In-Context Learning with Large Language Models: A Survey

Abstract

In-Context Learning (ICL) is a paradigm that empowers large language models (LLMs) to perform tasks by leveraging contextual information from input data without explicit retraining. This survey examines the concept and significance of ICL, highlighting its transformative potential in enhancing the adaptability and cognitive capabilities of LLMs. The survey addresses key motivations, including the variability in ICL performance due to input example selection and the need for robust evaluation frameworks. It explores theoretical foundations, pre-training, and prompting techniques that underpin ICL, emphasizing the importance of example selection and retrieval strategies. Empirical studies and benchmarks are reviewed to assess ICL effectiveness across diverse tasks, revealing the robustness of LLMs despite label variations. The survey also delves into the challenges and limitations of ICL, such as model robustness, sensitivity to input variations, and computational costs. Future directions are proposed, including expanding ICL applications, improving robustness, developing novel evaluation methods, and integrating external knowledge and multimodal inputs. By synthesizing these insights, the survey aims to advance the understanding and application of ICL, contributing significantly to the field of natural language processing.

1 Introduction

1.1 Concept and Significance of In-Context Learning

In-Context Learning (ICL) empowers large language models (LLMs) to execute tasks by utilizing contextual information from input data without necessitating explicit retraining. This emergent capability significantly enhances performance in downstream tasks, as LLMs learn to accomplish target tasks based on provided examples within the input context [1]. ICL involves conditioning on input-label pairs (demonstrations) for predictions, which is vital for comprehending LLMs' capabilities and behaviors [2].

The significance of ICL lies in its enhancement of predictive accuracy for various downstream tasks through the integration of a limited number of relevant input-label examples into the model's prompt. This non-parametric learning strategy, akin to k-nearest neighbors, informs the model's predictions based on contextual relationships among nearby examples. By optimizing the selection of few-shot demonstrations based on task-specific relevance, ICL improves performance and enables models to adapt to new tasks without parameter adjustments, revealing the intricate dynamics of label information utilization during inference [3, 4, 5, 6, 7]. This approach balances the utilization of context from input data with the pretrained knowledge embedded in the model, representing a transformative method that enhances the adaptability and cognitive capabilities of LLMs, thus positioning it as a cornerstone of contemporary NLP research and applications.

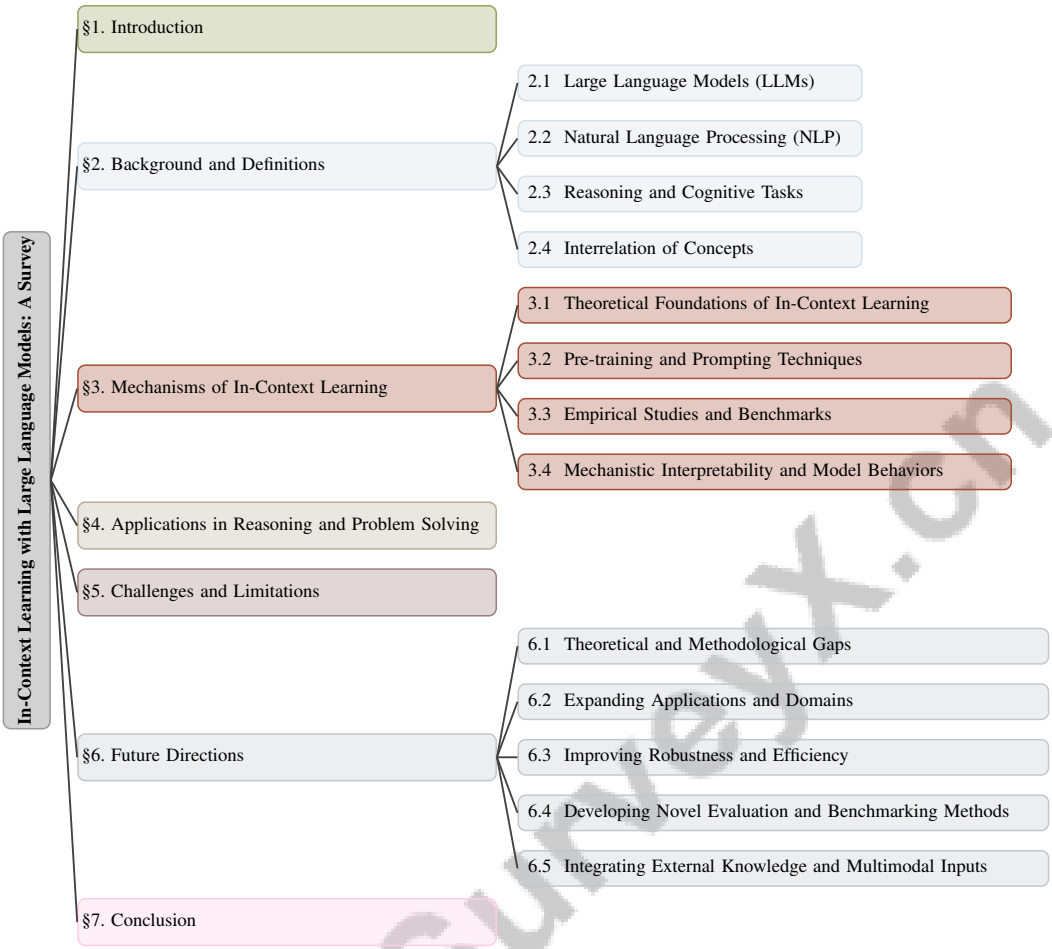


Figure 1: chapter structure

1.2 Motivation Behind the Survey

This survey on ICL within LLMs is motivated by several pressing challenges and gaps in existing research. A significant concern is the high variability in ICL performance, which is sensitive to the selection and order of input examples, leading to inconsistent evaluations and potentially misleading results [1]. The survey aims to address these inconsistencies by providing a robust framework for evaluating ICL capabilities.

Another fundamental challenge is understanding how ICL manifests in LLMs trained on unstructured data, which often does not align with typical ICL prompts. This survey investigates the mechanisms of ICL across various data environments, crucial for enhancing the effectiveness and adaptability of LLMs by identifying how task-specific relevance can optimize the selection of few-shot examples, thereby broadening ICL applications in diverse NLP contexts [6, 4, 8]. The current scope of ICL is largely confined to natural language tasks and basic instruction-following, highlighting the need for benchmarks that assess ICL across a wider array of tasks and facilitate continuous interaction with real-world environments.

The lack of comprehensive understanding regarding ICL mechanisms and their role in enhancing predictive accuracy further motivates this survey. By addressing these knowledge gaps, the survey aims to deepen theoretical comprehension of ICL processes. Additionally, the strategic selection of in-context examples is crucial, as performance can vary significantly depending on example selection. This survey will explore advanced strategies for optimizing example selection in ICL, focusing on methods that enhance task performance by ensuring selected examples effectively represent salient aspects of test instances, thus reducing redundancy and improving overall quality [9, 10].

Understanding the influence of context scaling on ICL capabilities is essential for enhancing LLM performance. Recent studies indicate that as the number of contextual examples increases, the model's ability to leverage these examples improves, leading to more effective predictions. This underscores the importance of task-specific relevance in few-shot example selection, which can significantly impact ICL outcomes and the overall efficacy of LLMs in NLP tasks [3, 11, 8, 4, 6]. The survey will examine how context scaling affects learning outcomes and model efficacy. Additionally, bridging the gap between ICL and Instruction Tuning (IT) in LLMs, which have been studied predominantly in isolation, is essential. By synthesizing insights from both domains, the survey aims to provide a more holistic understanding of LLM capabilities.

Furthermore, the survey intends to explore foundational principles, diverse applications, and intricate training processes of LLMs, addressing existing knowledge gaps and offering insights into their mechanics and ethical considerations. The limitations of current knowledge-based question answering (KBQA) methods, which often rely on complex training techniques and architectures requiring extensive labeled datasets for effective parameter convergence, underscore the urgency for this survey. Recent advancements, particularly the introduction of ICL capabilities in LLMs, present a promising alternative by enabling simpler, training-free approaches to semantic parsing. However, challenges persist, such as high error rates in generating logical forms due to insufficient exposure during pre-training, necessitating innovative solutions and comprehensive examinations of these evolving methodologies [12, 13, 14, 15, 16].

This survey aims to deepen understanding of ICL in LMs by investigating how various aspects of demonstration examples influence task performance. It seeks to elucidate mechanisms behind LMs' ability to learn from limited input-output pairs while examining factors such as label space representation, input distribution, and sequence formatting. Additionally, the survey addresses variability in demonstration effectiveness across different reasoning questions and proposes novel methods for prompt engineering to optimize demonstration selection, ultimately contributing to improved performance across various NLP tasks [2, 8, 17, 18]. By synthesizing research on ICL, the survey aims to advance the understanding and application of ICL, making a substantial contribution to the broader field of natural language processing.

1.3 Survey Objectives

The survey aims to provide a comprehensive understanding of ICL in LLMs, focusing on multiple facets that contribute to advancing this field. A primary objective is to evaluate multilingual ICL abilities of LLMs across diverse tasks and languages, specifically examining the impact of demonstrations [19]. The survey also proposes an adaptive in-context learning (AICL) method that dynamically selects the number of examples based on the specific requirements of the instance being processed, thereby enhancing ICL adaptability [6].

Additionally, the survey seeks to summarize ICL techniques, discuss training strategies, prompt design, and applications, while addressing inherent challenges [20]. It introduces a consistent evaluation framework for comparing ICL improvements across various models and datasets, focusing on the impact of prompt templates [21].

The survey examines challenges and advancements in LLMs, particularly regarding training methodologies and ethical implications, providing insights into the broader landscape of LLM research [22]. Another critical objective is to explore the relationship between ICL and instruction tuning, providing empirical evidence suggesting ICL operates similarly to instruction tuning without updating model parameters [23].

By addressing key objectives, this survey aims to significantly advance the field of natural language processing (NLP) by providing a comprehensive overview of ICL in LLMs, including its formal definition, advanced techniques such as training and prompt design strategies, various application scenarios like data engineering and knowledge updating, and the challenges that remain in this emerging paradigm [6, 8, 20].

1.4 Structure of the Survey

This survey on ICL with LLMs is structured to provide a comprehensive exploration of the topic by organizing research findings into distinct yet interrelated fields. The survey begins with an

Introduction that outlines the concept and significance of ICL, the motivation behind conducting the survey, and the objectives it aims to achieve. This is followed by a detailed **Background and Definitions** section that elucidates key concepts such as LLMs, natural language processing, reasoning, and cognitive tasks, highlighting their interrelations and foundational importance to the survey.

In the **Mechanisms of In-Context Learning** section, the survey delves into the theoretical and empirical underpinnings of ICL, examining how LLMs utilize context from input data to perform tasks without explicit retraining. This section is organized into subsections covering theoretical foundations, pre-training and prompting techniques, empirical studies and benchmarks, and mechanistic interpretability.

The next section, **Applications in Reasoning and Problem Solving**, investigates the practical applications of ICL in LLMs, focusing on reasoning tasks, logical problems, and mathematical problem-solving. This section highlights specific examples and case studies, discussing both strengths and limitations observed in these applications.

Challenges and Limitations are addressed in a subsequent section, where the survey identifies and discusses issues such as model robustness, sensitivity to input variations, scalability, computational costs, and benchmarking challenges. This section provides a critical assessment of the current state of ICL research and its limitations.

The survey concludes with a section on **Future Directions**, outlining potential research avenues and advancements in the field of ICL with LLMs. This includes discussions on theoretical and methodological gaps, expanding applications and domains, improving robustness and efficiency, developing novel evaluation and benchmarking methods, and integrating external knowledge and multimodal inputs.

By organizing the survey into these comprehensive sections, the paper aims to provide a holistic understanding of ICL, encompassing its theoretical foundations, practical applications, and future prospects [8].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Large Language Models (LLMs)

Large Language Models (LLMs) are central to advancements in natural language processing (NLP), exemplified by models like GPT-3, which leverage the Transformer architecture to handle extensive datasets and excel in diverse NLP tasks [22]. A notable feature of LLMs is in-context learning (ICL), enabling adaptation to new tasks with minimal examples, thus reducing retraining needs [24]. LLMs' efficacy in ICL is rooted in their ability to utilize co-occurrence and structural patterns in data, enhancing performance in tasks such as knowledge-based question answering through code-style ICL methods [15]. Despite their adaptability, LLMs' effectiveness in ICL is contingent on the strategic selection and ordering of input examples, as demonstrated in various datasets [1]. Their application extends to fields like cancer pathology image classification and language translation, though challenges remain in multimodal contexts due to limited training on single-image datasets [22]. Ongoing research continues to refine our understanding of LLM capabilities, focusing on principles, training methodologies, and ethical considerations [22].

2.2 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a crucial AI domain that facilitates interactions between computers and humans through natural language. LLMs perform a range of tasks, from language understanding and generation to complex functions like multilingual translation and sentiment analysis [25, 26, 27, 14, 28]. The integration of NLP with LLMs has revolutionized task execution, allowing for remarkable accuracy in language tasks by learning linguistic patterns from vast text data [29]. ICL enhances LLM capabilities by leveraging contextual information from input data, minimizing retraining needs. However, challenges persist in tasks requiring deep understanding, such as Math Word Problems (MWPs) [30]. Demonstration selection in ICL is critical, as it significantly impacts performance across tasks, as evidenced by studies using datasets like MRPC, QNLI, and SST2 [31]. NLP serves as the foundation for LLM and ICL advancements, fostering methodologies

that enhance machine understanding and generation of human language, including task-specific relevance definitions and retrieval-based approaches [25, 8, 4, 6, 32].

2.3 Reasoning and Cognitive Tasks

LLMs exhibit significant capabilities in reasoning and cognitive tasks, essential for effective in-context learning (ICL). These tasks include deductive, inductive, and abductive reasoning, vital for performance across learning scenarios. Despite their proficiency, LLMs face challenges in tasks like knowledge-based question answering, where high format error rates hinder logical form generation [15]. ICL effectiveness is heavily influenced by prompt template design, as variations can lead to substantial performance differences. Research highlights the role of model architecture in ICL capabilities and the benefits of strategically selecting semantically similar demonstrations [6, 4, 33, 3]. Benchmark evaluations across models, tasks, and languages further emphasize the diverse applicability and inherent challenges in cognitive tasks. Addressing these challenges involves optimizing prompt design and enhancing ICL techniques, which are crucial for improving LLM cognitive capabilities and expanding their application across domains [32, 6, 4, 8].

2.4 Interrelation of Concepts

The interplay between in-context learning (ICL), large language models (LLMs), and natural language processing (NLP) forms a foundational framework for understanding LLM capabilities and limitations. ICL's sensitivity to demonstration choice, format, and order significantly impacts LLM performance [34]. This necessitates strategic selection and arrangement of examples to enhance knowledge transfer and cognitive task execution [35]. LLMs' ICL capabilities are linked to their architectural design, particularly attention mechanisms that facilitate context maintenance and retrieval, mirroring human memory processes [5]. Challenges persist in understanding the correlation between ICL mechanisms and textual features, especially concerning noisy labels [36]. Evaluating LLMs' contextual understanding is complicated by knowledge hijacking, where models overly rely on in-context information, leading to incorrect predictions [37]. Benchmarks assessing LLMs' ability to resolve linguistic features beyond individual sentences are crucial for evaluating their comprehension capabilities [38]. In multimodal contexts, optimizing in-context example retrieval requires careful consideration, as many multimodal LLMs are trained on single-image datasets [39]. The survey explores various methodologies for retrieval models, focusing on retrieval-based approaches and their implications in ICL [40]. These interconnected concepts highlight the need for refining retrieval and demonstration strategies to improve model performance, suggesting task-specific retrieval of semantically similar examples can significantly enhance ICL effectiveness [8, 33, 4, 6, 7]. By elucidating these interconnections, the survey provides a comprehensive perspective on how these concepts collectively contribute to LLM development and optimization, advancing the field of NLP.

In recent years, the exploration of in-context learning mechanisms within large language models has garnered significant attention in the field of artificial intelligence. To elucidate this complex topic, Figure 2 provides a comprehensive illustration of the hierarchical structure underpinning these mechanisms. This figure meticulously details the theoretical foundations, pre-training and prompting techniques, as well as empirical studies and the crucial aspect of mechanistic interpretability. Notably, it highlights the integration of transformer architecture and the optimization of example selection, while also addressing the impact of multimodal and rectification approaches. Furthermore, the figure showcases the role of adaptive methods, reinforcement learning, and calibration in enhancing model performance. Empirical insights into demonstration selection and model robustness are also presented, enriching our understanding of the intricate dynamics involved in these advanced learning systems.

3 Mechanisms of In-Context Learning

3.1 Theoretical Foundations of In-Context Learning

In-Context Learning (ICL) in Large Language Models (LLMs) is grounded in the Transformer architecture, particularly its attention mechanisms, which enhance task recognition by focusing on relevant input segments [22]. The induction head mechanism balances in-context and global knowledge, integrating new contextual data with pre-existing information [37]. Recent studies

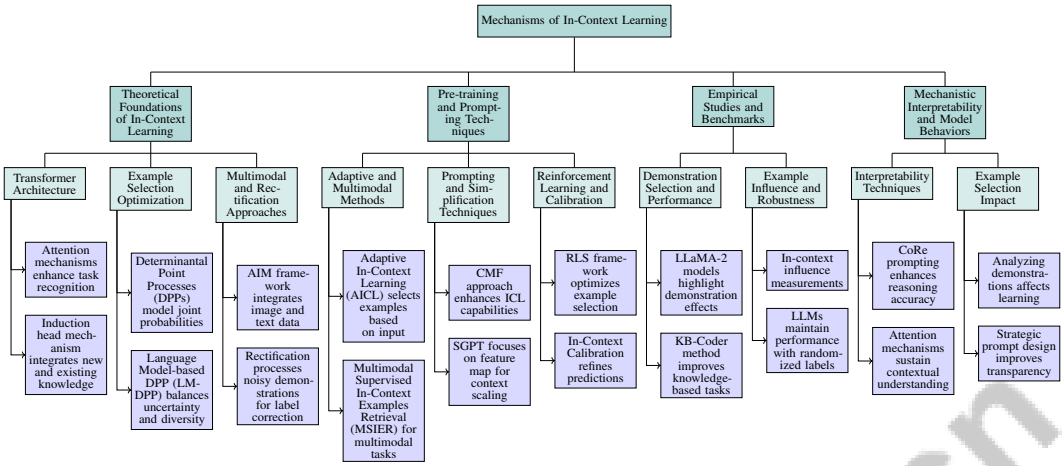


Figure 2: This figure illustrates the hierarchical structure of in-context learning mechanisms in large language models, detailing theoretical foundations, pre-training and prompting techniques, empirical studies, and mechanistic interpretability. It highlights the integration of transformer architecture, optimization of example selection, and the impact of multimodal and rectification approaches. It also showcases the role of adaptive methods, reinforcement learning, and calibration in enhancing model performance, along with empirical insights into demonstration selection and model robustness.

question the necessity of ground truth labels in demonstrations, highlighting the importance of selecting influential examples, both positive and negative, to optimize ICL performance [2, 1].

The theoretical framework of ICL is enriched by information retrieval concepts like Determinantal Point Processes (DPPs), which optimize example selection by modeling joint probabilities [35]. The Language Model-based DPP (LM-DPP) method further refines this by balancing uncertainty and diversity, ensuring informative example selection from unlabeled data [41]. In multimodal contexts, frameworks like AIM integrate image information into the latent space of textual labels, enhancing Multimodal Large Language Models (MLLMs) [39].

A notable theoretical innovation is the rectification approach, which processes noisy demonstrations simultaneously, allowing for effective label correction [36]. This method is crucial for maintaining LLM accuracy in diverse contexts. The theoretical landscape of ICL is shaped by architectural innovations, retrieval strategies, and adaptive learning techniques, enhancing our understanding of how LLMs leverage task-specific examples to generate predictions. This comprehension elucidates ICL mechanisms and highlights its potential for optimizing example selection through supervised ranking models, paving the way for more sophisticated LLMs across various natural language processing tasks [6, 8, 32].

As illustrated in Figure 3, the theoretical foundations of ICL encompass key components such as the Transformer architecture, example selection strategies, and innovative approaches like rectification and multimodal integration. This figure serves to visually reinforce the intricate relationships among these elements, providing a comprehensive overview of the ICL framework.

3.2 Pre-training and Prompting Techniques

Method Name	Method Optimization	Task Adaptability	Model Structuring
ICL-WT[42]	Prompt Tuning	Unseen Tasks	Input Data
Rect[36]	Useful Demonstrations	Various Tasks	Corrected Labels
SGPT[11]	Feature Map	Unseen Tasks	Identity Matrices
PICL[43]	Enhance Icl Capabilities	Diverse Intrinsic Tasks	Retrieving Paragraphs Sharing
ICS[33]	Multiple Prompt Inputs	Various Domains	Prompt Input Configurations

Table 1: Comparison of various pre-training and prompting methods for enhancing in-context learning (ICL) capabilities in large language models (LLMs). The table outlines the optimization techniques, task adaptability, and model structuring strategies employed by each method, highlighting their unique contributions to improving ICL performance.

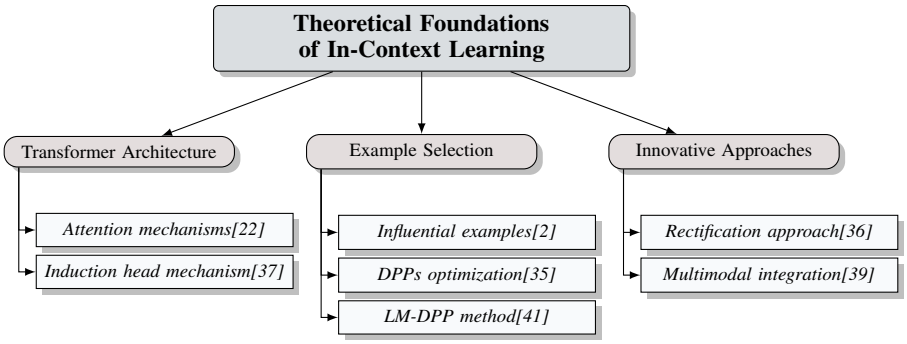


Figure 3: This figure illustrates the theoretical foundations of In-Context Learning (ICL), highlighting key components such as the Transformer architecture, example selection strategies, and innovative approaches like rectification and multimodal integration.

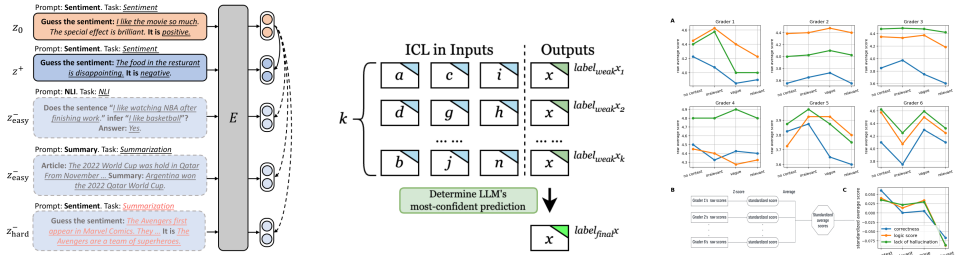
Pre-training and prompting techniques are pivotal for enhancing ICL capabilities in LLMs, focusing on optimizing example selection and organization. These techniques significantly influence ICL effectiveness [20]. The Adaptive In-Context Learning (AICL) method exemplifies this by selecting examples based on input characteristics, improving task adaptability. The Multimodal Supervised In-Context Examples Retrieval (MSIER) method refines example selection for multimodal tasks, enhancing performance. Relative positional encoding (RPE) further enhances transformers by addressing knowledge hijacking [37].

Prompting techniques refine ICL by structuring inputs to align with model expectations. The CMF approach gathers outputs across datasets, significantly enhancing ICL capabilities. The ICL-WT method involves a warmup training phase with demonstrations of speech-label pairs, illustrating innovative audio data utilization for ICL [42]. Rectification methods enhance ICL performance by correcting noisy labels through generative modeling [36]. SGPT simplifies transformer models by focusing on a feature map that enhances context scaling [11].

The policy-based reinforcement learning framework (RLS) optimizes example selection in incomplete utterance rewriting (IUR) tasks by leveraging feedback from LLMs. This method enhances selection through a language model selector that encodes candidate examples into dense representations, demonstrating LLM adaptability by utilizing outputs to compute rewards and policy gradients for continuous improvement. RLS significantly outperforms traditional example selection methods and supervised fine-tuning models in few-shot settings, highlighting the importance of balancing example abundance and similarity for effective ICL [22, 9, 44, 33, 45]. The In-Context Calibration method refines LLM predictions by recalibrating the influence of semantic priors, enhancing accuracy and reliability.

These pre-training and prompting techniques bolster LLM robustness and adaptability in ICL, enabling more accurate and contextually aware outputs across various domains. By refining these strategies, LLMs advance the field of natural language processing, demonstrating improved efficiency in complex tasks [46]. Table 1 provides a comprehensive overview of pre-training and prompting methods that are instrumental in advancing in-context learning (ICL) within large language models (LLMs), detailing their optimization strategies, adaptability to different tasks, and model structuring techniques.

As illustrated in Figure 4, exploring the mechanisms underlying ICL, particularly through pre-training and prompting techniques, is crucial for advancing deep learning models' capabilities. The figures exemplify various facets of these techniques: the first figure illustrates a deep learning model adept at multitask learning, showcasing its versatility in processing diverse data types, including text and images. The second figure presents a schematic of a machine learning model that processes a sequence of inputs to produce a unified output, emphasizing the structural components of input and output layers. The third figure provides insights into the evaluation of logical reasoning tasks under varying contextual conditions, underscoring context's impact on performance through different grader assessments. Collectively, these examples highlight the potential of pre-training and prompting techniques in enhancing model performance across diverse scenarios [43, 33, 47].



(a) Deep Learning Model Demonstrates Multitask Learning Ability[43]

(b) ICL in Inputs and Outputs[33]

(c) Grading of Logical Reasoning Tasks with Different Contexts[47]

Figure 4: Examples of Pre-training and Prompting Techniques

3.3 Empirical Studies and Benchmarks

Benchmark	Size	Domain	Task Format	Metric
ICL-Bench[48]	10,000	Linear Regression	Regression	Mean Squared Error
ICL-Benchmark[49]	1,057	Text Processing	Functional And Behavioral Tasks	Accuracy, Generation Entropy
LOFT[50]	1,000,000	Information Retrieval	Multi-turn Question Answering	Recall@1, Exact Match
SelfAware[51]	3,369	Question Answering	Unanswerable Question Identification	F1 score
CSS-ICL-IT[52]	30,000	Computational Social Science	Text Classification	Accuracy, F1-score
RBK[53]	12,800	Text Classification	Binary Classification	RBK, weighted F1
MambaFormer[54]	1,280	Regression	In-context Learning	Accuracy, F1-score
ICL-Benchmark[55]	18	Event Detection	Named Entity Recognition	F1-score

Table 2: This table presents a comprehensive overview of key benchmarks employed in the evaluation of In-Context Learning (ICL) within Large Language Models (LLMs). It details the benchmark names, dataset sizes, application domains, task formats, and performance metrics, providing a structured insight into the diverse methodologies and evaluation criteria used in contemporary research.

Empirical studies and benchmarks are essential for evaluating ICL's effectiveness within LLMs, offering insights into their adaptability and performance across tasks. As illustrated in Figure 5, this figure highlights the key empirical studies and benchmarks in In-Context Learning (ICL) for Large Language Models (LLMs), emphasizing the effects of demonstration similarity, specialized techniques like KB-Coder, and influence-based example selection methods. Table 2 provides a detailed overview of the various benchmarks used in empirical studies to assess the effectiveness of In-Context Learning (ICL) strategies in Large Language Models (LLMs). Experiments with LLaMA-2 models on datasets like SST2 for sentiment analysis and ENCS for English-Czech translation demonstrate the significant effects of demonstration similarity and quantity on model performance [23]. These findings emphasize strategic demonstration selection and organization to optimize ICL outcomes.

Investigations using datasets such as WebQSP, GraphQ, and GrailQA assess the KB-Coder method, revealing substantial improvements over baseline methods in knowledge-based question answering tasks [15]. These studies underscore the potential of specialized ICL techniques to enhance model capabilities within specific domains.

The influence of in-context example selection has been explored through methods that calculate in-context influences, allowing for the measurement and ranking of examples' impact on ICL performance [1]. This nuanced understanding of example contributions facilitates informed decisions in example selection.

Empirical evidence suggests that performance drops minimally when labels are randomized, indicating that LLMs can recover expected input-label correspondences without direct mappings [2]. This finding underscores LLMs' robustness in maintaining performance despite variations in label configurations.

Collectively, these empirical studies and benchmarks provide valuable insights into the strengths and limitations of ICL within LLMs, informing future research directions and enhancing LLM

adaptability and effectiveness across diverse applications. They highlight advanced techniques such as task-specific relevance in example selection, the potential of multiple prompt constructions, and the benefits of retrieval-based demonstration strategies [6, 4, 33, 8]. Leveraging these findings allows researchers to refine ICL strategies, ultimately advancing the field of natural language processing.

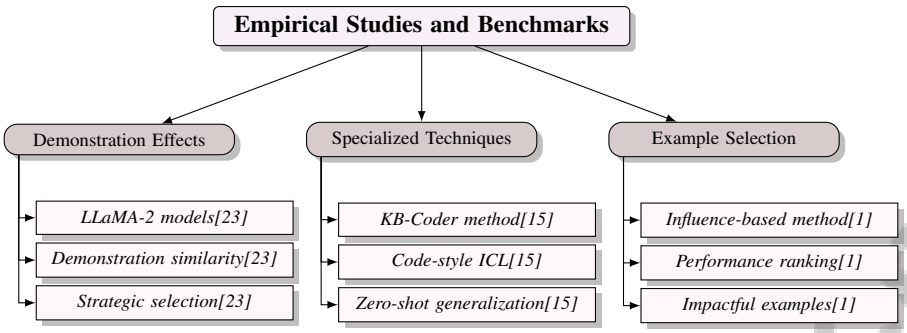


Figure 5: This figure illustrates the key empirical studies and benchmarks in In-Context Learning (ICL) for Large Language Models (LLMs), highlighting the effects of demonstration similarity, specialized techniques like KB-Coder, and influence-based example selection methods.

3.4 Mechanistic Interpretability and Model Behaviors

Mechanistic interpretability in LLMs is crucial for understanding how these models process input data and make predictions, particularly regarding ICL. The interpretability of mechanisms within LLMs directly influences their behavior, enabling researchers to decipher the processes that drive model decisions and outputs, which is pivotal for optimizing performance and ensuring reliable outputs across various tasks [56].

One approach to enhancing interpretability is through prompting techniques like CoRe, which guides LLMs to identify essential conditions of a problem before applying zero-shot reasoning. This method emphasizes condition recognition, allowing models to focus on critical task aspects and improve reasoning accuracy. By structuring problem-solving processes, CoRe enhances transparency in model reasoning pathways, providing insights into how LLMs navigate complex tasks without explicit retraining [56].

LLM behavior in ICL is influenced by architectural features such as attention mechanisms, which facilitate selective focus on relevant input segments. These mechanisms enhance the model's capacity to sustain contextual understanding and effectively retrieve relevant information by utilizing ICL techniques that draw on localized examples and extensive training knowledge, thereby mimicking human memory processes [6, 37, 18]. Understanding these behaviors is essential for optimizing LLM adaptability and performance across diverse applications.

Furthermore, interpretability can be enhanced by investigating the impact of example selection and organization on ICL outcomes. Analyzing how different demonstrations affect learning processes provides insights into the factors driving model behavior and informs the strategic design of prompts and examples. This approach enhances the transparency of LLMs by clarifying their decision-making processes and informs the development of more effective ICL techniques by leveraging task-specific relevance criteria to optimally select few-shot examples, ultimately leading to improved performance in natural language processing tasks [25, 33, 4, 6, 32].

4 Applications in Reasoning and Problem Solving

The integration of large language models (LLMs) into reasoning and problem-solving, particularly through in-context learning (ICL), marks a significant advancement. This section explores methodologies enhancing LLM efficacy in complex reasoning tasks, focusing on adaptive engagement with mathematical and logical challenges, laying the groundwork for understanding their broader implications.

4.1 Mathematical and Logical Problem Solving

Utilizing LLMs for mathematical and logical problem-solving via ICL signifies a pivotal shift in natural language processing, enabling adaptive engagement with intricate reasoning tasks. The strategic selection of demonstrations, as seen in curriculum demonstration selection (CDS), enhances performance on datasets like MATH by tailoring examples based on complexity [2]. Iterative demonstration selection (IDS) frameworks, such as Deep-Thinking, show improved handling of complex scenarios, underscoring LLMs' potential in intricate problem-solving [38].

Dynamic uncertainty ranking methods further refine LLM performance on long-tail questions, surpassing baseline methods and enhancing prediction accuracy [36]. In logical reasoning, integrating programming languages like Python with structured domain descriptions advances ICL for semantic parsing, improving accuracy and compositional generalization [57, 58, 59, 60]. Despite these advancements, challenges persist in fully understanding LLMs' logical reasoning abilities. Analyses suggest that LLMs' impressive performance may rely more on ICL than genuine understanding, with context changes affecting outputs [61, 62].

LLMs also utilize reasoning capabilities through natural language explanations (NLEs), improving decision-making in complex tasks. Experiments across tasks like Theory of Mind and symbolic reasoning using datasets such as SOCIALIQA and GSM8K illustrate LLMs' capabilities in both associative and complex reasoning. Chain-of-Thought prompt learning further emphasizes LLMs' potential for multi-step reasoning, raising questions about their engagement in cognitive processes associated with human intelligence [63, 64]. The demonstration notebook approach enhances performance by dynamically selecting question-specific demonstrations, beneficial for mathematical and logical problem-solving.

Research into LLMs' capabilities in mathematical and logical problem-solving through ICL underscores context length and reasoning structures' importance while introducing innovative methods like Reasoning Graph-enhanced Exemplar Retrieval and Inferential Exclusion Prompting. These developments are crucial for improving LLM adaptability and effectiveness in complex reasoning tasks, addressing limitations, and enhancing generalizability across benchmarks [65, 60, 66, 56, 6]. These methodologies illustrate LLMs' transformative potential in advancing natural language processing and cognitive computing.

4.2 Enhancing In-Context Learning Capabilities

Enhancing LLM capabilities in reasoning and problem-solving through ICL requires advanced methods for optimizing demonstration selection and retrieval. Adaptive retrieval methods improve ICL efficiency by dynamically selecting relevant examples, boosting performance, and reducing computational overhead [40]. Training demonstration retrievers is critical for refining algorithms that select and organize in-context examples, exposing LLMs to high-quality demonstrations that maximize learning outcomes [40].

Integrating multimodal data into ICL processes enhances LLM capabilities, particularly in addressing challenges faced by current multimodal large language models (MLLMs). Frameworks like AIM, which aggregate image information into textual labels, effectively upgrade MLLMs for multimodal ICL tasks, improving performance in applications like image captioning and visual question answering [6, 39]. Incorporating diverse data types into retrieval and demonstration selection enables LLMs to achieve a comprehensive understanding of tasks requiring cross-modal reasoning.

To fully leverage LLMs' reasoning and problem-solving capabilities, continuous refinement of retrieval methods and demonstration strategies is essential. Fine-tuning LLMs with high-quality Chain-of-Thought (CoT) rationales and leveraging past mistakes significantly enhance performance. The COTERRORSET benchmark facilitates systematic error analysis, revealing that self-rethinking prompting and mistake tuning improve reasoning abilities. Retrieval-based ICL, sourcing semantically similar demonstrations, proves more effective than random selection methods. These advancements underscore iterative improvements in retrieval and demonstration techniques to maximize LLM efficacy in complex reasoning tasks [4, 62]. Focusing on adaptive retrieval techniques and effective demonstration retriever training enhances LLM in-context learning capabilities, paving the way for advanced applications in natural language processing.

5 Challenges and Limitations

Understanding the challenges and limitations of large language models (LLMs) in the context of in-context learning (ICL) is essential for evaluating their performance across cognitive tasks. This section explores specific issues such as demonstration example selection, generalizability, and the impact of noisy labels, setting the stage for further exploration.

5.1 Challenges and Limitations in Cognitive Tasks

LLMs face significant obstacles in executing cognitive tasks through ICL, affecting their effectiveness and dependability. A critical issue is the selection of demonstration examples, which is crucial for ICL success. Current approaches lack consensus on key indicators, often resulting in suboptimal solutions that fail to consistently improve model performance [1]. Moreover, the narrow focus of many studies on specific natural language processing (NLP) benchmarks raises concerns about the generalizability of findings across various tasks [2], potentially hindering the development of adaptable ICL methods.

The presence of noisy labels in demonstrations further complicates cognitive task execution, negatively impacting model performance and stability. Effective strategies for managing label noise are critical for maintaining accuracy and reliability. Additionally, knowledge hijacking, where models overly depend on in-context information at the expense of global knowledge, leads to incorrect predictions and underscores the need to balance in-context and pre-trained knowledge [37].

Generating correctly formatted logical forms from limited demonstration examples is another challenge, especially in tasks requiring precise logical reasoning [15]. Variability in prompt formats complicates ICL evaluation, making it challenging to draw valid conclusions about effectiveness [21]. Addressing these challenges involves establishing comprehensive theoretical frameworks, refining generalization strategies, and implementing advanced bias mitigation techniques. Learning from past errors through self-rethinking prompting and mistake tuning can enhance model reliability and fairness. Incorporating diverse and representative samples in ICL prompts can significantly improve fairness without sacrificing predictive accuracy [67, 62].

To visualize these challenges and potential improvement strategies, Figure 6 illustrates the hierarchical structure of challenges and improvement strategies in cognitive tasks involving LLMs, focusing on demonstration selection, knowledge management, and improvement strategies. By tackling these issues, researchers can improve LLM efficacy in complex cognitive tasks.

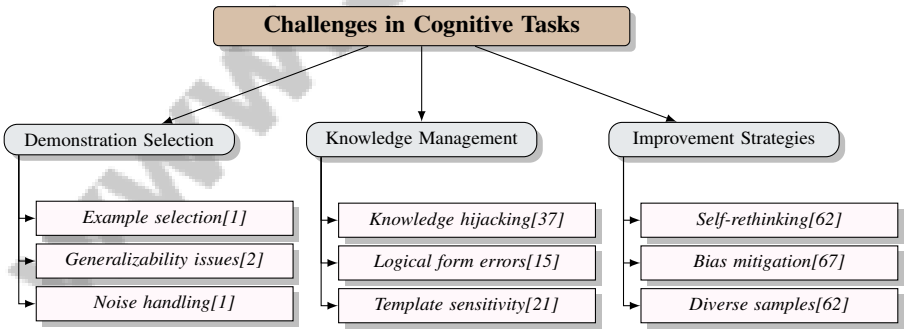


Figure 6: This figure illustrates the hierarchical structure of challenges and improvement strategies in cognitive tasks involving LLMs, focusing on demonstration selection, knowledge management, and improvement strategies.

5.2 Model Robustness and Sensitivity

The robustness and sensitivity of LLMs in ICL are critical for their performance across applications. A major challenge is the majority label bias in benchmarks, which skews the understanding of model performance under varied label distributions [53]. This bias necessitates developing evaluation frameworks that account for input data and label distribution variability.

Moreover, LLM sensitivity to input variations is complicated by current data poisoning strategies, which are less effective in ICL contexts where models do not adhere to fixed training principles

[68]. Innovative approaches are needed to assess and enhance LLM robustness against adversarial inputs and data manipulation. The In-Context Retrieval (ICR) method offers a promising solution by leveraging candidate pools and LLM judgments without relying on external knowledge bases or extensive computational resources [69]. This approach enhances robustness by focusing on LLMs' intrinsic capabilities to process diverse input contexts, ensuring consistent performance across scenarios.

Enhancing LLM robustness and sensitivity in ICL requires a multifaceted approach that integrates advanced evaluation techniques, innovative data handling strategies, and context-aware retrieval methods. By systematically examining the impact of context type, retrieval methods, and model variations on LLM performance, researchers can improve the reliability and effectiveness of these models in addressing complex input data challenges, advancing natural language processing [26, 25].

5.3 Scalability and Computational Costs

Scalability and computational costs are critical challenges in deploying and optimizing LLMs for ICL. A primary concern is the computational expense associated with training energy-based models, which require extensive differentiation and backpropagation, significantly increasing computational burden and complicating efficient scaling [70]. The selection of demonstration examples also presents scalability challenges; while effective selection can enhance performance, the computational demands of current methods may be prohibitive for very large models [71].

In multimodal contexts, the AIM framework offers significant efficiency gains in memory usage and parameter training, enabling better scalability in handling complex inputs without incurring prohibitive computational costs [39]. Addressing scalability and computational cost issues requires developing innovative training methodologies and optimization techniques to reduce resource consumption while maintaining or enhancing performance. By focusing on context retrieval, knowledge unlearning, and efficient training methods, researchers can enhance the practicality and accessibility of LLMs across diverse applications, tackling challenges in low-resource language translation and the retention of harmful knowledge [26, 22, 14].

5.4 Benchmarking and Evaluation Challenges

Benchmarking and evaluating LLMs in ICL presents challenges that impede accurate assessment of capabilities and performance. A primary challenge is the inherent variability in the selection and organization of demonstration examples, which can significantly impact ICL task performance. Research indicates that retrieving semantically similar demonstrations from a curated pool enhances model outcomes, underscoring the importance of task-specific relevance criteria for maximizing predictive accuracy [6, 4]. This variability complicates establishing standardized benchmarks for consistent performance measurement across contexts and tasks.

Moreover, majority label bias in existing benchmarks further complicates evaluation by skewing perceptions of model robustness and effectiveness across diverse label distributions. This bias highlights the need for nuanced evaluation frameworks that account for variability in input data [68]. The ICLPoison approach offers a novel perspective on assessing robustness by exploiting ICL's vulnerabilities to subtle data manipulations, providing insights into LLMs' resilience and adaptability to adversarial inputs [68].

To effectively address the challenges of benchmarking and evaluating LLMs in ICL, it is essential to develop sophisticated evaluation techniques and frameworks that capture the intricate dynamics of ICL tasks, particularly given the reliance on few-shot examples and the importance of task-specific relevance in selecting examples for optimal performance [6, 4, 8]. By focusing on these areas, researchers can enhance the reliability and validity of LLM assessments, ultimately advancing the field of natural language processing.

6 Future Directions

Advancements in large language models (LLMs) drive the need to explore future directions for refining in-context learning (ICL) methodologies. This section highlights theoretical and methodological

gaps, providing a foundation for enhancing ICL practices and enabling robust LLM applications across domains.

6.1 Theoretical and Methodological Gaps

Addressing theoretical and methodological gaps in ICL is crucial for enhancing LLM performance and versatility. ICL, a pivotal approach in natural language processing (NLP), allows LLMs to predict based on contextual prompts. Research emphasizes areas needing exploration, such as refining model-generated demonstrations and integrating strategies like SEC to boost ICL efficacy [8, 20]. Analyzing demonstration quality and template design across languages and tasks is essential for optimizing ICL outcomes.

Future research should examine diverse tasks and datasets, focusing on how demonstration types affect ICL performance [72]. Developing models that predict the optimal number of examples and methods for incorporating diversity in example selection is critical [6]. Enhancing the distillation process by aligning language modeling with ICL objectives could significantly improve performance [73].

Understanding effective demonstrations and retrieval methods in smaller language models remains a significant gap [40]. Future efforts could focus on multitask tuning for unified retrievers applicable to new tasks without retraining, addressing CEIL framework limitations [35]. Optimizing caching mechanisms for demonstration storage and improving aggregation processes can enhance ICL efficiency [39].

Exploring meta-in-context learning for complex tasks and larger datasets, along with its application in multimodal models, presents promising research avenues [74]. Enhancing ICL efficiency, investigating low-resource settings, and developing frameworks to understand ICL mechanisms are vital areas for future inquiry [20]. Exploring ICL and Instruction Tuning (IT) relationships, improving entity linking, and relation matching processes could significantly boost performance [23, 15]. Findings on other tasks, such as generation, should focus on optimizing demonstrations for improved performance [2]. Predicting dynamically generated example performance and extending influence frameworks to other NLP tasks are potential research avenues [1]. Enhancing model efficiency, contextual understanding, and bias detection methods are crucial [22].

6.2 Expanding Applications and Domains

Expanding ICL applications and domains is essential for advancing LLM adaptability across diverse tasks. Research should expand datasets to cover broader scenarios, improving model adaptability [46]. This expansion will enhance LLM utility across sectors.

Incorporating active retrieval methods and enhancing retrieval models for smaller language models are promising directions. Investigating effective demonstration selection can optimize in-context example selection, improving LLM performance [40].

In speech processing, exploring ICL on advanced speech LMs and developing warmup training strategies are vital for enhancing performance in automated transcription and voice-activated systems [42].

Applying Iterative Demonstration Selection (IDS) to tasks like summarization and translation could enhance ICL effectiveness, broadening IDS utility in optimizing demonstration selection and highlighting new research domains in NLP [6, 4, 75]. The Deep-Thinking framework's application across tasks like math reasoning and code generation could validate its effectiveness and expand ICL applicability.

Enhancing ICL adaptability and performance can be achieved by developing adaptive algorithms that modify selection strategies based on task complexity and integrating advanced retrieval techniques prioritizing few-shot examples' relevance. This aligns with findings suggesting ICL effectiveness improves with task-specific relevance definitions and self-adaptive mechanisms for example selection and ordering, leading to better predictions and performance metrics [6, 4, 76, 8]. Exploring NLP tasks beyond classification, such as natural language inference and generative tasks, could further advance ICL understanding and application.

6.3 Improving Robustness and Efficiency

Enhancing LLM robustness and efficiency in ICL is crucial for expanding applicability across domains. Future research should optimize evaluation frameworks and prompt design for real-world application performance assessment. Key aspects include refining demonstration selection to improve robustness and efficiency in handling diverse inputs [71].

Exploring dynamic demonstration controllers, like the D2 Controller, ensures models consistently approach optimal performance across datasets, enhancing adaptability through dynamic input condition adjustments [77]. Integrating symbolic reasoning and enhancing control mechanisms in reasoning are promising directions for addressing scalability and expanding applications [64].

Future work should enhance long-context understanding capabilities and explore new assessment methodologies for tasks requiring extensive contextual processing [78]. Expanding data augmentation and integrating fairness metrics can enhance FairICL robustness, ensuring equitable and reliable outputs [79].

The Ensemble SuperICL method improves LLM robustness in domain-specific tasks by leveraging specialized knowledge from multiple small models, demonstrating ensemble approaches' efficacy in enhancing ICL [80]. Future research should explore additional reasoning tasks, improve LLM inhibitory control, and replicate studies with closed-source models to enhance robustness [81].

In low-resource translation contexts, expanding datasets and refining evaluation metrics enhance LLM robustness, enabling accurate translations in underrepresented languages and domains [26]. Future work will focus on enhancing robustness through targeted training on specific datasets to improve contextual accuracy [82].

Focusing on ICL's theoretical and empirical aspects will enhance LLM robustness and efficiency within this innovative paradigm. This approach facilitates better ICL utilization across diverse NLP tasks and aids in identifying and mitigating technology risks, such as bias and misinformation. Advancements in training strategies, prompt design, and task-specific relevance metrics can lead to more effective and versatile LLM applications, driving NLP progress [6, 8, 18].

6.4 Developing Novel Evaluation and Benchmarking Methods

Developing novel evaluation and benchmarking methods for LLMs in ICL is crucial for advancing capabilities and ensuring reliable performance across applications. Current frameworks inadequately capture ICL task complexities, necessitating innovative assessment methodologies considering task-specific relevance and leveraging neural ranking model advances to optimize few-shot demonstration selection. Redefining relevance to enhance downstream performance provides a more accurate measure of capabilities, improving ICL effectiveness in NLP applications [6, 4, 8].

Integrating ethical considerations, such as bias, fairness, privacy, and misinformation risks, into benchmarking ensures LLMs are evaluated on technical performance and ethical alignment, promoting responsible AI development [22].

To evaluate LLM performance in real-world applications effectively, innovative metrics and methodologies addressing ICL's unique challenges and downstream tasks' specific requirements are crucial, as these factors significantly influence LLM output effectiveness [83, 9, 25, 33, 6]. Developing benchmarks that account for input data and label distribution variability ensures comprehensive understanding of model robustness and effectiveness across contexts. Focusing on these areas improves LLM assessment reliability and validity, advancing NLP.

Integrating sophisticated evaluation techniques, such as adversarial and stress testing, is essential for understanding LLM resilience and adaptability in ICL scenarios. These methodologies identify potential biases skewing evaluation results and reveal vulnerabilities, such as data poisoning susceptibility, informing robust defense development. Employing these techniques enables better LLM performance assessment under varied conditions, enhancing reliability and effectiveness in real-world applications [68, 84, 83, 4]. These methods identify vulnerabilities and areas for improvement, guiding future research directions and informing more robust model development.

By introducing innovative evaluation and benchmarking methodologies for LLMs in ICL, researchers can deepen understanding. This advancement addresses potential evaluation biases, enhancing assessment reliability. As ICL evolves, it opens new LLM application avenues across domains, including

data engineering and machine translation, leading to more effective and versatile applications in NLP and beyond [83, 32, 25, 8].

6.5 Integrating External Knowledge and Multimodal Inputs

Integrating external knowledge and multimodal inputs enhances LLM capabilities in ICL. By incorporating various data types and external information sources, LLMs develop a deeper understanding of complex tasks, improving adaptability and performance across fields. Research on Math Word Problems (MWPs) shows LLM struggles with longer contexts; however, tailored prompts and auxiliary tasks can enhance reasoning capabilities. Multimodal models' effectiveness is influenced by careful in-context example selection, highlighting textual and visual data's importance in optimizing performance [85, 86, 26, 56].

Multimodal large language models (MLLMs) leverage text, images, and audio to perform cross-modal reasoning tasks. This capability benefits scenarios where single-modal inputs lack sufficient context for accurate execution. For example, the AIM framework aggregates image information from multimodal demonstrations into textual labels' latent space, enhancing MLLM ICL efficiency and accuracy [39].

Integrating external knowledge sources significantly enhances LLM reasoning capabilities. By accessing structured databases, ontologies, or domain-specific repositories, LLMs supplement pre-trained knowledge with relevant external data, leading to informed decision-making and problem-solving. This approach is advantageous in knowledge-intensive tasks, like knowledge-based question answering (KBQA), where external knowledge bridges training data gaps [15].

Developing advanced retrieval models and inference algorithms is crucial for optimizing external knowledge and multimodal input selection and integration. Researchers should refine these processes to ensure LLMs access high-quality, contextually relevant information enhancing learning outcomes. This involves exploring various retrieval strategies and algorithms to determine effective methods for different tasks and datasets [40].

7 Conclusion

This survey comprehensively examines in-context learning (ICL) and its crucial role in enhancing the cognitive abilities of large language models (LLMs). The findings highlight ICL as a transformative approach that enables LLMs to effectively leverage contextual information from input data, thereby improving adaptability and performance across diverse tasks. The quality and organization of demonstrations are critical factors in optimizing ICL strategies and enhancing model performance [20]. The exploration of the Knowledgeable In-Context Tuning (KICT) framework further illustrates its potential to significantly boost LLM performance by utilizing factual knowledge, suggesting promising avenues for future research [16].

Moreover, the survey emphasizes the societal implications of knowledge hijacking, which is vital for the safe and reliable use of ICL capabilities in LLMs [37]. Understanding this phenomenon is essential for mitigating risks associated with ICL and ensuring the ethical deployment of these models. The need for robust evaluation metrics and methodologies also emerges as a key takeaway, underscoring the importance of developing comprehensive frameworks that accurately assess model capabilities and address the complexities inherent in ICL tasks [18].

In the broader context of natural language processing, the survey illustrates ICL's potential to drive advancements in LLMs, paving the way for more sophisticated and versatile applications. By addressing the identified challenges and pursuing future research directions, the field can continue to evolve, ultimately enhancing the efficacy and reliability of LLMs in processing complex and variable input data.

References

[1] Tai Nguyen and Eric Wong. In-context example selection with influences, 2023.

[2] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work?, 2022.

[3] Ivan Lee, Nan Jiang, and Taylor Berg-Kirkpatrick. Is attention required for icl? exploring the relationship between model architecture and in-context learning ability, 2024.

[4] Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y Zhao. Dr.icl: Demonstration-retrieved in-context learning, 2023.

[5] Jannik Kossen, Yarin Gal, and Tom Rainforth. In-context learning learns label relationships but is not conventional learning, 2024.

[6] Andrew Parry, Debasis Ganguly, and Manish Chandra. "in-context learning" or: How i learned to stop worrying and love "applied information retrieval", 2024.

[7] Yu Bai, Heyan Huang, Cesare Spinoso-Di Piano, Marc-Antoine Rondeau, Sanxing Chen, Yang Gao, and Jackie Chi Kit Cheung. Identifying and analyzing performance-critical tokens in large language models, 2025.

[8] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning, 2024.

[9] Yinheng Li. A practical survey on zero-shot prompt design for in-context learning, 2023.

[10] Shivanshu Gupta, Matt Gardner, and Sameer Singh. Coverage-based example selection for in-context learning, 2023.

[11] Amirhesam Abedsoltan, Adityanarayanan Radhakrishnan, Jingfeng Wu, and Mikhail Belkin. Context-scaling versus task-scaling in in-context learning, 2024.

[12] Yoonsang Lee, Pranav Atreya, Xi Ye, and Eunsol Choi. Crafting in-context examples according to lms' parametric knowledge, 2024.

[13] Zixian Huang, Jiaying Zhou, Gengyang Xiao, and Gong Cheng. Enhancing in-context learning with answer feedback for multi-span question answering, 2023.

[14] Nianwen Si, Hao Zhang, Heyu Chang, Wenlin Zhang, Dan Qu, and Weiqiang Zhang. Knowledge unlearning for llms: Tasks, methods, and challenges, 2023.

[15] Zhijie Nie, Richong Zhang, Zhongyuan Wang, and Xudong Liu. Code-style in-context learning for knowledge-based question answering, 2024.

[16] Jianing Wang, Chengyu Wang, Chuanqi Tan, Jun Huang, and Ming Gao. Knowledgeable in-context tuning: Exploring and exploiting factual knowledge for in-context learning, 2024.

[17] Yiming Tang and Bin Dong. Demonstration notebook: Finding the most suited in-context learning example from interactions, 2024.

[18] Yuxiang Zhou, Jiazheng Li, Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. The mystery of in-context learning: A comprehensive survey on interpretation and analysis, 2024.

[19] Miaoran Zhang, Vagrant Gautam, Mingyang Wang, Jesujoba O. Alabi, Xiaoyu Shen, Dietrich Klakow, and Marius Mosbach. The impact of demonstrations on multilingual in-context learning: A multidimensional analysis, 2024.

[20] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[21] Anton Voronov, Lena Wolf, and Max Ryabinin. Mind your format: Towards consistent evaluation of in-context learning improvements, 2024.

[22] Milad Moradi, Ke Yan, David Colwell, Matthias Samwald, and Rhona Asgari. Exploring the landscape of large language models: Foundations, techniques, and challenges, 2024.

[23] Hanyu Duan, Yixuan Tang, Yi Yang, Ahmed Abbasi, and Kar Yan Tam. Exploring the relationship between in-context learning and instruction tuning, 2023.

[24] Kevin Christian Wibisono and Yixin Wang. From unstructured data to in-context learning: Exploring what tasks can be learned and when, 2024.

[25] Javad Pourmostafa Roshan Sharami, Dimitar Shterionov, and Pieter Spronck. Guiding in-context learning of llms through quality estimation for machine translation, 2024.

[26] Sara Court and Micha Elsner. Shortcomings of llms for low-resource translation: Retrieval and understanding are both the problem, 2024.

[27] Jérémie Cabessa, Hugo Hernault, and Umer Mushtaq. In-context learning and fine-tuning gpt for argument mining, 2024.

[28] Jiaao Chen, Xiaoman Pan, Dian Yu, Kaiqiang Song, Xiaoyang Wang, Dong Yu, and Jianshu Chen. Skills-in-context: Unlocking compositionality in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13838–13890, 2024.

[29] Yoav Levine, Noam Wies, Daniel Jannai, Dan Navon, Yedid Hoshen, and Amnon Shashua. The inductive bias of in-context learning: Rethinking pretraining example design, 2022.

[30] Gulsum Yigit and Mehmet Fatih Amasyali. Data augmentation with in-context learning and comparative evaluation in math word problem solving, 2024.

[31] Dong Shu and Mengnan Du. Comparative analysis of demonstration selection algorithms for llm in-context learning, 2024.

[32] Nicholas Kroeger, Dan Ley, Satyapriya Krishna, Chirag Agarwal, and Himabindu Lakkaraju. In-context explainers: Harnessing llms for explaining black box models, 2024.

[33] Bingsheng Yao, Guiming Chen, Ruishi Zou, Yuxuan Lu, Jiachen Li, Shao Zhang, Yisi Sang, Sijia Liu, James Hendler, and Dakuo Wang. More samples or more prompts? exploring effective in-context sampling for llm few-shot prompt engineering, 2024.

[34] Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. *Advances in Neural Information Processing Systems*, 36:15614–15638, 2023.

[35] Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning, 2023.

[36] Junyong Kang, Donghyun Son, Hwanjun Song, and Buru Chang. In-context learning with noisy labels, 2024.

[37] Shuo Wang and Issei Sato. Understanding knowledge hijack mechanism in in-context learning through associative memory, 2024.

[38] Fan Wang, Chuan Lin, Yang Cao, and Yu Kang. Benchmarking general-purpose in-context learning, 2024.

[39] Jun Gao, Qian Qiao, Ziqiang Cao, Zili Wang, and Wenjie Li. Aim: Let any multi-modal large language models embrace efficient in-context learning, 2024.

[40] Man Luo, Xin Xu, Yue Liu, Panupong Pasupat, and Mehran Kazemi. In-context learning with retrieved demonstrations for language models: A survey, 2024.

[41] Peng Wang, Xiaobin Wang, Chao Lou, Shengyu Mao, Pengjun Xie, and Yong Jiang. Effective demonstration annotation for in-context learning via language model-based determinantal point process, 2024.

[42] Ming-Hao Hsu, Kai-Wei Chang, Shang-Wen Li, and Hung yi Lee. Exploring in-context learning of textless speech language model for speech classification tasks, 2024.

[43] Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. Pre-training to learn in context, 2023.

[44] Abdelhakim Benechehab, Youssef Attia El Hili, Ambroise Odonnat, Oussama Zekri, Albert Thomas, Giuseppe Paolo, Maurizio Filippone, Ievgen Redko, and Balázs Kégl. Zero-shot model-based reinforcement learning using large language models, 2024.

[45] Haowei Du and Dongyan Zhao. In-context learning with reinforcement learning for incomplete utterance rewriting, 2024.

[46] Pengshuo Qiu, Frank Rudzicz, and Zining Zhu. Scenarios and approaches for situated natural language explanations, 2024.

[47] Xiang Li, Haoran Tang, Siyu Chen, Ziwei Wang, Ryan Chen, and Marcin Abram. Why does in-context learning fail sometimes? evaluating in-context learning on open and closed questions. *arXiv preprint arXiv:2407.02028*, 2024.

[48] Kartik Ahuja and David Lopez-Paz. A closer look at in-context learning under distribution shifts, 2023.

[49] Madeline Brumley, Joe Kwon, David Krueger, Dmitrii Krasheninnikov, and Usman Anwar. Comparing bottom-up and top-down steering approaches on in-context learning tasks, 2024.

[50] Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Sébastien M. R. Arnold, Vincent Perot, Siddharth Dalmia, Hexiang Hu, Xudong Lin, Panupong Pasupat, Aida Amini, Jeremy R. Cole, Sebastian Riedel, Iftekhar Naim, Ming-Wei Chang, and Kelvin Guu. Can long-context language models subsume retrieval, rag, sql, and more?, 2024.

[51] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know?, 2023.

[52] Taihang Wang, Xiaoman Xu, Yimin Wang, and Ye Jiang. Instruction tuning vs. in-context learning: Revisiting large language models in few-shot computational social science, 2024.

[53] Karan Gupta, Sumegh Roychowdhury, Siva Rajesh Kasa, Santhosh Kumar Kasa, Anish Bhanushali, Nikhil Pattisapu, and Prasanna Srinivasa Murthy. How robust are llms to in-context majority label bias?, 2023.

[54] Jongho Park, Jaeseung Park, Zheyang Xiong, Nayoung Lee, Jaewoong Cho, Samet Oymak, Kangwook Lee, and Dimitris Papailiopoulos. Can mamba learn how to learn? a comparative study on in-context learning tasks, 2024.

[55] Hao Peng, Xiaozhi Wang, Jianhui Chen, Weikai Li, Yunjia Qi, Zimu Wang, Zhili Wu, Kaisheng Zeng, Bin Xu, Lei Hou, et al. When does in-context learning fall short and why? a study on specification-heavy tasks. *arXiv preprint arXiv:2311.08993*, 2023.

[56] Xin Xu, Tong Xiao, Zitong Chao, Zhenya Huang, Can Yang, and Yang Wang. Can llms solve longer math word problems better? *arXiv preprint arXiv:2405.14804*, 2024.

[57] Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. Large language models are in-context semantic reasoners rather than symbolic reasoners, 2023.

[58] Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. Large language models are in-context semantic reasoners rather than symbolic reasoners. *arXiv preprint arXiv:2305.14825*, 2023.

[59] Ben Bogin, Shivanshu Gupta, Peter Clark, and Ashish Sabharwal. Leveraging code to improve in-context learning for semantic parsing, 2024.

[60] Yukang Lin, Bingchen Zhong, Shuoran Jiang, Joanna Siebert, and Qingcai Chen. Reasoning graph enhanced exemplars retrieval for in-context learning. *arXiv preprint arXiv:2409.11147*, 2024.

[61] Junbing Yan, Chengyu Wang, Jun Huang, and Wei Zhang. Do large language models understand logic or just mimick context?, 2024.

[62] Yongqi Tong, Dawei Li, Sizhe Wang, Yujia Wang, Fei Teng, and Jingbo Shang. Can llms learn from previous mistakes? investigating llms' errors to boost for reasoning. *arXiv preprint arXiv:2403.20046*, 2024.

[63] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. Reasoning with large language models, a survey, 2024.

[64] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*, 2024.

[65] Jiayu Liu, Zhenya Huang, Chaokun Wang, Xunpeng Huang, Chengxiang Zhai, and Enhong Chen. What makes in-context learning effective for mathematical reasoning: A theoretical analysis. *arXiv preprint arXiv:2412.12157*, 2024.

[66] Yongqi Tong, Yifan Wang, Dawei Li, Sizhe Wang, Zi Lin, Simeng Han, and Jingbo Shang. Eliminating reasoning via inferring with planning: A new framework to guide llms' non-linear thinking. *arXiv preprint arXiv:2310.12342*, 2023.

[67] Jingyu Hu, Weiru Liu, and Mengnan Du. Strategic demonstration selection for improved fairness in llm in-context learning, 2024.

[68] Pengfei He, Han Xu, Yue Xing, Hui Liu, Makoto Yamada, and Jiliang Tang. Data poisoning for in-context learning, 2024.

[69] Shangqing Xu and Chao Zhang. Misconfidence-based demonstration selection for llm in-context learning, 2024.

[70] Rylan Schaeffer, Mikail Khona, and Sanmi Koyejo. In-context learning of energy functions, 2024.

[71] Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. Revisiting demonstration selection strategies in in-context learning, 2024.

[72] Quanyu Long, Yin Wu, Wenya Wang, and Sinno Jialin Pan. Does in-context learning really learn? rethinking how large language models respond and solve tasks via in-context learning, 2024.

[73] Yukun Huang, Yanda Chen, Zhou Yu, and Kathleen McKeown. In-context learning distillation: Transferring few-shot learning ability of pre-trained language models, 2022.

[74] Julian Coda-Forno, Marcel Binz, Zeynep Akata, Matthew Botvinick, Jane X. Wang, and Eric Schulz. Meta-in-context learning in large language models, 2023.

[75] Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. In-context learning with iterative demonstration selection, 2024.

[76] Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering, 2023.

[77] Fei Zhao, Taotian Pang, Zhen Wu, Zheng Ma, Shujian Huang, and Xinyu Dai. Dynamic demonstrations controller for in-context learning, 2024.

[78] Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. Long-context llms struggle with long in-context learning, 2024.

[79] Karuna Bhaila, Minh-Hao Van, Kennedy Edemacu, Chen Zhao, Feng Chen, and Xintao Wu. Fair in-context learning via latent concept variables, 2024.

[80] M. Mehdi Mojarradi, Lingyi Yang, Robert McCraith, and Adam Mahdi. Improving in-context learning with small language model ensembles, 2024.

[81] Pengrui Han, Peiyang Song, Haofei Yu, and Jiaxuan You. In-context learning may not elicit trustworthy reasoning: A-not-b errors in pretrained language models. *arXiv preprint arXiv:2409.15454*, 2024.

[82] Milad Fotouhi, Mohammad Taha Bahadori, Oluwaseyi Feyisetan, Payman Arabshahi, and David Heckerman. Fast training dataset attribution via in-context learning, 2024.

[83] Mingyang Song, Mao Zheng, Xuan Luo, and Yue Pan. Can many-shot in-context learning help llms as evaluators? a preliminary empirical study, 2025.

[84] Wai Man Si, Michael Backes, and Yang Zhang. Iclguard: Controlling in-context learning behavior for applicability authorization, 2024.

[85] Yi Lu, Jing Nathan Yan, Songlin Yang, Justin T. Chiu, Siyu Ren, Fei Yuan, Wenting Zhao, Zhiyong Wu, and Alexander M. Rush. A controlled study on long context extension and generalization in llms, 2024.

[86] Yang Luo, Zangwei Zheng, Zirui Zhu, and Yang You. How does the textual information affect the retrieval of multimodal in-context learning?, 2024.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.