the limitations of ICL in handling complex topics. The thematic diversity and domain specificity of news texts lead to ambiguous class boundaries, while long-text features disperse key information, collectively constraining ICL's effectiveness. The low accuracy of question classification tasks at 40% highlights ICL's shortcomings in fine-grained classification, particularly when tasks require external knowledge support or suffer from imbalanced training data distributions.

Intrinsic Properties of Language Models The intrinsic properties of language models play a decisive role in ICL performance. Pre-training data composition bias is a primary factor: the abundance of sentiment-expressing texts in pre-training corpora provides a solid foundation for sentiment analysis tasks, whereas specialized question-type texts are relatively scarce. From a model architecture perspective, the self-attention mechanism of Transformers excels at capturing global semantic features, explaining why sentiment analysis tasks—which require holistic contextual understanding—perform well. However, tasks requiring precise entity recognition and fine-grained reasoning show limited effectiveness. Additionally, language models' insufficient ability to handle long-range dependencies directly impacts their performance on long-text news classification.

PROMPT ENGINEERING FACTORS Prompt engineering factors exhibit differential effects across tasks. For sentiment analysis tasks, simple prompt templates paired with representative examples yield good results, owing to the intuitive nature of sentiment judgment tasks. News classification tasks, however, require carefully designed example paragraphs containing category-defining content, which is challenging to ensure in practice due to variations in information density and representativeness. Question classification tasks are the most sensitive to prompt design, necessitating not only examples of typical question structures but also explicit explanations of category definitions, placing higher demands on the completeness of prompt engineering.

This study yields several important theoretical insights: First, ICL's capability boundaries exhibit clear patterns, excelling in pattern-recognition tasks but showing limited effectiveness in tasks requiring precise knowledge or complex reasoning. Second, the alignment between task characteristics and ICL's strengths is a key determinant of performance, with subjectively and semantically demanding tasks being more suitable for ICL. Finally, ICL and traditional methods should complement rather than replace each other. Future research should develop more systematic task-method matching frameworks to achieve optimal performance across scenarios. These findings provide valuable references for deepening the understanding of ICL's mechanisms and advancing its practical applications.

### 3.1.4 CONCLUSION

Experimental results on structured datasets demonstrate the outstanding classification performance of traditional machine learning methods: KNN achieves 100% accuracy on the Iris dataset, SVM reaches 99.17% on the Digits dataset, and random forest achieves perfect classification on the Wine dataset, while ICL only attains an average accuracy of 70% (dropping to 40% on the Digits dataset). Results on text datasets show a divergent trend: traditional methods (KNN and GBDT) slightly outperform ICL on AG News (75% vs. 60%) and TREC (62% vs. 40%), but ICL significantly surpasses traditional methods (95% vs. 65%) in the SST2 sentiment analysis task.

These findings reveal three key theoretical insights: First, ICL's capability boundaries exhibit clear patterns—it excels in pattern recognition tasks but struggles with tasks requiring precise knowledge or complex reasoning. Second, the alignment between task characteristics and ICL's strengths is a critical factor, with tasks involving subjectivity and deep semantic understanding being more suitable for ICL. Third, ICL and traditional methods should form a complementary rather than substitutive relationship, and future research should develop systematic task-method matching frameworks to optimize performance across different scenarios. These conclusions provide valuable insights for understanding ICL's working mechanisms and advancing its practical applications.

## 3.2 COMPARISON OF ICL ACCURACY ACROSS DIFFERENT LARGE LANGUAGE MODELS

In this experiment, we aim to investigate how the size and type of large language models (LLMs) influence the performance of in-context learning (ICL) on a standard text classification task. We select three different LLMs with varying model sizes and architectures: Qwen/Qwen2-1.5B-Instruct, Qwen/Qwen2-7B-Instruct, and deepseek-chat. The target task is the AG News dataset, a widely used benchmark for multi-class news categorization.

The primary goal is to quantitatively compare the ICL accuracy of these models under the same experimental settings, providing insights into whether larger model sizes or different architectures translate into improved ICL performance.

### 3.2.1 EXPERIMENT PROCEDURE

We use the AG News dataset, which consists of text samples labeled into four news categories. The dataset was split into training and testing subsets with an 80:20 ratio.

The experiment follows these steps:

1) Dataset Preparation:

We loaded the AG News dataset using the TextDataLoader class. A total of 20 training samples and 20 testing samples were selected for ICL, simulating a few-shot learning scenario.

2) Model Setup:

Three LLMs were selected: Qwen/Qwen2-1.5B-Instruct (1.5 billion parameters) , Qwen/Qwen2-7B-Instruct (7 billion parameters), deepseek-chat.

Qwen models were accessed through the SiliconFlow API, while deepseek-chat was accessed via the DeepSeek API.

3) In-Context Learning:

For each model, a prompt was constructed that included 20 labeled training samples followed by an unlabeled test input. The model was asked to predict the output label based on the context provided.

4) Evaluation: Each model's prediction accuracy was computed based on its performance on the 20 test samples. Results were recorded in a CSV file and visualized as a bar chart for easy comparison.

### 3.2.2 Quantitative Results

The experiment results are summarized below:

| Model | Accuracy |
|---|---|
| Qwen/Qwen2-1.5B-Instruct | 0.40 |
| Qwen/Qwen2-7B-Instruct | 0.65 |
| deepseek-chat | 0.75 |

Table 3: ICL Accuracy of Different LLMs on AG News Dataset

The bar chart in Figure 1 (see below) provides a visual comparison of model accuracies:

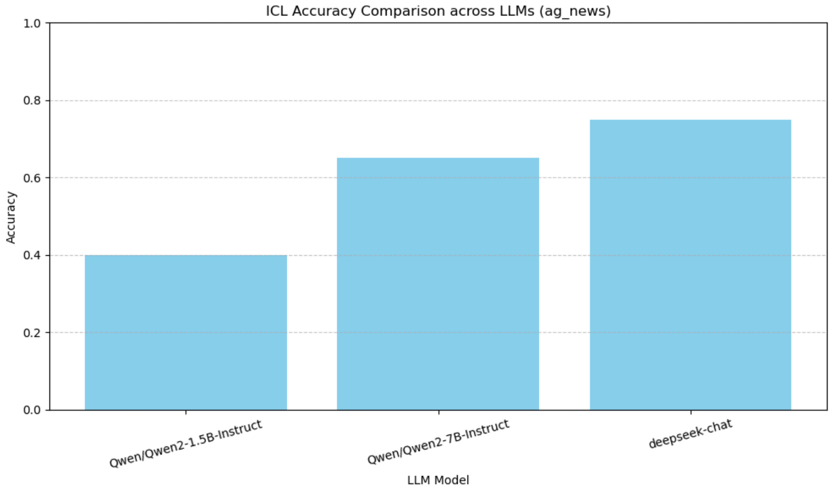

Figure 7: ICL Accuracy Comparison across LLMs (AG News Dataset)

### 3.2.3 Results Analysis

The results reveal several important observations:

1) Model Size Impact:
The Qwen2-7B-Instruct (7B parameters) significantly outperforms the Qwen2-1.5B-Instruct (1.5B

parameters), achieving 0.65 vs. 0.40 in accuracy. This demonstrates a clear benefit of increased model size in ICL performance, aligning with the general understanding that larger models can capture more complex patterns and generalize better in few-shot settings.

2) Model Architecture and Instruction-Following:
Interestingly, deepseek-chat outperformed both Qwen models, achieving 0.75 accuracy despite not having the largest parameter size. This suggests that architecture differences and instruction-following capabilities play a critical role in ICL effectiveness, not just model size alone.

3) Performance Plateau:
While moving from 1.5B to 7B showed a large improvement (+25 points), the difference between Qwen2-7B and deepseek-chat is smaller (+10 points), hinting that beyond a certain size, architectural optimizations and training strategies may matter more than sheer size.

4) Instruction-Following Quality:
The performance of deepseek-chat may also reflect stronger alignment and robustness in following ICL prompts, indicating that prompt design and model fine-tuning are crucial factors influencing ICL success.

### 3.2.4 Conclusion

This experiment highlights the following key insights:

• Increasing the parameter size of LLMs generally improves ICL accuracy, as shown by the jump from 0.40 (1.5B) to 0.65 (7B).

• Model architecture and prompt-following capabilities can outweigh raw parameter size, as evidenced by deepseek-chat’ s superior performance (0.75 accuracy).

• For practical applications, both model size and architecture should be considered when selecting LLMs for ICL tasks.

• Further exploration is warranted to test the generality of these findings across different datasets and prompt settings.

These results support the hypothesis that while size matters, model quality, prompt tuning, and underlying architecture are equally (if not more) important in maximizing the performance of in-context learning.

## 3.3 impact of prompt length on ICL accuracy

This part aims to delve into the impact of prompt length on the classification performance of ICL to discuss whether longer prompts invariably improve performance.

### 3.3.1 Experiment Procedure

The study investigated the impact of prompt length on in-context learning (ICL) accuracy using the AG News dataset and the DeepSeek Chat model. The experimental design systematically varied the number of training examples (5 to 20 samples with a step size of 5) while maintaining a fixed test set of 20 samples, with an 80/20 train-test split ratio. The implementation involved loading the model API, preprocessing the dataset, and executing ICL across different prompt lengths. The experimental loop iterated through different training sample sizes, measuring classification accuracy for each condition. Results were visualized through accuracy curves to examine the relationship between prompt length and model performance.

### 3.3.2 Quantitative Results

The experimental data revealed a non-monotonic relationship between prompt length and classification accuracy. The model achieved its peak performance (75% accuracy) with 10 training samples, representing a 25% improvement over the 5-sample condition (50% accuracy). However, increasing the prompt length beyond this point led to diminishing returns, with accuracy declining to 60% at 15 samples and 55% at 20 samples. This pattern suggests the existence of an optimal prompt length for this specific classification task. The complete numerical results were preserved in structured format, with the 10-sample condition demonstrating superior performance compared to both shorter and longer prompt configurations.

| Length of Few-Shot | Accuracy |
|---|---|
| 5 | 50% |
| 10 | 75% |
| 15 | 60% |
| 20 | 55% |

Table 4: ICL Accuracy of Different Few-Shot Length on AG News Dataset

The bar chart in Figure 1 (see below) provides a visual comparison of model accuracies:

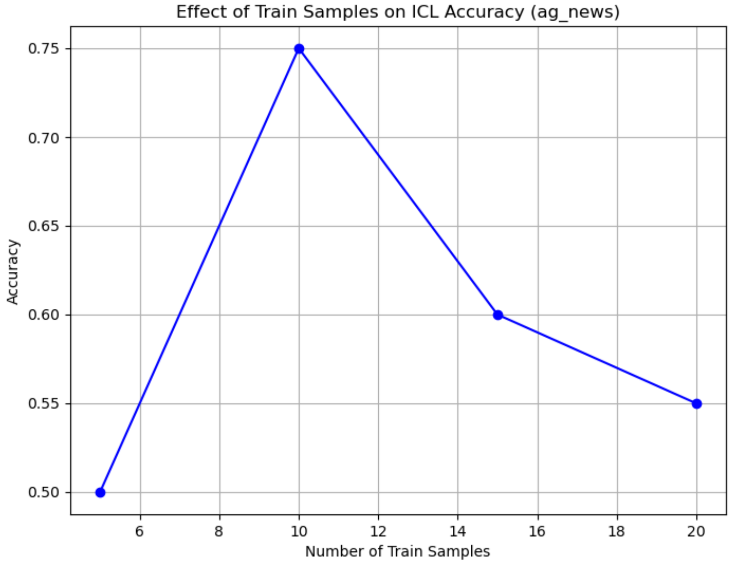

Figure 8: ICL Accuracy Comparison across Few-Shot Length (AG News Dataset)

### 3.3.3 Results Analysis

The observed performance trajectory indicates that prompt length exerts a significant but non-linear influence on ICL effectiveness. The initial accuracy improvement from 5 to 10 samples suggests that moderate prompt expansion enhances the model's ability to discern classification patterns. Subsequent performance degradation at higher sample counts may stem from cognitive overload mechanisms, where excessive contextual information dilutes the model's attention to discriminative features. The inverse U-shaped accuracy curve implies that AG News classification represents an intermediate-complexity task where excessive demonstrations introduce noise rather than beneficial signal. This phenomenon aligns with emerging literature on the trade-off between contextual information quantity and processing efficiency in transformer-based architectures.

### 3.3.4 Conclusion

This empirical investigation demonstrates that prompt length optimization constitutes a critical factor in ICL implementation, with the 10-sample condition emerging as the optimal configuration for AG News classification. The findings challenge the conventional assumption that longer prompts invariably improve performance, instead highlighting the importance of balanced context provision. Future research should explore the interaction between prompt length and content quality across diverse task domains, while practitioners should adopt a systematic approach to prompt length tuning.

The results underscore the need for task-specific calibration of ICL parameters to achieve optimal model performance without unnecessary computational overhead or information overload.

## 4 CONCLUSION

This study systematically compares the performance of in-context learning (ICL) with traditional machine learning methods across different types of tasks, revealing the applicability boundaries and optimization pathways of ICL technology. Experimental results demonstrate that ICL exhibits significant advantages in text-based tasks requiring semantic understanding and pattern recognition (e.g., sentiment analysis), achieving accuracy rates as high as 95%, far surpassing traditional methods. However, in structured data processing and tasks requiring precise reasoning, traditional machine learning methods still maintain a clear lead. This finding confirms the importance of aligning task characteristics with model capabilities, suggesting that researchers should select appropriate technical approaches based on specific task requirements.

Further investigation reveals a notable complementary relationship between ICL and traditional machine learning methods. In complex language tasks such as text comprehension and semantic analysis, ICL can capture deep features that are difficult for traditional methods to identify, while for classification and prediction tasks involving structured data, traditional methods demonstrate higher accuracy and stability. This complementary nature indicates that the future direction of machine learning should not involve simple technological substitution but rather the establishment of a more systematic task-method matching framework to maximize the advantages of different techniques.

Regarding model optimization, the study uncovers multiple factors influencing ICL performance. While increasing model size can improve performance, such improvements exhibit diminishing marginal returns, and the final outcomes are significantly moderated by model architecture and instruction-following capabilities. Notably, the quality of prompt engineering impacts ICL performance no less than model size itself, with the optimal prompt length varying by task type. Excessive increases in prompt content may even degrade performance. These findings provide critical guidance for the practical application of ICL: model selection should comprehensively consider scale, architecture, and prompt design rather than solely pursuing parameter expansion.