In-Context Learning in Large Language Models: Mechanisms, Challenges, and Frontiers

Yuqing Yuan[1*]

[1]School of Telecommunity Engineering, Xi'an Jiaotong University, 710049 Xi'an, China

Abstract. In recently years, large language models (LLMs) show an ability to learn directly from examples embedded in their input, a process known as in-context learning (ICL). This learning approach enables the model to comprehend examples within the context and generalize to new tasks without the need for modifying model parameters. This indicates that the learning process can be fully realized during the inference phase. The phenomenon was first observed in GPT-3 and has since become central to understanding reasoning in large models. Studies describe ICL as implicit Bayesian inference, as internal simulation of learning algorithms, or as the operation of induction heads within attention circuits. Recent work extends these perspectives: The Implicit Dynamics of ICL links activation updates to posterior inference, In-Context Learning with Long-Context Models explores many-shot scaling, and VL-ICL Bench evaluates multimodal adaptation. Overall, these studies indicate that ICL integrates multiple features such as statistical inference, algorithmic approximation, and mechanistic emergence. However, challenges persist in theoretical integration, reproducibility, and robustness. This article will review the theoretical framework, empirical results, and the latest research progress of ICL in 2025. The aim is to clarify its internal mechanism, reveal the current limitations, and envision the future development direction, in order to promote a more systematic and coherent understanding of ICL.

1 Introduction

Artificial intelligence has experienced several major stages of development. It started with symbolic reasoning, later moved into probabilistic learning, and has now entered the era of large neural networks that can handle complex and abstract patterns. Large Language Models (LLMs), trained on massive text data, have changed how people think about learning and reasoning. Among the many new abilities shown by these models, in-context learning (ICL) is perhaps the most surprising. When a model is given only a few examples in a prompt, it can recognise the pattern and make accurate predictions for new inputs. What is special is that this happens without changing the model's parameters — the system seems to learn directly from the context itself.

The discovery of this behaviour in GPT-3 changed how researchers understood machine learning [1]. Before that, adapting to new tasks was thought to require gradient-based

* Corresponding author's email: 2213420425@stu.xjtu.edu.cn

parameter updates. But experiments showed that a pretrained transformer could perform translation, question answering, and classification simply by reading examples in the input. These findings suggested that large-scale pretraining not only stores linguistic information but also builds an internal mechanism that allows quick adaptation. The model appears to do more than recall facts; it performs a kind of temporary reasoning during its forward process.

Different studies have tried to explain how this happens. One common view describes ICL as a kind of implicit Bayesian inference, where pretraining builds a prior over tasks and relationships, and the examples in the prompt act as evidence to update this prior [2]. The next-token prediction can then be seen as the posterior mean. Another explanation is that transformers imitate known learning algorithms such as linear regression or gradient descent [3]. In this view, the attention and feed-forward layers behave like an inner learning loop. A third perspective focuses on specific attention circuits, called induction heads, which copy and align patterns across tokens, creating a basic form of meta-learning. Each explanation captures part of the story, but none can yet give a full picture of where ICL comes from or how far it can go.

Subsequent experiments have shown that the performance of ICL depends not only on the model architecture or scale, but also on the way data is organized. A dataset with long-range structure or repeatability can more effectively prompt models to reason, while scrambled or noisy data can weaken this ability. This finding supports the view that patterns in training data strongly shape the way the model learns relationships. Researchers also found that the order and selection of examples in prompts significantly affect accuracy. The use of automatic retrieval, selection, and testing of similar examples in queries can bring more stable results, indicating that aligning the context with the query is more important than simply increasing the number of examples [4]. This type of sensitivity means that ICL is controlled by a subtle balance between the input distribution and the internal representation.

Context length also plays an important role. When the number of examples increases, accuracy first rises, then levels off, and can even drop if attention spreads too thinly. This pattern—often called the “lost-in-the-middle” effect—shows that transformers still struggle to maintain focus across very long sequences. Recent long-context models try to fix this problem by changing attention design or adding memory components. These improvements suggest that ICL ability can grow when models are given larger context windows and better positional encoding [5].

At the same time, new interpretability studies have begun connecting theory with mechanism. Researchers analysing activation paths found that token representations change across layers in a way that looks like Bayesian updating. Rather than mapping input to output once and for all, the model seems to adjust its internal guesses after reading each example. This dynamic behaviour implies that ICL involves a real reasoning process inside activations, not just pattern copying. It also strengthens the link between probabilistic inference and neural computation.

In recent years, ICL research has extended beyond text to multimodal and cross-domain settings. With the rise of vision-language models, researchers asked whether similar contextual adaptation could occur in image or audio tasks. The VL-ICL Bench evaluated this question and found that multimodal prompts can produce limited generalisation when both visual and textual information are combined. Yet performance remains unstable and depends heavily on input alignment. Even so, these results imply that contextual reasoning might be a general property of attention-based systems, not something limited to language.

By 2025, the field had become an active intersection of theory, experiments, and applications. New works such as The Implicit Dynamics of ICL explored how transformer activations perform approximate posterior inference, while In-Context Learning with Long-Context Models examined scaling effects in thousand-shot settings. These studies suggest

that probabilistic reasoning, algorithmic simulation, and mechanistic structure may be different layers of the same phenomenon. Yet many basic questions are still open.

This paper aims to explore these issues through a full review of recent studies. It discusses theoretical views, summarises empirical findings, and analyses current challenges that shape the future of ICL research. The goal is to show how contextual learning arises, what limits it, and what future work might overcome those limits. The paper is organised as follows: Section 2 reviews theoretical and technical progress; Section 3 analyses research findings; Section 4 highlights key challenges; Section 5 proposes future directions; and Section 6 concludes with broader reflections on artificial intelligence.

2 Related Theoretical Framework and Technological Progress

The study of in-context learning has evolved from a descriptive observation into a rigorous investigation of mechanism and theory. As the phenomenon became more central to understanding generalisation in large language models, researchers developed a set of conceptual frameworks that attempt to explain how a fixed network can behave as a learner during inference. These frameworks, though diverse, increasingly converge toward a view of ICL as an emergent result of scale, data structure, and architectural design.

2.1 Bayesian and Probabilistic Interpretations

The Bayesian perspective remains one of the most influential explanations of ICL. It treats the transformer as an amortised inference engine that implicitly represents distributions over tasks. During pretraining, the model learns to encode priors over linguistic patterns, semantic relations, and compositional rules. When presented with a sequence of examples in context, it updates these priors into posteriors conditioned on observed evidence, producing predictions equivalent to the posterior mean, highlighting the statistical nature of ICL as an approximate Bayesian inference process. In this sense, ICL can be viewed as statistical inference through forward computation.

Recent analyses extend this framework to quantify how accuracy scales with the number of demonstrations. The Bayesian Scaling Laws for ICL proposed that model performance follows predictable curves analogous to sample-efficient learners, where improvements diminish as contextual evidence saturates. The Implicit Dynamics of ICL connected this statistical process to neural activations, showing that hidden states across transformer layers evolve according to equations similar to Bayesian belief updates. Together, these works suggest that what appears as learning within context is, in fact, probabilistic reasoning distributed across the network's forward path.

2.2 Algorithmic Simulation and Implicit Meta-Learning

A second major interpretation views ICL as the simulation of classical learning algorithms within network activations. Under this paradigm, attention and feed-forward layers mimic the computations of gradient descent [6], regression fitting, or kernel methods. Each example in context modifies the residual stream, producing intermediate representations that function like parameter updates, even though the model's weights remain fixed. When the prompt contains multiple examples, the cumulative effect of attention reweighting approximates iterative optimisation.

This interpretation links ICL closely to meta-learning. Both aim to adapt quickly to new tasks from limited data. Yet meta-learning explicitly optimises over tasks through gradient updates, while ICL performs the same adaptation implicitly in a single forward pass. Researchers have argued that pretraining on diverse tasks equips the transformer with a learned learning algorithm embedded in its structure. Dong et al. summarised this

equivalence, noting that the key difference lies in time scale: meta-learning occurs across training episodes, whereas ICL enacts it instantaneously within context. This analogy helps bridge neural computation with algorithmic theory, grounding intuitive notions of "learning without updates" in formal reasoning. Recent case studies on simple function classes further demonstrate that transformers can learn linear and non-linear mappings in-context under appropriate prompt design.

2.3 Mechanistic Explanations and Attention Circuits

While probabilistic and algorithmic views capture the functional level, mechanistic analysis seeks to identify concrete structures inside the model that enable contextual adaptation. Research on attention heads shows that they have different functions, such as copying, aligning, and abstracting information. Early work described induction heads as the key mechanism transferring patterns from previous tokens to later ones. By attending from a token to its predecessor and reproducing its associated completion, these heads effectively implement next-token prediction conditioned on contextual examples.

Further investigation has refined this understanding. Researchers discovered that not all heads contribute equally: some act as feature-value heads maintaining coherence in token embeddings, while others specialise in pattern completion or reasoning over longer spans. The 2025 analysis Which Attention Heads Matter for In-Context Learning extended this line of work, identifying functional clusters of heads that cooperate to form hierarchical circuits. Another study, Beyond Induction Heads, revealed that ICL emerges in stages—first through surface-level copying, then through relational abstraction distributed across deeper layers. These insights suggest that contextual reasoning results from multiple interacting sub-circuits rather than a single structural unit.

2.4 Architectural Advances and Long-Context Scaling

Technical progress in architecture has also transformed how ICL operates. Traditional transformers suffer from quadratic attention cost, which limits sequence length. To overcome this, new models employ sparse, sliding-window, or retrieval-based attention to extend effective context beyond 100 000 tokens. The In-Context Learning with Long-Context Models and related Many-Shot ICL studies tested models under thousand-shot conditions [7] and found that performance stabilises or even improves when examples are semantically coherent and attention is hierarchically structured. Long-context designs reveal that scaling context does not simply increase computational load—it changes the model's capacity to reason across document-level dependencies.

Moreover, long-context transformers demonstrate that memory segmentation is crucial. Instead of attending to all tokens uniformly, the model benefits from dividing the sequence into logical blocks and propagating summary tokens upward. This structure mirrors hierarchical reasoning, where local adaptation informs global inference. Such results connect hardware-efficient design with cognitive-like processing, marking a step toward more interpretable contextual computation.

2.5 Adaptive Demonstration Retrieval and Prompt Automation

As ICL matured, prompt design evolved from manual art to automated strategy. Early prompting relied on human-curated examples, but this approach proved inconsistent. Retrieval-based methods now automatically select demonstrations aligned with the query distribution. Algorithms compute embedding similarity or gradient matching to choose samples that best represent the target task. For instance, gradient-matching retrieval minimises the divergence between contextual updates and full fine-tuning gradients, ensuring that in-context adaptation mirrors actual learning dynamics. Bandit-style approaches, as

proposed in the CASE framework, treat demonstration selection as an exploration problem, balancing novelty and relevance.

These techniques improve both stability and efficiency. By turning demonstration selection into a data-driven process, they reveal that much of ICL's success depends not on the model's inherent intelligence but on the organisation of its context. When examples are chosen adaptively, large models act less like static memory banks and more like dynamic reasoning systems capable of reconfiguring attention based on input structure.

2.6 Multimodal and Cross-Domain Extensions

The boundaries of ICL extend beyond text-only applications. Recent work introduced multimodal prompts that integrate visual, auditory, and symbolic information. The VL-ICL Bench evaluated image-text settings and showed that contextual learning generalises partially across modalities. Alignment between embeddings is critical: when feature spaces mismatch, performance collapses. Additional experiments, such as ConText, applied ICL principles to vision tasks like text removal and segmentation, illustrating that transformers can learn spatial rules from visual demonstrations. These results imply that contextual reasoning is not exclusive to language but a general property of attention-based architectures exposed to structured data.

Multimodal ICL introduces new research opportunities but also amplifies complexity. Differences in token density, noise, and spatial hierarchy create imbalance across channels, making stability difficult to maintain. Nevertheless, the trend toward cross-domain adaptation suggests that contextual inference could unify diverse modalities under shared computational principles.

In summary, theoretical and technological progress portrays ICL as a multi-layered phenomenon: probabilistic at the conceptual level, algorithmic in its dynamics, and mechanistic in its implementation. Advances in architecture, retrieval strategy, and multimodal integration further demonstrate that contextual learning is not an anomaly but a structural consequence of scale and attention. These frameworks together form the intellectual foundation for the empirical analyses that follow in Section 3.

3 Research Findings and Analysis

3.1 Dependence on Data Regularity

A growing body of work demonstrates that the structure of pretraining data critically shapes ICL performance. When models are trained on corpora exhibiting thematic consistency and long-range coherence, they develop stronger contextual reasoning. Conversely, when document order or topic continuity is disrupted, their ability to infer relationships from few examples diminishes. Empirical studies have shown that data distributional regularities drive emergent ICL behaviour [8].

This evidence suggests that ICL emerges not only from architecture but also from the statistical topology of language data.

More recent studies expanded this observation beyond language. When transformers are pretrained on structured multimodal datasets, such as paired text–image corpora, they show cross-modal contextual inference. The same regularity principle appears to apply: models learn to recognise and extend relational structures when input domains share consistent mappings. This finding implies that contextual reasoning might reflect a universal inductive bias toward pattern continuation.

3.2 Sensitivity to Prompt Structure and Demonstration Order

Prompt composition exerts a disproportionate influence on ICL outcomes. The order of examples, their spacing, and the formatting of input–output pairs can shift performance by large margins. Controlled ablation experiments revealed that rearranging demonstrations or altering separators changes accuracy even when semantic information remains constant. This instability underscores the absence of a canonical prompting protocol. Retrieval-based strategies mitigate variance by aligning demonstrations with query embeddings, improving reproducibility. The KATE retrieval system demonstrated that choosing examples close to the target distribution yields consistent gains.

However, sensitivity persists even under automated retrieval. When examples contain conflicting labels or stylistic inconsistencies, performance degrades sharply. Researchers attribute this to the transformer's reliance on surface-level regularities: attention patterns track syntactic similarity more than semantic content. As a result, ICL sometimes reproduces superficial correlations rather than underlying rules. The development of Gradient-Matching ICL offered partial relief by selecting examples whose gradient directions approximate those of ideal updates, but full robustness remains elusive.

3.3 Context Length, Scaling, and Attention Decay

Scaling behaviour has become a key focus of recent research. As the number of in-context examples grows, performance follows a non-linear curve—initial improvement, saturation, and eventual decline. This "many-shot" scaling law mirrors the Bayesian saturation predicted in earlier theory. The In-Context Learning with Long-Context Models experiments confirmed that performance stabilises when prompts exceed hundreds of examples, provided that context windows are hierarchically organised. Without segmentation, attention weight diffuses and middle tokens receive minimal focus, leading to the well-known "lost-in-the-middle" effect [9].

Studies on The Implicit Dynamics of ICL provided mechanistic evidence for these trends. Activation analyses revealed that each example updates internal representations that serve as temporary priors for subsequent tokens. As the number of examples increases, updates accumulate until representational space saturates, causing diminishing returns. These findings link neural dynamics directly to probabilistic inference curves. They also imply that scaling limitations arise not from model size alone but from the intrinsic capacity of attention mechanisms to store and update contextual information.

3.4 Multimodal Behaviour and Cross-Domain Adaptation

The extension of ICL into multimodal environments has exposed both its versatility and fragility. The VL-ICL Bench provided systematic evaluation across image captioning, visual question answering, and text–image matching tasks. Results showed that multimodal ICL can generalise within domains but struggles across domain boundaries. Performance is highly dependent on alignment between visual and textual feature spaces; even minor mismatches in embedding scale or normalisation degrade accuracy. Models often overfit to dominant modalities, relying on textual cues while underutilising visual information.

The ConText framework introduced in 2025 further examined visual ICL by applying contextual prompting to text-removal and segmentation tasks. The system achieved significant improvements when example images shared similar spatial layouts, supporting the claim that ICL depends on representational homogeneity. These experiments highlight that multimodal contextual reasoning is feasible but unstable. Cross-domain coherence—ensuring balanced interaction between modalities—appears to be the decisive factor for successful adaptation.

3.5 Length Biases and Adaptive Unlearning

Beyond performance metrics, researchers have begun investigating the behavioural dynamics of ICL. Schoch and Ji reported that models display length biases—a tendency to prefer shorter or longer outputs depending on prompt distribution. Interestingly, the same models can "unlearn" these biases within a single context when provided counterexamples. This reversible adjustment indicates that transformers perform a form of error-driven learning during inference. Such findings connect neural sequence modelling to cognitive processes of recalibration and adaptation. They also support the hypothesis that attention reweighting functions as a lightweight update rule parallel to human pattern correction.

3.6 Out-of-Distribution Generalisation

A persistent question is whether ICL generalises beyond the distributions observed during pretraining. Studies conducted at ICML and ICLR 2025 challenged this assumption by evaluating models on unseen task schemas. Li and colleagues demonstrated that ICL performance collapses when label semantics diverge from familiar categories, implying that contextual reasoning is largely interpolative rather than extrapolative. Similar patterns appeared in tasks requiring abstract rule transfer, where models replicated surface patterns without inferring deeper structure. Fang and collaborators proposed that enforcing invariance constraints during pretraining could enhance generalisation, though empirical validation remains limited. These findings indicate that ICL's generality may stem from pattern continuation rather than true task-level understanding.

3.7 Benchmarking and Evaluation Practices

As the literature expanded, inconsistencies in evaluation became a bottleneck. Metrics and datasets varied across studies, making replication difficult. ICLEval established a more systematic framework that separates pattern copying from rule induction and measures robustness under prompt perturbations. By introducing fine-grained subtests, it exposed weaknesses in tasks previously considered solved. Adoption of such benchmarks has already improved transparency, yet widespread standardisation is still emerging. Open access to model checkpoints and uniform reporting of prompt templates are now recognised as essential for reliable comparison.

In summary, empirical research has revealed that ICL depends on structured data, context-sensitive attention, and adaptive retrieval, while remaining vulnerable to noise, bias, and distributional shifts. The phenomenon embodies both strength and instability: it generalises swiftly within known regimes but struggles outside them. These results shape the central puzzle for ongoing research—understanding how contextual reasoning can become both flexible and reliable, and how theory can predict its behaviour across domains.

4 Current Limitations and Challenges

Despite the impressive breadth of progress, research on in-context learning remains fragmented. The field faces conceptual, methodological, and practical challenges that restrict deeper understanding and robust application. These limitations reveal not only the complexity of the phenomenon but also the immaturity of the theoretical tools currently used to explain it.

4.1 Fragmented Theoretical Foundations

The first and perhaps most persistent difficulty lies in the absence of a unified theoretical account. Bayesian, algorithmic, and mechanistic frameworks each provide partial

explanations but remain difficult to reconcile. Some evidence supports probabilistic inference as the underlying principle, whereas other analyses show algorithmic simulation dynamics that contradict purely statistical interpretations. Mechanistic studies, focusing on attention heads and circuit behaviour, describe micro-level operations without clarifying how they correspond to abstract inference. The resulting landscape resembles a set of overlapping but disconnected models.

This fragmentation hinders progress in two ways. It prevents cumulative theory-building, as results derived from one interpretation often fail to generalise under another, and it complicates evaluation design because metrics rooted in one paradigm may not capture the essence of others. Without a shared mathematical formulation, even strong empirical patterns remain conceptually ambiguous. Scholars have proposed hybrid frameworks that embed Bayesian inference within algorithmic simulation, yet these remain exploratory. Establishing a coherent foundation is now considered a prerequisite for the next phase of ICL research.

4.2 Reproducibility and Experimental Instability

A second obstacle concerns reproducibility. Experimental outcomes in ICL often vary dramatically with minor changes in prompt templates, sampling seeds, or tokenisation schemes. Reproduction across different model checkpoints yields inconsistent results even when nominal conditions match. This instability arises partly from the sensitivity of transformers to contextual distribution and partly from undocumented implementation details in public models. Reproducing large-scale experiments is also expensive, requiring computational resources that few institutions possess.

The absence of shared benchmark standards exacerbates the problem. Unlike traditional supervised tasks with fixed datasets and metrics, ICL evaluation involves free-form prompts and varying numbers of demonstrations. Minor wording differences can shift accuracy by several percentage points, making it difficult to draw meaningful comparisons across studies [10]. Although new frameworks like ICLEval attempt to impose structure, universal adoption remains limited. Large community benchmarks such as BIG-bench further reveal evaluation variance across tasks and model scales. Until evaluation procedures become more standardised, reproducibility will continue to constrain empirical progress.

4.3 Robustness, Sensitivity, and Bias

ICL's most celebrated feature—its ability to adapt quickly—also exposes its fragility. Models show extreme sensitivity to input order, domain shifts, and adversarial noise [11]. Rubin and colleagues demonstrated that even well-designed retrieval systems can fail when the contextual examples contain contradictory cues. Small variations in punctuation or case formatting sometimes produce disproportionate errors. Such brittleness undermines confidence in the stability of contextual reasoning.

Moreover, contextual adaptation often amplifies hidden biases embedded in pretraining data. Because demonstrations act as immediate conditioning signals, any skew in their composition can propagate directly to outputs. The BiasICL study revealed that prompts containing gender- or ethnicity-coded names influence model predictions even in neutral tasks. This problem extends beyond fairness concerns to reliability itself: a model that mirrors contextual bias cannot be trusted to perform consistent reasoning. Developing bias-resilient prompting or counter-conditioning methods is therefore a priority for both ethical and technical reasons.

4.4 Computational and Resource Constraints

The scaling requirements of ICL poses another challenge. Many-shot contexts demand long-sequence processing with quadratic attention cost. Although recent architectures have

introduced sparse or memory-augmented attention, inference over 100k-token prompts remains computationally prohibitive. Empirically, LongICLBench shows that long-context LLMs still falter on realistic many-shot settings with large label spaces.

Storage and latency grow rapidly with sequence length, limiting deployment in real-world applications [12]. Techniques like retrieval-augmented generation alleviate some burden but introduce new complexity, including cache management and semantic drift between retrieved and original contexts.

Furthermore, the lack of parameter updates in ICL means that contextual information must be reprocessed from scratch for each input. This leads to redundancy and inefficiency compared to fine-tuning approaches that internalise task structure. Designing systems that balance the flexibility of in-context learning with the persistence of parametric adaptation remains an open engineering problem.

4.5 Multimodal Complexity and Cross-Domain Fragility

Although multimodal in-context learning (ICL) has become an important and promising direction, it also brings distinctive challenges. Text–image models often show unstable alignment between the two modalities, and even small embedding shifts may lead to errors in cross-modal reasoning. Results from the VL-ICL Bench indicate that multimodal prompts are highly sensitive to noise from either the visual or textual channel. Practical VLM systems (e.g., MMICL) further expose instability and alignment issues under multi-modal ICL settings. Depending on the preprocessing method, visual features can dominate textual ones or the opposite may occur. When ICL is extended to video, audio, or embodied environments, the situation becomes more complex because temporal and causal relationships are introduced, which are rarely present in text-only models.

Generalisation across different domains remains a major difficulty. When tasks differ from those used in pretraining, the ability of the model to perform contextual reasoning drops sharply. This suggests that current systems mainly perform interpolation within known distributions rather than true abstract reasoning. Although fine-tuned multimodal models have achieved limited improvement, their performance still relies heavily on repeating learned patterns instead of understanding general principles. Bridging this gap will require both new architectural designs and richer training objectives that encourage models to learn deeper structural representations.

5 Future Research Directions

Research on in-context learning (ICL) is now moving into a phase that requires integration rather than simple expansion. In recent years, models have shown an impressive ability to develop contextual reasoning, but the internal mechanisms behind this behaviour are still not fully understood. To establish a solid and consistent foundation, future studies should aim to connect different theoretical perspectives, improve evaluation methods, and design new model architectures that make contextual learning more stable, transparent, and interpretable.

5.1 Toward Unified Theoretical Models

A central goal for future work is the integration of existing theoretical frameworks into a single formal system. Bayesian inference explains statistical efficiency, algorithmic simulation describes dynamic behaviour, and mechanistic analysis exposes structural circuits. However, these perspectives often operate at different levels of abstraction. Integrating them into a unified model could clarify how probabilistic reasoning emerges from deterministic computation [13]. One promising direction is to model transformer activations as an energy landscape where inference corresponds to local optimisation steps—an approach already

explored in preliminary work on Energy-Based ICL. Such hybrid formulations might provide a mathematical bridge linking the probabilistic and mechanistic layers of explanation.

A second theoretical priority involves the relationship between ICL and human learning. Comparing contextual reasoning with cognitive processes of analogy, memory recall, and pattern completion could reveal shared inductive biases. Empirical research on neuro-symbolic integration suggests that transformers approximate associative memory systems capable of reconstructing knowledge from sparse cues. Establishing formal analogies between neural inference and cognitive adaptation may yield insight into the limits of algorithmic learning.

5.2 Controlled and Reproducible Experimentation

The instability of ICL results calls for controlled environments where experiments can be replicated precisely. Synthetic data generation offers one solution. By constructing datasets with explicit structure and known rules, researchers can isolate the conditions under which ICL emerges. The SyntheticBench project demonstrated that rule-governed corpora produce predictable scaling patterns absent in natural text. Expanding such controlled benchmarks will help distinguish genuine learning dynamics from artefacts of pretraining data.

Open-sourcing large-scale evaluation suites is equally important. Transparent sharing of prompts, model checkpoints, and scoring scripts should become standard practice [14]. Future competitions might mirror the structure of reinforcement learning challenges, where fixed evaluation environments enable direct comparison across architectures. These practices not only enhance reproducibility but also encourage theoretical convergence by forcing diverse models to perform under consistent conditions. Recent analyses also examine gradient-based criteria as proxies for ICL dynamics.

5.3 Robust and Adaptive Demonstration Selection

Another priority lies in automating the process of contextual example selection. Current retrieval methods optimise similarity metrics but often ignore causal dependencies among examples. Reinforcement learning could be employed to train agents that adaptively curate demonstrations based on feedback from model outputs. Such systems would treat prompting as an interactive decision process rather than a static configuration. Integration with active learning frameworks may further reduce redundancy by identifying the minimal subset of examples that yield maximal adaptation.

Developing context compression techniques will complement this goal. Hierarchical summarisation or latent tokenisation can retain essential information while reducing computational load. When combined with adaptive retrieval, compressed prompting could scale ICL to ultra-long sequences without prohibitive memory cost.

5.4 Multimodal and Interactive ICL

Expanding ICL into multimodal and interactive domains remains a frontier. The 2025 Cross-Modal ICL study revealed that transformers can perform relational reasoning when text and vision are jointly encoded within the same context [15]. Extending this principle to speech, video, and embodied action could unify perception and reasoning under a single contextual paradigm. Interactive ICL, where the model modifies its prompt through dialogue or feedback, may transform static inference into dynamic problem-solving. Such systems would not merely interpret context but construct it, iteratively refining their own learning environment.

Interactive multimodal ICL will also require new evaluation methods. Static accuracy scores cannot capture the temporal adaptation that occurs through feedback. Metrics based on information gain or regret reduction may provide more realistic measures of progress.

These developments could redefine contextual learning as a general framework for adaptation across cognitive modalities.

6 Conclusion

In-context learning has evolved from an incidental observation in early language models into one of the most revealing phenomena in contemporary artificial intelligence. It exposes how large neural systems can display adaptive behaviour without changing their parameters, bridging the gap between static memory and dynamic reasoning. Across the literature, researchers have approached ICL from probabilistic, algorithmic, and mechanistic angles, each uncovering a facet of its operation. These perspectives collectively show that contextual learning emerges when scale, structure, and data regularity converge to produce implicit inference inside the model's forward computation.

Empirical work has broadened the understanding of where and when ICL succeeds. Models perform best when pretraining data exhibit coherent patterns, prompts are semantically aligned, and context length matches the model's attention capacity. Yet their sensitivity to order, noise, and distributional shift remains a defining limitation. Advances in retrieval automation, long-context architecture, and multimodal integration have expanded the scope of contextual reasoning, but stability and reproducibility continue to challenge practical deployment.

The theoretical landscape remains fragmented. The field now requires synthesis—a unifying formulation that connects statistical inference, algorithmic simulation, and attention-based mechanisms within a single framework. Future research must also integrate ethical, interpretability, and resource considerations to ensure that contextual adaptation remains transparent and equitable. As work on controlled benchmarks, adaptive prompting, and multimodal reasoning advances, ICL is likely to transform from a heuristic curiosity into a general principle of machine learning.

In the broader trajectory of artificial intelligence, ICL stands as both a capability and a clue. It suggests that intelligence may not depend on continual parameter updates but on the flexible recombination of knowledge in context. Understanding this process will not only refine the design of large language models but also illuminate fundamental questions about how systems—biological or artificial—learn to think from what they already know.

References

1. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal et al., Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901 (2020)
2. S. Xie, A. Raghunathan, P. Liang, T. Ma, An explanation of in-context learning as implicit Bayesian inference. Int. Conf. Learn. Represent. (ICLR 2022)
3. S. Garg, P. Nakkiran, D. Tsipras, What can transformers learn in-context? a case study of simple function classes. Adv. Neural Inf. Process. Syst. 35 (2022)
4. O. Rubin, D. Herzig, A. Berant, Learning to retrieve prompts for in-context learning. Conf. North Amer. Ch. Assoc. Comput. Linguist. (NAACL 2022)
5. O. Press, N. A. Smith, M. Lewis, Train short, test long: Attention with linear biases (ALiBi) enables length extrapolation. arXiv preprint arXiv:2108.12409 (2021)
6. J. von Oswald, V. V. Ramasesh, W. Czarnecki, J. Kirkpatrick, C. Clopath, Transformers learn in-context by gradient descent. Nat. Mach. Intell. 5, 850–862 (2023)
7. R. Agarwal, S. Kundu, H. Shi, J. Zhao, Many-shot in-context learning. Adv. Neural Inf. Process. Syst. (NeurIPS 2024 Spotlight)

8. M. Hahn, N. Goyal, A theory of emergent in-context learning as implicit structure induction. arXiv preprint arXiv:2303.07971 (2023)
9. H. Liu, S. Liu, Z. Liang, Y. Ruan, J. Wei, D. Zhou, Lost in the middle: how language models use long contexts. Trans. Assoc. Comput. Linguist. 12, 157–173 (2024)
10. R. Schaeffer, B. Miranda, S. Koyejo, Are emergent abilities of large language models a mirage. arXiv preprint arXiv:2304.15004 (2023)
11. K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, M. Fritz, Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. arXiv:2302.12173 (2023)
12. Y. Li, G. Zhang, Q. D. Do, X. Yue, W. Chen, LongICLBench: Benchmarking long-context in-context learning. arXiv preprint arXiv:2404.02060 (2024)
13. F. Falck, Z. Wang, C.C. Holmes, Is in-context learning in large language models Bayesian? A martingale perspective. Proc. 41st Int. Conf. Mach. Learn. (ICML 2024), PMLR 235, 12784–12805 (2024)
14. A. Srivastava, A. Rastogi, A. Misra, et al., Beyond the Imitation Game: Quantifying and extrapolating behaviors of language models. arXiv preprint arXiv:2206.04615 (2022).
15. J.-B. Alayrac, J. Donahue, P. Luc et al., Flamingo: a visual language model for few-shot learning. Adv. Neural Inf. Process. Syst. (2022)