

The Mystery of In-Context Learning: A Comprehensive Survey on Interpretation and Analysis

Yuxiang Zhou[♡], Jiazheng Li[♡], Yanzheng Xiang[♡], Hanqi Yan[♡], Lin Gui[♡], Yulan He^{♡♦}

[♡]Department of Informatics, King’s College London [♦]The Alan Turing Institute

{yuxiang.zhou,jiazheng.li,yanzheng.xiang}@kcl.ac.uk

{hanqi.yan,lin.1.gui,yulan.he}@kcl.ac.uk

Abstract

Understanding in-context learning (ICL) capability that enables large language models (LLMs) to excel in proficiency through demonstration examples is of utmost importance. This importance stems not only from the better utilization of this capability across various tasks, but also from the proactive identification and mitigation of potential risks, including concerns regarding truthfulness, bias, and toxicity, that may arise alongside the capability. In this paper, we present a thorough survey on the interpretation and analysis of in-context learning. First, we provide a concise introduction to the background and definition of in-context learning. Then, we give an overview of advancements from two perspectives: 1) the theoretical perspective, emphasizing studies on mechanistic interpretability and delving into the mathematical foundations behind ICL; and 2) the empirical perspective, concerning studies that empirically analyze factors associated with ICL. We conclude by discussing open questions and the challenges encountered and, by suggesting potential avenues for future research. We believe that our work establishes the basis for further exploration into the interpretation of in-context learning. To aid this effort, we have created a repository¹ containing resources that will be continually updated.

1 Introduction

The concept of in-context learning (ICL) was originally introduced by Brown et al. (2020), defined as ‘*the model is conditioned on a natural language instruction and/or a few demonstrations of the task and is then expected to complete further instances of the task simply by predicting what comes next*’. ICL is receiving increasing attention due to its remarkable adaptability and parameter-free nature. As shown in Figure 1, LLMs such as GPT-4 (OpenAI, 2024), Llama3 (AI@Meta, 2024), and

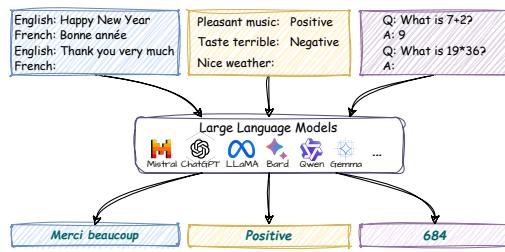


Figure 1: Illustration of In-context Learning.

Qwen2 (QwenTeam, 2024) have exhibited proficiency across various tasks, such as machine translation, sentiment analysis, and question answering, with a minimal set of task-oriented examples, all without re-training. While ICL has been dominantly deployed in the Natural Language Processing (NLP) community, our understanding of ICL remains limited. Recently, an increasing number of studies have attempted to interpret and analyze ICL. Garg et al. (2022), Dai et al. (2023), and Akyürek et al. (2023) explained ICL through the lens of linear regression formulation. Xie et al. (2022), Wang et al. (2023b), and Hahn and Goyal (2023) provided an interpretation of ICL rooted in latent variable models. Meanwhile, a distinct line of research has aimed to understand the influential factors affecting ICL through empirical analyses. Min et al. (2022), Wei et al. (2023), Wang et al. (2023a), and Yoo et al. (2022) demonstrated that the ICL performance is influenced by task-specific characteristics and multiple facets of ICL instances, including quantities, order, and flipped labels. Consequently, it is essential to systematically categorize and summarize these studies, not only for a deeper understanding and more effective utilization of ICL across various tasks, but also to assist in anticipating and mitigating potential risks. These risks encompass concerns related to truthfulness, bias, and toxicity, that may arise alongside ICL.

In this paper, we present a thorough and organized survey of the research on the interpretation

¹<https://github.com/zyxnlp/ICL-Interpretation-Analysis-Resources>

and analysis of ICL. First, we provide a brief introduction of the background and offer the definition of ICL. Then, we present a comprehensive overview of advancements, from two distinct viewpoints: 1) the *theoretical perspective*, encapsulating studies focused on mechanistic interpretability and mathematical investigations into the foundations of ICL; and 2) the *empirical perspective*, pertaining to studies that prioritize empirical analysis by probing factors associated with ICL. In conclusion, we highlight the existing challenges and suggest potential avenues for further research.

2 Background and Notation

In this section, we define the ICL paradigm using the following notations. A *task* \mathcal{T}_D consists of two components: a demonstration space \mathcal{D} , which encompasses all possible demonstrations, and a joint probability distribution $P_{(X,Y)}$. A *task demonstration* $D = \{(x_i, y_i)\}_{i=1}^n \in \mathcal{D}$ contains n example pairs sampled from the joint distribution. In NLP, these example pairs could be Question-Answering pairs for QA tasks, parallel text for machine translation tasks, or sentence-label pairs for text classification tasks. For example, for machine translation, an example pair (x, y) could be (*English: Happy New Year, French: Bonne année*). In contrast to traditional supervised learning approaches which aim to generalize from a fixed training dataset to predict future instances, ICL leverages continual exposure to demonstration examples and is guided by *task query* to adapt the model dynamically to different contexts. In this survey, we denote the *query* as $\mathcal{T}_Q = \{\mathcal{Q}, P_Q\}$, consisting a query space \mathcal{Q} and marginal distribution P_Q . A *task query* $Q = \{q_j\}_{j=1}^m \in \mathcal{Q}$ contains m instances sampled from the marginal distribution. For example, q could be “*English: Thank you French:*” in the machine translation task. Additionally, we define $A = \{a_j\}_{j=1}^m$ which represents the gold label for Q . Let an LLM be defined as a function F_θ pre-trained on large-scale text corpora. ICL can be defined as follows:

Definition (In-Context Learning) *In the context of a task query Q , in-context learning refers to the capability of F_θ to predict the correct answer A , conditioned on a task demonstration D .*

Based on the above definition, the process of ICL can be formally expressed as follows:

$$\begin{aligned} D &\sim P_{(X,Y)} & Q &\sim P_Q \\ \hat{A} &\leftarrow F_\theta(D, Q) \end{aligned} \quad (1)$$

The performance of ICL can be measured by:

$$\mathcal{S} = \mathbb{E}_{D,Q,A}[M(\hat{A}, A)] \quad (2)$$

\hat{A} denotes the model-predicted output, and M is an evaluation metric chosen based on the *task query* Q and its gold label A .

Based on our definition, we organize the existing literature on the interpretation and analysis of ICL into theoretical and empirical perspectives, as summarized in Table 1. Researchers in the theoretical category focus on interpreting the connections among F_θ , D , Q and A to explain the fundamental mechanism behind the ICL process. In contrast, those in the empirical category primarily centre on analyzing the relationship between performance \mathcal{S} and the characteristics of the demonstration D .

3 Theoretical Interpretation of ICL

3.1 Mechanistic Interpretability

With the goal of reverse-engineering components of LLMs models into more understandable algorithms, Elhage et al. (2021) developed a mathematical framework to decompose operations within Transformers (Vaswani et al., 2017) for explaining ICL. They discovered that one-layer attention-only Transformers can perform very primitive ICL by assessing patterns (e.g., bigrams) from parameters. Furthermore, they found that two-layer Transformers manifest a more general ICL capability using *induction head*. The induction heads are composed of attention heads that implement an algorithm to complete token sequences by copying and generating sequences that have occurred before. Building on this foundation, Olsson et al. (2022) later investigated the internal structures responsible for ICL by analyzing the induction head in a full Transformer architecture. They implemented circuits consisting of induction head and *previous token head*, which copies information from one token to its successor. Their study revealed a phase change occurring early in the training of LLMs of various sizes and found that circuits play a crucial role in implementing most ICL in LLMs. One pivotal insight from Olsson et al. (2022) presented comprehensive arguments supporting their hypothesis that induction heads may serve as the primary mechanistic source of ICL in a significant portion of LLMs, particularly those based on transformer architectures.

Later, Edelman et al. (2024) extended Olsson et al. (2022) by introducing a Markov Chain sequence modeling task, where demonstrations are sampled from a Markov chain. They showed that

Work	Key Words	Models	Tasks
Theoretical Perspective			
(Elhage et al., 2021)	Mechanistic Interpretability	Transformer [†]	-
(Olsson et al., 2022)	Mechanistic Interpretability	Transformer	-
(Edelman et al., 2024)	Mechanistic Interpretability	Transformer [†]	Markov Chain modeling
(Swaminathan et al., 2023)	Mechanistic Interpretability	Transformer	Next token prediction
(Todd et al., 2024)	Mechanistic Interpretability	Llama 2, GPT [‡]	Antonym, etc.
(Bai et al., 2023)	Mechanistic Interpretability	Transformer [†]	Regression
(Garg et al., 2022)	Regression Function Learning	Transformer [†]	Regression
(Li et al., 2023b)	Regression Function Learning	Transformer [†]	Regression
(Li et al., 2023a)	Regression Function Learning	Transformer [†]	Regression
(Akyürek et al., 2023)	Regression Function Learning	Transformer [†]	Regression
(Guo et al., 2024)	Regression Function Learning	GPT [‡]	Representation Regression
(Dai et al., 2023)	Gradient Descent, Meta-Optimization	GPT [‡]	Classification
(von Oswald et al., 2023a)	Gradient Descent, Meta-Optimization	Transformer [†]	Regression
(von Oswald et al., 2023b)	Gradient Descent, Meta-Optimization	Transformer [†]	Regression
(Deutch et al., 2024)	Gradient Descent, Meta-Optimization	Transformer [†]	Classification
(Shen et al., 2024)	Gradient Descent, Meta-Optimization	LLaMa, GPT [‡]	Classification
(Fu et al., 2024)	Gradient Descent, Meta-Optimization	Transformer, LSTM	Regression
(Xie et al., 2022)	Bayesian Inference	Transformer [†] , LSTM	Next token prediction
(Wang et al., 2023b)	Bayesian Inference	GPT [‡]	Classification
(Wies et al., 2023)	Bayesian Inference	-	-
(Jiang, 2023)	Bayesian Inference	GPT [†]	Sythetic Generation
(Zhang et al., 2023b)	Bayesian Inference	Transformer [†]	-
(Panwar et al., 2024)	Bayesian Inference	Transformer	Regression
(Jeon et al., 2024)	Bayesian Inference	Transformer [†]	Regression
(Bigelow et al., 2024)	Bayesian Inference	GPT [‡]	Sequence generation
Empirical Perspective			
(Shin et al., 2022)	DATA Domain	GPT-3	Classification, etc.
(Han et al., 2023)	DATA Domain, DATA Distribution	GPT [‡]	Classification
(Raventós et al., 2023)	Task Diversity	GPT-2	Regression
(Razeghi et al., 2022)	DATA Term frequency	GPT [‡]	Reasoning
(Kandpal et al., 2023)	DATA Term frequency	GPT-3 [‡]	Question Answering
(Chan et al., 2022)	DATA Distribution	Transformer	Classification
(Yadlowsky et al., 2023)	DATA Distribution	GPT-2 [‡]	Regression
(Hendel et al., 2023)	Task Diversity	LLaMA	Translation, etc.
(Wei et al., 2022b)	Model Scale	GPT-3 [‡]	Classification
(Schaeffer et al., 2023)	Model Scale, Evaluation Metric	GPT-3 [‡]	Classification
(Tay et al., 2023)	Pre-training Objective	UL2 [‡]	Classification, etc.
(Kirsch et al., 2024)	Model Architecture	Transformer [†]	Classification
(Singh et al., 2023)	Model Architecture	LLaMA [†]	Synthetic generation
(Yousefi et al., 2024)	Embeddings	Llama, Vicuna	Regression
(Akyürek et al., 2024)	Model Architecture	Transformer, LSTM, etc.	Language learning
(Lu et al., 2022)	Demonstration Order	GPT [‡]	Classification
(Liu et al., 2024)	Demonstration Order	GPT-3.5 [‡]	Question Answering
(Zhao et al., 2021)	Demonstration	GPT [‡]	Classification, IE, IR
(Liu et al., 2022)	Demonstration Order	GPT-3	Classification, QA, etc.
(Min et al., 2022)	Input-Label Mapping	GPT [‡]	Classification, etc.
(Yoo et al., 2022)	Input-Label Mapping	GPT [‡]	Classification
(Wei et al., 2023)	Input-Label Mapping	GPT-3 [‡]	Classification, QA, etc.
(Pan et al., 2023)	Input-Label Mapping	GPT-3 [‡]	Classification, etc.
(Lin and Lee, 2024)	Input-Label Mapping	Llama, LSTM, etc.	Classification, etc.
(Kossen et al., 2024)	Input-Label Mapping	LLaMA [‡]	Classification
(Tang et al., 2023)	Input-Label Mapping	GPT [‡]	Classification
(Si et al., 2023)	Input-Label Mapping	GPT-3 [‡]	Classification, QA, etc.
(Wang et al., 2023a)	Input-Label Mapping	GPT2-XL	Classification

Table 1: Summary of research studies on the interpretation of ICL. QA stands for Question Answering. DATA refers to pre-training data. The symbol [†] denotes specifically designed models. The [‡] denotes that either various models or models from different families were used.

transformers learn statistical induction heads to approach the Bayes-optimal by computing the correct posterior probability of the next token, given all previous occurrences of the prior token. On the contrary, Swaminathan et al. (2023) adopted an alternative approach to elucidate the principles underpinning the mechanisms of ICL by studying clone-structured causal graphs (CSCGs), a sequence-learning model. They demonstrated that LLMs and CSCGs share similar mechanisms underlying ICL, which consist of: (a) learning template circuits for pattern completion, (b) retrieving

relevant templates, and (c) rebinding novel tokens within the templates. Following on, Todd et al. (2024) measured causal mediators across a distribution of different tasks to identify the information transported by the attention heads in ICL. They discovered *function vectors* (FVs), a small number of attention heads transport information of the demonstrations within the Transformer’s hidden states during ICL. Bai et al. (2023) unveiled a general mechanism, *in-context algorithm selection*, to interpret ICL from a statistical viewpoint. They first demonstrated that transformers can implement

a broad class of standard machine learning algorithms, such as least squares, ridge regression, and Lasso. Then, they theoretically demonstrated that transformers can adaptively select from these algorithms to learn more complex ICL tasks.

3.2 Regression Function Learning

Several research studies posited that the emergence of ICL can be attributed to the intrinsic capability of models to approximate regression functions for a novel task query Q based on the task demonstration D . Garg et al. (2022) first formally defined ICL as a problem of learning functions and explored whether Transformers can be trained from scratch to learn simple and well-defined function classes, such as linear regression functions. To achieve this, they derived $D = \{(x_i, f(x_i))\}_{i=1}^n$ using a function f sampled from a linear function class and trained models to predict the function value $A = \{f(q_j)\}_{j=1}^m$ for the corresponding $Q = \{q_j\}_{j=1}^m$. Their empirical findings revealed that Transformers exhibited ICL abilities, as they manifested to “learn” previously unseen linear functions from examples, achieving an average error comparable to that of the optimal least squares estimator. Furthermore, Garg et al. (2022) demonstrated that ICL can be applied to more complex function classes, such as sparse linear functions and decision trees. They posited that the capability to learn a function class through ICL is an inherent property of the model F_θ . Later, Li et al. (2023b) extended Garg et al. (2022) to interpret ICL from a statistical perspective. They derived generalization bounds for ICL, considering two types of input examples: sequences that are independently and identically distributed (i.i.d.) and trajectories originating from a dynamical system. They established a multitask generalization rate for both types of examples, addressing temporal dependencies by associating generalization to algorithmic stability and framing ICL as an algorithm learning problem. They found that Transformers can implement near-optimal algorithms on classical regression problems and proved that self-attention has favourable stability properties by quantifying the influence of individual tokens on one another.

At the same time, Li et al. (2023a) took a further step from the work of (Garg et al., 2022) to gain a deeper understanding of the role of the softmax unit within the attention mechanism of LLMs. They sought to mathematically interpret ICL based on the softmax regression formulation represented as $\min_{\mathbf{x}} \|\langle \exp(A\mathbf{x}), \mathbf{1}_n \rangle^{-1} \exp(A\mathbf{x}) - \mathbf{b}\|_2$. They

established the upper bounds for data transformations effected by a single self-attention layer and theoretically demonstrated that LLMs perform ICL in a way that is highly similar to gradient descent (GD). Akyürek et al. (2023) took a different approach by delving into the process through which ICL learns linear functions, rather than analysing the types of functions that ICL can learn. Through an examination of the inductive biases and algorithmic attributes inherent in Transformer-based ICL, they discerned that ICL can be understood in algorithmic terms, and linear learners within the model may essentially rediscover standard estimation algorithms. They showed that trained in-context learners (ICLs) closely align with the predictors derived from GD, ridge regression, and exact least-squares regression. While the transition between these predictors varies with model depth and training set noise, they will converge to Bayesian estimators at large hidden sizes and depths. Additionally, Akyürek et al. (2023) provided a theoretical proof demonstrating that Transformers can implement learning algorithms for linear models using GD and closed-form ridge regression.

To understand ICL in a more complex scenario, Guo et al. (2024) studying ICL in the setting of *learning with representations*. They extended Garg et al. (2022) to consider a more general task demonstration $D = \{(x_i, \Phi^*(x_i))\}_{i=1}^n$ where the label depends on the instance x through representation function $\Phi^*(x)$ rather than f . They theoretically demonstrated the existence of Transformers that can implement an approximately optimal ICL algorithm with mild depth and size. Furthermore, Guo et al. (2024) trained Transformers to analyzing ICL with various mechanisms, such as copying (Olsønn et al., 2022) and *post-ICL representation selection* (Bai et al., 2023). Their empirical results showed that lower layers transform the dataset and upper layers perform linear ICL.

3.3 Gradient Descent & Meta-Optimization

In the realm of gradient descent (GD), Dai et al. (2023) adopted a perspective of viewing LLMs as meta-optimizers and interpreting ICL as a form of implicit fine-tuning. They first conducted a qualitative analysis of Transformer attention, representing it in a relaxed linear attention form, and identified a dual relationship between it and GD. Through a comparative analysis between ICL and explicit fine-tuning, Dai et al. (2023) interpreted ICL as a meta-optimization process. They further provided evi-

dence that the Transformer attention head possesses a dual nature similar to GD (Irie et al., 2022), where the optimizer produces meta-gradients based on the provided examples for ICL through forward computation. Concurrently, von Oswald et al. (2023a) also proposed a connection between the training of Transformers on auto-regressive objectives and gradient-based meta-learning formulations. They examined how Transformers define a loss function based on the given examples and the mechanisms by which Transformers assimilate knowledge using the gradients of this loss function. Their findings suggest that ICL may manifest as an emergent property, approximating gradient-based ICL within the forward pass of the model.

Following on, von Oswald et al. (2023b) extended von Oswald et al. (2023a) to uncover underlying gradient-based mesa-optimization algorithms driving model predictions by reverse-engineering autoregressive Transformers trained on sequence modeling tasks. They showed that these models exhibit ICL capability enabled by the *mesa-layer*, a novel attention layer that efficiently solves a least-squares optimization problem. On the contrary, Deutch et al. (2024) revisited the hypothesis that GD approximates ICL and highlighted core issues in the evaluation metrics and baselines of Dai et al. (2023). Their findings suggested a weak correlation between ICL and GD and revealed major discrepancies in the flow of information throughout the model between ICL and GD. In a similar vein, Shen et al. (2024) highlight that prior studies verify their hypothesis by training models explicitly for ICL, which differs from practical setups in real model training. They showed that the hand-constructed weights used in these studies possess properties that do not match those in real world scenarios. Furthermore, they observed that ICL and GD have different sensitivities to the order in which they observe demonstrations in natural settings. Fu et al. (2024) also presented evidence showing that Transformers learn to perform ICL by implementing a higher-order optimization rather than GD. They theoretically demonstrated that Transformer circuits can efficiently implement Newton’s method (Gautschi, 2011) and empirically showed that Transformers achieve the same convergence rate as Newton’s method while being exponentially faster than GD.

3.4 Bayesian Inference

Xie et al. (2022) were the first to interpret ICL through the lens of Bayesian inference, positing

that LLMs have the capability to perform implicit Bayesian inference via ICL. Specifically, they synthesized a small-scale dataset to examine how ICL emerges in Transformer models during pre-training on text with extended coherence. Their findings revealed that Transformers are capable of inferring latent concepts to generate coherent subsequent tokens during pre-training. Additionally, these models were shown to perform ICL by identifying a shared latent concept among examples during the inference process. Their theoretical analysis confirms that this phenomenon persists even when there is a distribution mismatch between the examples and the data used for pre-training, particularly in settings where the pre-training distribution is derived from a mixture of Hidden Markov Models (HMMs) (Baum and Petrie, 1966).

Following on, Wang et al. (2023b) and Wies et al. (2023) expanded the investigation of ICL by relaxing the assumptions made by Xie et al. (2022). Wies et al. (2023) assumed that there is a lower bound on the probability of any mixture component, alongside distinguishable downstream tasks with sufficient label margins. They proved that ICL is guaranteed to happen when the pre-training distribution is a mixture of downstream tasks. Wang et al. (2023b) posited that ICL in LLMs essentially operates as a form of topic modeling that implicitly extracts task-relevant information from examples to aid in inference. They characterized the data generation process using a causal graph and imposed no constraints on the distribution or quantity of samples. Their theoretical investigations revealed that ICL can approximate the Bayes optimal predictor when a finite number of samples are chosen based on the latent concept variable. At the same time, Jiang (2023) also introduced a novel latent space theory extending the idea of Xie et al. (2022) to explain ICL in LLMs. Instead of focusing on specific data distributions generated by HMMs, they delved into general sparse data distributions and employed LLMs as a universal density approximator for the marginal distribution, allowing them to probe these sparse structures more broadly. They also demonstrated that ICL in LLMs can be ascribed to Bayesian inference operating on the broader sparse joint distribution of languages.

To shed light on the significance of the attention mechanism for ICL from a Bayesian view, Zhang et al. (2023b) defined ICL as the task of predicting a response that aligns with a given covariate based on examples derived from a latent variable model.

They demonstrated that certain attention mechanisms converge towards the conventional softmax attention as the number of examples goes to infinity. These attentions, due to their encoding of Bayesian Model Averaging (BMA) algorithm (Wasserman, 2000) within their structure, empower the Transformer model to perform ICL. Panwar et al. (2024) extended the previous setup in (Garg et al., 2022; Akyürek et al., 2023) by testing the Bayesian hypothesis for ICL over both linear and nonlinear function families. They found that Transformers mimic the Bayesian predictor to perform ICL, including generalizing new function classes not seen during pre-training. Furthermore, they demonstrated that the simplicity bias in ICL arises from the pre-training distribution and provided empirical evidence that Transformers solve mixtures of tasks, suggesting the Bayesian perspective could offer a unified understanding of ICL.

Concurrently, Jeon et al. (2024) took a different approach to revisit ICL as Bayesian inference without restrictive assumptions by introducing information-theoretic tool. They decomposed error for the Bayes optimal predictor into meta-learning error and intra-task error, and derived an in-context error upper bound of $\log(N)/\tau$ for the sparse mixture of transformers, where N is the number of mixture components and τ is the in-context length. To analyze ICL as Bayesian model selection in a practical setting, Bigelow et al. (2024) modeled latent concepts evoked in LLMs by different contexts. They adopted random binary sequences as context and examined dynamics of ICL by manipulating properties of the data, such as sequence length, based on the cognitive science of human randomness perception. They defined *subjective randomness* to investigate model behaviour and demonstrated sharp phase changes, where LLMs suddenly shift from one pattern of behaviour to another during text generation, supporting the theories of ICL as model selection.

4 Empirical Analysis of ICL

4.1 Pre-training Data

There has been controversy among researchers regarding the effect of pre-training data properties on the performance of ICL. To analyze the correlation between the domain of a corpus and ICL performance, Shin et al. (2022) evaluated LLMs pre-trained with subcorpora from diverse sources (e.g., blog, community website, news articles) within the HyperCLOVA corpus (Kim et al., 2021). They

found that the corpus sources significantly influenced ICL performance; however, a pre-training corpus aligned with the downstream task’s domain does not always guarantee competitive ICL performance. For instance, while LLMs trained on subcorpora from blog posts exhibited superior ICL performance, a model trained on news-related dataset did not sustain this superiority in ICL scenarios. Han et al. (2023) found that the effectiveness of pre-training data for ICL is not necessarily tied to its domain relevance to downstream tasks. By using MAUVAE score (Pillutla et al., 2021) to quantify information divergence between pre-training data and target task data, they observed that pre-training data containing low-frequency tokens and long-tail information tend to have greater impact on ICL. Conversely, Razeghi et al. (2022) and Kandpal et al. (2023) identified a positive correlation between ICL performance and the term frequency within pre-training data, suggesting the memorization capabilities can significantly influence ICL.

By the manipulation of pre-training tasks to be a uniform distribution, Raventós et al. (2023) identified a *diversity threshold* - quantified by the number of tasks seen during pre-training - that indicates the emergence of ICL. They empirically demonstrated that LLMs cannot perform a new task through ICL if the diversity of the pre-training task falls below the threshold. Chan et al. (2022) have identified three critical distributional properties of pre-training data that drive ICL: 1) the training data exhibits *bursty distribution* (Sarkar et al., 2005), where tokens appear in clusters rather than being uniformly distributed over time; 2) the marginal distribution across tokens is highly skewed, exhibiting a high prevalence of infrequently occurring classes, following a *ZipFian distribution* (Zipf, 1949); 3) the token meanings or interpretations are dynamic rather than fixed, where a token can have multiple interpretations (e.g., polysemy) or multiple tokens may correspond to the same interpretation.

Yadlowsky et al. (2023) explored the impact of pre-training data composition on the ability ICL. Building on the setup of Garg et al. (2022), they showed empirical evidence that the Transformers can perform model selection among pre-trained function classes during ICL with minimal additional cost. However, there was no evidence that the models were able to generalize beyond their pre-training data through ICL. To shed light on the mechanisms of ICL, Hendel et al. (2023) investigated the relationship between demonstrations and

the parameters of functions in certain hypothesis classes by examining the top tokens in the output distribution. They revealed that ICL functions by compressing training data into a task vector, which then guides Transformers to generate outputs.

4.2 Pre-training Model

The attributes of pre-training models have been shown to be significant factors affecting ICL. Wei et al. (2022b) focused on training computation (e.g., FLOPs (Hoffmann et al., 2022)) and model size(e.g., number of model parameters). They analyzed the emergent manner in which ICL manifests with the scaling of LLMs and highlighted the positive correlation between the model’s scale and ICL performance. On the contrary, Schaeffer et al. (2023) empirically analyzed the effect of the choice of evaluation metric on the emergence of ICL ability. By controlling for factors such as downstream task, model family, and model outputs, they found that this ability appears due to the choice of metric, rather than as a result of fundamental changes in models with scaling. At the same time, Tay et al. (2023) have suggested that the pre-training objective is a pivotal factor influencing ICL performance. They observed that continued pre-training with varied objectives enables robust ICL performance. Kirsch et al. (2024) have posited that aspects of model architecture, such as the dimension of the hidden size, play a more critical role than model size in the emergence of ICL.

Singh et al. (2023) suggested that the emergence of ICL should be viewed as a *transient* rather than a *persistent* phenomenon. They demonstrated that ICL may not persist as the model continues to be trained. By examining model sizes, pre-training data size, and domain, they found that ICL first emerges, then disappears, giving way to in-weights learning (IWL). Yousefi et al. (2024) introduced a neuroscience-inspired framework to empirically analyze how LLM embeddings and attention representations change following ICL. They measured the ratio of attention information over parameterized probing classifiers based on representational similarity analysis (RSA) and found a meaningful correlation between improvements in behaviour after ICL and changes in both embeddings and attention weights across LLM layers. Akyürek et al. (2024) proposed a novel dataset, REGBENC, to systematically study in context language learning (ICLL) in the setting of regular languages, the class of formal languages generated by finite automata (Hopcroft,

1971). They investigated ICLL in relation to model classes and mechanisms by examining essential features such as structured outputs, probabilistic predictions, and compositional reasoning about input data. They found that Transformers are the most efficient and can develop higher-order variants of induction heads (Olsson et al., 2022).

4.3 Demonstration Order

The order of the demonstrations has a significant impact on the ICL. Lu et al. (2022) designed demonstrations containing four samples with a balanced label distribution and conducted experiments involving all 24 possible permutations of sample orders. Their experimental results demonstrated that the ICL performance varies across different permutations and model sizes. In addition, they found that effective prompts are not transferable across models, indicating that the optimal order is model-dependent, and what works well for one model does not guarantee good results for the other models. Both Zhao et al. (2021) and Liu et al. (2024) identified a similar phenomenon where LLMs tend to repeat answers found at the end of provided demonstrations in ICL. Their results indicated that ICL performs optimally when the relevant information is positioned at the beginning or end of the demonstrations and the performance degraded when the LLMs are compelled to use information from the middle of the input. Liu et al. (2022) delved deeper and analyzed the underlying reasons for how the order of demonstration influences ICL. They proposed retrieving examples semantically similar to a test example for creating its demonstration and found that the demonstration order appears to be dependent on the specific dataset in use.

4.4 Input-Label Mapping

Some studies have explored the impact of input-label mappings on the performance of ICL in LLMs. Min et al. (2022) empirically showed that substituting the ground-truth labels in demonstrations with random ones results in a marginal performance decrease across various tasks. This indicates that ICL exhibits a low sensitivity to the accuracy of labels in the demonstration. This finding contradicts the conclusions in Yoo et al. (2022), Wei et al. (2023), and Kossen et al. (2024), who argued that LLMs rely significantly on accurate input-label mappings to perform ICL. For example, Yoo et al. (2022) highlighted that averaging performance across multiple datasets fails to ac-

curately reflect the insensitivity observed within specific datasets. They introduced two novel metrics *label correctness sensitivity* and *ground-truth label effect ratio*, to extensively quantify the impact of ground-truth labels on ICL performance. Their empirical findings confirmed that ground-truth label significantly influences ICL, revealing a strong correlation between sensitivity to label correctness and the complexity of the downstream task.

To investigate the effect of semantic priors and input-label mappings on ICL, Wei et al. (2023) discovered that LLMs can prioritize input-label mappings from demonstrations over pre-training semantic priors, leading LLMs to drop below random guessing when all the labels in the demonstrations are flipped. Additionally, their research indicated that smaller models predominantly utilize the semantic meanings of labels rather than the input-label mappings presented in ICL demonstrations. Pan et al. (2023) investigated how ICL leverages demonstrations by characterizing *task recognition* and *task learning* in LLMs. They reported that LLMs exhibit a significantly better ability to learn input-label mappings through ICL when compared to smaller models. Moreover, they observed that the ability of ICL to discern tasks from demonstrations does not substantially improve with increased model size. Following on, Lin and Lee (2024) extended Pan et al. (2023) by introducing multiple task groups and task-dependent input distributions to investigate the factor of pre-training data. They showed that ICL demonstrations with biased labels contain sufficient information to retrieve a correct pretrained task. Tang et al. (2023) revealed that LLMs may rely on shortcuts in ICL demonstrations for downstream tasks. These shortcuts consist of spurious correlations between ICL examples and their associated labels.

Si et al. (2023) identified that LLMs exhibited feature bias when provided with *underspecified* ICL demonstrations in which two features are equally predictive of the labels. Their experiment suggested that interventions such as employing instructions or incorporating semantically relevant label words could effectively mitigate bias in ICL. To understand the underlying mechanism behind LLMs performing ICL from an information flow perspective, Wang et al. (2023a) discovered that semantic information is concentrated within the representations of label words in the shallow computation layers. Furthermore, they showed that the consolidated information within label words acts

as a reference for LLMs' final predictions, highlighting the importance of label words in ICL.

5 Open Questions

Despite ongoing endeavours to interpret and analyse ICL, we are still far from fully understanding it due to the open-ended nature of some questions. For example, while Elhage et al. (2021) and Olsson et al. (2022) contribute to our understanding of ICL by probing the internal architecture of LLMs, it is important to note that their findings represent initial steps towards the comprehensive reverse-engineering of LLMs. It becomes particularly intricate when dealing with LLMs characterized by complex structures comprising hundreds of layers and spanning billions to trillions of parameters. This complexity introduces significant challenges. Similarly, although studies have provided theoretical proofs and empirical evidence on the relation of ICL and regression function learning (Garg et al., 2022; Akyürek et al., 2023; Li et al., 2023a,b), gradient descent or meta-optimization (von Oswald et al., 2023a; Dai et al., 2023), and Bayesian inference (Xie et al., 2022; Wang et al., 2023b; Wies et al., 2023; Jiang, 2023; Zhang et al., 2023b), their conclusions are limited to simplified model architectures and controlled synthetic experimental settings. This raises the open question of whether these findings hold in the context of standard model architectures without approximations and if they can be directly applied to real-world scenarios.

On the other hand, there are contradictory findings from researchers analysing the factors affecting ICL. For instance, while Razeghi et al. (2022) and Kandpal et al. (2023) identified a positive relation between ICL performance and term frequency in pre-training data, Han et al. (2023) found that pre-training data with long-tail and rarely occurring tokens contribute more significantly to ICL. Additionally, while Min et al. (2022) suggested that ICL exhibits low sensitivity to labels in the demonstrations, other studies revealed that accurate input-label mappings play an important role in performing ICL (Yoo et al., 2022; Wei et al., 2023; Kossen et al., 2024). One of the core challenges in analysing ICL empirically lies in the necessity of controlling for numerous relevant variables, leading most existing conclusions to typically rely on correlations rather than causal relations. This raises open questions about the reliability of these findings and the extent to which confounding factors may influence the results.

6 Future Directions

Correlation vs Causation Most existing studies have interpreted ICL in LLMs primarily through correlational analyses, leading to biased conclusions that may not be broadly applicable. One core challenge lies in various underlying factors that interact with each other and influence ICL (Wei et al., 2022b; Lu et al., 2024). A potential approach involves designing qualitative and quantitative analysis for ICL by systematically accounting for a range of potential factors. For example, Biderman et al. (2023) controlled for training variables including model architecture, training scale, model checkpoints, hyper-parameters, and source libraries to investigate the influence of data frequency on ICL and bias behaviour in LLMs. Another key challenge is the absence of benchmark datasets that effectively justify the causal effects of the investigating factors on ICL (Zhang et al., 2023a; Jin et al., 2024). One possible solution is to generate synthetic datasets with domain-specific expertise and develop methods in causal discovery and inference (Swaminathan et al., 2023; Kiciman et al., 2024) for interpreting ICL through a causal lens.

Evaluation Current research efforts typically measure ICL by assessing task performance or optimizing criteria such as gradient (von Oswald et al., 2023a) and token loss (Olsson et al., 2022) during the pre-training stage. However, Schaeffer et al. (2023) have recently argued that emergent abilities (e.g., ICL) discussed in some prior studies appear to be mirages, due to the researchers’ choice of evaluation metrics. Their hypothesis is that choosing a metric that nonlinearly or discontinuously deforms per-token error rates and the limited size of test datasets may not provide an accurate estimation of the performance of smaller models. The core challenge is that aggregated performance metrics do not adequately assess ICL ability across various scenarios or predict their behaviour in new tasks alongside data distributions drifting. Dedicated criteria explicitly designed for the assessment of ICL are currently lacking. In addition, the evaluation process typically encompasses multiple models with a distinct objective in existing LLM paradigms, such as reinforcement learning from human feedback (RLHF) (Christiano et al., 2017). A potential solution involves identifying and addressing specific capability gaps to enhance predictions of model performance on novel tasks. For example, Burden et al. (2023) incrementally inferred the

capability of LLMs with subsets of tasks before assessing more complex dependencies.

Demonstration Selection While research exploring the relationship between demonstration and ICL is expanding (Min et al., 2022; Xiang et al., 2024), it is mostly limited to interpretation based on a finite number of demonstrations. The core challenge lies in the exponential growth of possible demonstrations with the increase in examples. A potential approach is to formalize the demonstration selection as a sequential decision-making problem (Zhang et al., 2022), aiming to learn an approximation of the expected reward from demonstrations to identify those most beneficial for control experiments. Other potential solutions can include disentangling features supportive for performing ICL on downstream tasks, and controlling for these features in demonstrations (Si et al., 2023).

Trustworthiness Trustworthiness issues such as fairness, truthfulness, robustness, bias, and toxicity are significant concerns for LLMs. Exploring these properties within ICL in LLMs is particularly challenging due to their unanticipated nature (Kenthapadi et al., 2023). It is difficult to analyze the relationship between various aspects of ICL and factors relating to trustworthiness when LLM training objectives and downstream tasks are inconsistent. Safety concerns have also become one of the most pressing issues. Studies (Perez and Ribeiro, 2022; Bai et al., 2022) have shown that LLMs can be manipulated to perform harmful and dangerous ICL through exposure to toxic demonstrations. Understanding ICL could play a crucial role in addressing the trustworthiness and safety issues associated with LLMs. For example, knowledge of how LLMs incorporate biases during ICL can guide the development of debiasing models. Furthermore, insights gained from studying how LLMs respond to toxic demonstrations can inform the design of countermeasures aimed at detecting and filtering out harmful input.

7 Conclusion

This paper presents a comprehensive review of current research efforts focused on interpreting and analyzing ICL in LLMs. We organize these advancements into theoretical and empirical perspectives, highlighting existing challenges and discussing potential avenues for further research in this area. We believe this survey will serve as a valuable resource for encouraging further exploration into the interpretation of ICL of LLMs.

Limitation

While we referenced numerous studies to interpret and analyze in-context learning, many of them are only briefly described due to space limitations. Our aim was to provide an overview of existing research efforts into ICL interpretation, and to organize previous research within a principled framework. Moreover, the survey primarily focuses on the ICL ability, which has been extensively investigated in previous studies. Nevertheless, there are other intriguing capabilities that have emerged in LLMs, such as chain-of-thought (Chu et al., 2024; Feng et al., 2023) and instruction following (Wei et al., 2022a; Chung et al., 2024; Ouyang et al., 2022), which are not included in this survey.

Acknowledgements

This work was supported in part by the UK Engineering and Physical Sciences Research Council (EPSRC) through a Turing AI Fellowship (grant no. EP/V020579/1, EP/V020579/2) and a New Horizons grant (grant no. EP/X019063/1), and Innovate UK through the Accelerating Trustworthy AI programme (grant no. 10093055).

References

- AI@Meta. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. [What learning algorithm is in-context learning? investigations with linear models](#). In *Proc. of ICLR*.
- Ekin Akyürek, Bailin Wang, Yoon Kim, and Jacob Andreas. 2024. [In-context language learning: Architectures and algorithms](#). In *Proc. of ICML*.
- Yu Bai, Fan Chen, Haiquan Wang, Caiming Xiong, and Song Mei. 2023. [Transformers as statisticians: Provably in-context learning with in-context algorithm selection](#). In *Proc. of NeurIPS*.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, T. J. Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, Benjamin Mann, and Jared Kaplan. 2022. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#). *arXiv preprint arXiv:2204.05862*.
- Leonard E. Baum and Ted Petrie. 1966. [Statistical inference for probabilistic functions of finite state markov chains](#). *Annals of Mathematical Statistics*, 37:1554–1563.
- Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aftab Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. [Pythia: a suite for analyzing large language models across training and scaling](#). In *Proc. of ICML*.
- Eric J. Bigelow, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, and Tomer David Ullman. 2024. [In-context learning dynamics with random binary sequences](#). In *Proc. of ICLR*.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Proc. of NeurIPS*.
- John Burden, Konstantinos Voudouris, Ryan Burnell, Danaja Rutar, Lucy Cheke, and Jos'e Hern'andez-Orallo. 2023. [Inferring capabilities from task performance with bayesian triangulation](#). *arXiv preprint arXiv:2309.11975*.
- Stephanie C. Y. Chan, Adam Santoro, Andrew Kyle Lampinen, Jane X. Wang, Aaditya K Singh, Pierre H. Richemond, Jay McClelland, and Felix Hill. 2022. [Data distributional properties drive emergent in-context learning in transformers](#). In *Proc. of NeurIPS*.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. [Deep reinforcement learning from human preferences](#). In *Proc. of NeurIPS*.
- Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. [Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future](#). In *Proc. of ACL*.
- Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharar Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny

- Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.
- Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. 2023. Why can gpt learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Proc. of ACL*.
- Gilad Deutch, Nadav Magar, Tomer Bar Natan, and Guy Dar. 2024. In-context learning and gradient descent revisited. In *Proc. of NAACL*.
- Benjamin L. Edelman, Ezra Edelman, Surbhi Goel, Eran Malach, and Nikolaos Tsilivis. 2024. The evolution of statistical induction heads: In-context learning markov chains. *arXiv preprint arXiv:2402.11004*.
- Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.
- Guaho Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. 2023. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *Proc. of NeurIPS*.
- Deqing Fu, Tian-Qi Chen, Robin Jia, and Vatsal Sharan. 2024. Transformers learn higher-order optimization methods for in-context learning: A study with linear models. *arXiv preprint arXiv:2310.17086*.
- Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. 2022. What can transformers learn in-context? a case study of simple function classes. In *Proc. of NeurIPS 2022*.
- Walter Gautschi. 2011. *Numerical analysis*. Springer Science & Business Media.
- Tianyu Guo, Wei Hu, Song Mei, Huan Wang, Caiming Xiong, Silvio Savarese, and Yu Bai. 2024. How do transformers learn in-context beyond simple functions? a case study on learning with representations. In *Proc. of ICLR*.
- Michael Hahn and Navin Goyal. 2023. A theory of emergent in-context learning as implicit structure induction. *arXiv preprint arXiv:2303.07971*.
- Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. 2023. Understanding in-context learning via supportive pre-training data. In *Proc. of ACL*.
- Roei Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *Proc. Findings of EMNLP*.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and L. Sifre. 2022. Training compute-optimal large language models. In *Proc. of NeurIPS*.
- John E. Hopcroft. 1971. An $n \log n$ algorithm for minimizing states in a finite automaton. In *Proc. of ISTMC*.
- Kazuki Irie, Róbert Csordás, and Jürgen Schmidhuber. 2022. The dual form of neural networks revisited: Connecting test time predictions to training patterns via spotlights of attention. In *Proc. of ICML*.
- Hong Jun Jeon, Jason D. Lee, Qi Lei, and Benjamin Van Roy. 2024. An information-theoretic analysis of in-context learning. In *Proc. of ICML*.
- Hui Jiang. 2023. A latent space theory for emergent abilities in large language models. *arXiv preprint arXiv:2304.09960*.
- Zhijing Jin, Jia-Rou Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Scholkopf. 2024. Can large language models infer causation from correlation? In *Proc. of ICLR*.
- Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *Proc. of ICML*.
- Krishnaram Kenthapadi, Himabindu Lakkaraju, and Nazneen Rajani. 2023. Generative ai meets responsible ai: Practical challenges and opportunities. In *Proc. of SIGKDD*.
- Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dong-pil Seo, Heungsuk Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Wooyoung Park, and Nako Sung. 2021. What changes can large-scale language models bring? intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers. In *Proc. of EMNLP*.
- Louis Kirsch, James Harrison, Jascha Narain Sohl-Dickstein, and Luke Metz. 2024. General-purpose in-context learning by meta-learning transformers. *arXiv preprint arXiv:2212.04458*.
- Jannik Kossen, Tom Rainforth, and Yarin Gal. 2024. In-context learning in large language models learns

- label relationships but is not conventional learning. *Proc. of ICLR*.
- Emre Kıcıman, Robert Osazuwa Ness, Amit Sharma, and Chenhao Tan. 2024. **Causal reasoning and large language models: Opening a new frontier for causality.** *Transactions on Machine Learning Research*.
- Shuai Li, Zhao Song, Yu Xia, Tong Yu, and Tianyi Zhou. 2023a. **The closeness of in-context learning and weight shifting for softmax regression.** *arXiv preprint arXiv:2304.13276*.
- Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. 2023b. **Transformers as algorithms: Generalization and stability in in-context learning.** In *Proc. of ICML*.
- Ziqian Lin and Kangwook Lee. 2024. **Dual operating modes of in-context learning.** In *Workshop of ICLR*.
- Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. **What makes good in-context examples for GPT-3?** In *Proc. of DeeLIO*.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. **Lost in the middle: How language models use long contexts.** *Transactions of the Association for Computational Linguistics*.
- Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2024. **Are emergent abilities in large language models just in-context learning?** In *Proc. of ACL*.
- Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. **Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity.** In *Proc. of ACL*.
- Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. **Rethinking the role of demonstrations: What makes in-context learning work?** In *Proc. of EMNLP*.
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, T. J. Henighan, Benjamin Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, John Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom B. Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Christopher Olah. 2022. **In-context learning and induction heads.** *arXiv preprint arXiv:2209.11895*.
- OpenAI. 2024. **Gpt-4 technical report.** *arXiv preprint arXiv:2303.08774*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. **Training language models to follow instructions with human feedback.** In *Proc. of NeurIPS*.
- Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023. **What in-context learning "learns" in-context: Disentangling task recognition and task learning.** In *Proc. Findings of ACL*.
- Madhuri Panwar, Kabir Ahuja, and Navin Goyal. 2024. **In-context learning through the bayesian prism.** In *Proc. of ICLR*.
- Fábio Perez and Ian Ribeiro. 2022. **Ignore previous prompt: Attack techniques for language models.** In *Workshop NeurIPS*.
- Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaïd Harchaoui. 2021. **Mauve: Measuring the gap between neural text and human text using divergence frontiers.** In *Proc. of NeurIPS*.
- QwenTeam. 2024. **Qwen2 technical report.** *arXiv preprint arXiv:2407.10671*.
- Allan Raventós, Mansheej Paul, F. Chen, and Surya Ganguli. 2023. **Pretraining task diversity and the emergence of non-bayesian in-context learning for regression.** In *Proc. of NeurIPS*.
- Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. **Impact of pretraining term frequencies on few-shot reasoning.** In *Proc. of EMNLP*.
- Avik Sarkar, Paul H. Garthwaite, and Anne N. De Roeck. 2005. **A bayesian mixture model for term re-occurrence and burstiness.** In *Proc. of CoNLL*.
- Rylan Schaeffer, Brando Miranda, and Oluwasanmi Koyejo. 2023. **Are emergent abilities of large language models a mirage?** In *Proc. of NeurIPS*.
- Lingfeng Shen, Aayush Mishra, and Daniel Khashabi. 2024. **Do pretrained transformers learn in-context by gradient descent?** In *Proc. of ICML*.
- Seongjin Shin, Sang-Woo Lee, Hwijeon Ahn, Sungdong Kim, Hyoungseok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woo Chul Park, Jung-Woo Ha, and Nako Sung. 2022. **On the effect of pretraining corpora on in-context learning by a large-scale language model.** In *Proc. of NAACL*.
- Chenglei Si, Dan Friedman, Nitish Joshi, Shi Feng, Danqi Chen, and He He. 2023. **Measuring inductive biases of in-context learning with underspecified demonstrations.** In *Proc. of ACL*.
- Aaditya K. Singh, Stephanie C. Y. Chan, Ted Moskovitz, Erin Grant, Andrew M. Saxe, and Felix Hill. 2023. **The transient nature of emergent in-context learning in transformers.** In *Proc. of NeurIPS*.

- Siva K. Swaminathan, Antoine Dedieu, Rajkumar Vasudeva Raju, Murray Shanahan, Miguel Lázaro-Gredilla, and Dileep George. 2023. [Schema-learning and rebinding as mechanisms of in-context learning and emergence](#). In *Proc. of NeurIPS*.
- Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. [Large language models can be lazy learners: Analyze shortcuts in in-context learning](#). In *Findings of ACL*.
- Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier García, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Neil Houlsby, and Donald Metzler. 2023. [Unifying language learning paradigms](#). In *Proc. of ICLR*.
- Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. [Function vectors in large language models](#). In *Proc. of ICLR*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Proc. of NeurIPS*.
- Johannes von Oswald, Eyyvind Niklasson, E. Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023a. [Transformers learn in-context by gradient descent](#). In *Proc. of ICML*.
- Johannes von Oswald, Eyyvind Niklasson, Maximilian Schlegel, Seijin Kobayashi, Nicolas Zucchet, Nino Scherrer, Nolan Miller, Mark Sandler, Blaise Agüera y Arcas, Max Vladymyrov, Razvan Pascanu, and João Sacramento. 2023b. [Uncovering mesa-optimization algorithms in transformers](#). *arXiv preprint arXiv:2309.05858*.
- Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023a. [Label words are anchors: An information flow perspective for understanding in-context learning](#). In *Proc. of EMNLP*.
- Xinyi Wang, Wanrong Zhu, Michael Stephen Saxon, and William Yang Wang. 2023b. [Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning](#). In *Proc. of ICML*.
- Larry A. Wasserman. 2000. [Bayesian model selection and model averaging](#). *Journal of mathematical psychology*, 44 1:92–107.
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022a. [Finetuned language models are zero-shot learners](#). In *Proc. of ICLR*.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. [Emergent abilities of large language models](#). *Transactions on Machine Learning Research*.
- Jerry W. Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. [Larger language models do in-context learning differently](#). *arXiv preprint arXiv:2303.03846*.
- Noam Wies, Yoav Levine, and Amnon Shashua. 2023. [The learnability of in-context learning](#). In *Proc. of NeurIPS*.
- Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. 2024. [Addressing order sensitivity of in-context demonstration examples in causal language models](#). In *Proc. Findings of ACL*.
- Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. [An explanation of in-context learning as implicit bayesian inference](#). In *Proc. of ICLR*.
- Steve Yadlowsky, Lyric Doshi, and Nilesh Tripuraneni. 2023. [Pretraining data mixtures enable narrow model selection capabilities in transformer models](#). *arXiv preprint arXiv:2311.00871*.
- Kang Min Yoo, Junyeob Kim, Hyuhng Joon Kim, Hyunsoo Cho, Hwiyeol Jo, Sang-Woo Lee, Sang-goo Lee, and Taeuk Kim. 2022. [Ground-truth labels matter: A deeper look into input-label demonstrations](#). In *Proc. of EMNLP*.
- Safoora Yousefi, Leo Betthauser, Hosein Hasaneig, Raphael Milliere, and Ida Momennejad. 2024. [Decoding in-context learning: Neuroscience-inspired analysis of representations in large language models](#). *arXiv preprint arXiv:2310.00313*.
- Cheng Zhang, Stefan Bauer, Paul Bennett, Jian chuan Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawłowski, and James Vaughan. 2023a. [Understanding causality with large language models: Feasibility and opportunities](#). *arXiv preprint arXiv:2304.05524*.
- Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. [Active example selection for in-context learning](#). In *Proc. of EMNLP*.
- Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. 2023b. [What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization](#). *arXiv preprint arXiv:2305.19420*.
- Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. [Calibrate before use: Improving few-shot performance of language models](#). In *Proc. of ICML*.
- George Kingsley Zipf. 1949. [Human behavior and the principle of least effort](#). *Journal of Consulting Psychology*.