A Survey on In-context Learning

Qingxiu Dong[1], **Lei Li**[1], **Damai Dai**[1], **Ce Zheng**[1], **Jingyuan Ma**[1], **Rui Li**[1], **Heming Xia**[2], **Jingjing Xu**[3], **Zhiyong Wu**[4], **Baobao Chang**[1], **Xu Sun**[1], **Lei Li**[5] **and Zhifang Sui**[1]
[1] Peking University [2] The Hong Kong Polytechnic University
[3] ByteDance [4] Shanghai AI Lab [5] Carnegie Mellon University
dqx@stu.pku.edu.cn, szf@pku.edu.cn

Abstract

With the increasing capabilities of large language models (LLMs), in-context learning (ICL) has emerged as a new paradigm for natural language processing (NLP), where LLMs make predictions based on contexts augmented with a few examples. It has been a significant trend to explore ICL to evaluate and extrapolate the ability of LLMs. In this paper, we aim to survey and summarize the progress and challenges of ICL. We first present a formal definition of ICL and clarify its correlation to related studies. Then, we organize and discuss advanced techniques, including training strategies, prompt designing strategies, and related analysis. Additionally, we explore various ICL application scenarios, such as data engineering and knowledge updating. Finally, we address the challenges of ICL and suggest potential directions for further research. We hope that our work can encourage more research on uncovering how ICL works and improving ICL.

1 Introduction

With the scaling of model size and data size (Brown et al., 2020; Chowdhery et al., 2023; OpenAI, 2023; Touvron et al., 2023a,b), large language models (LLMs) demonstrate the in-context learning (ICL) ability, that is, learning from a few examples in the context. Many studies have shown that LLMs can perform a series of complex tasks through ICL, such as solving mathematical reasoning problems (Wei et al., 2022c). These strong abilities have been widely verified as emerging abilities for large language models (Wei et al., 2022b).

The key idea of in-context learning is to learn from analogy. Figure 1 gives an example that describes how language models make decisions via ICL. First, ICL requires a few demonstration examples to form a prompt context. These examples are usually written in natural language templates. Then, ICL concatenates a query question and the piece of prompt context together to form the input, which is then fed into the language model for prediction. Different from supervised learning, which requires a training stage that uses backward gradients to update model parameters, ICL does not perform parameter updates. The model is expected to learn the pattern hidden in the demonstration and accordingly make the right prediction.

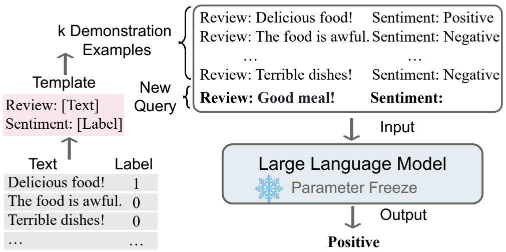

Figure 1: Illustration of in-context learning. ICL requires a prompt context containing a few demonstration examples written in natural language templates. Taking this prompt and a query as the input, large language models are responsible for making predictions.

As a new paradigm, ICL has multiple attractive advantages. First, since the demonstration is written in natural language, it provides an interpretable interface to communicate with LLMs (Brown et al., 2020). This paradigm makes it much easier to incorporate human knowledge into LLMs by changing the demonstration and templates (Liu et al., 2022; Lu et al., 2022; Wei et al., 2022c; Wu et al., 2023b). Second, in-context learning is similar to the decision process of human beings by learning from analogy (Winston, 1980). Third, compared to supervised training, ICL is a training-free learning framework. This could not only greatly reduce the computational costs for adapting the model to new tasks, but also make language-model-as-a-service (Sun et al., 2022) possible and can be easily applied to large-scale real-world tasks.

Despite being promising, there are also interesting questions and intriguing properties that require

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1107–1128
November 12-16, 2024 ©2024 Association for Computational Linguistics

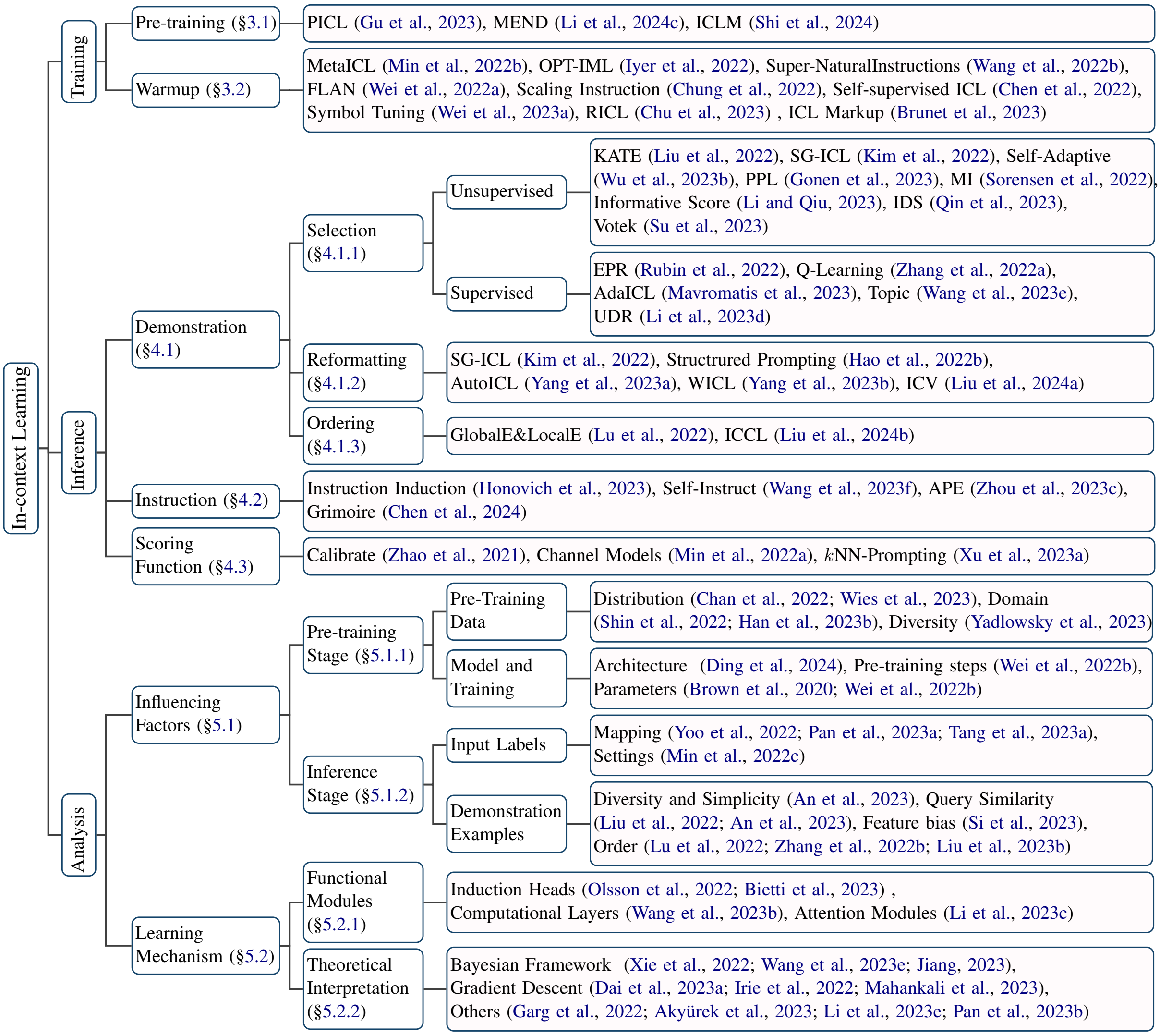
In-context Learning
Training
Pre-training (§3.1)
PICL (Gu et al., 2023), MEND (Li et al., 2024c), ICLM (Shi et al., 2024)
Warmup (§3.2)
MetaICL (Min et al., 2022b), OPT-IML (Iyer et al., 2022), Super-NaturalInstructions (Wang et al., 2022b), FLAN (Wei et al., 2022a), Scaling Instruction (Chung et al., 2022), Self-supervised ICL (Chen et al., 2022), Symbol Tuning (Wei et al., 2023a), RICL (Chu et al., 2023) , ICL Markup (Brunet et al., 2023)
Inference
Demonstration (§4.1)
Selection (§4.1.1)
Unsupervised
KATE (Liu et al., 2022), SG-ICL (Kim et al., 2022), Self-Adaptive (Wu et al., 2023b), PPL (Gonen et al., 2023), MI (Sorensen et al., 2022), Informative Score (Li and Qiu, 2023), IDS (Qin et al., 2023), Votek (Su et al., 2023)
Supervised
EPR (Rubin et al., 2022), Q-Learning (Zhang et al., 2022a), AdaICL (Mavromatis et al., 2023), Topic (Wang et al., 2023e), UDR (Li et al., 2023d)
Reformatting (§4.1.2)
SG-ICL (Kim et al., 2022), Structrured Prompting (Hao et al., 2022b), AutoICL (Yang et al., 2023a), WICL (Yang et al., 2023b), ICV (Liu et al., 2024a)
Ordering (§4.1.3)
GlobalE&LocalE (Lu et al., 2022), ICCL (Liu et al., 2024b)
Instruction (§4.2)
Instruction Induction (Honovich et al., 2023), Self-Instruct (Wang et al., 2023f), APE (Zhou et al., 2023c), Grimoire (Chen et al., 2024)
Scoring Function (§4.3)
Calibrate (Zhao et al., 2021), Channel Models (Min et al., 2022a), kNN-Prompting (Xu et al., 2023a)
Analysis
Influencing Factors (§5.1)
Pre-training Stage (§5.1.1)
Pre-Training Data
Distribution (Chan et al., 2022; Wies et al., 2023), Domain (Shin et al., 2022; Han et al., 2023b), Diversity (Yadlowsky et al., 2023)
Model and Training
Architecture (Ding et al., 2024), Pre-training steps (Wei et al., 2022b), Parameters (Brown et al., 2020; Wei et al., 2022b)
Inference Stage (§5.1.2)
Input Labels
Mapping (Yoo et al., 2022; Pan et al., 2023a; Tang et al., 2023a), Settings (Min et al., 2022c)
Demonstration Examples
Diversity and Simplicity (An et al., 2023), Query Similarity (Liu et al., 2022; An et al., 2023), Feature bias (Si et al., 2023), Order (Lu et al., 2022; Zhang et al., 2022b; Liu et al., 2023b)
Learning Mechanism (§5.2)
Functional Modules (§5.2.1)
Induction Heads (Olsson et al., 2022; Bietti et al., 2023) , Computational Layers (Wang et al., 2023b), Attention Modules (Li et al., 2023c)
Theoretical Interpretation (§5.2.2)
Bayesian Framework (Xie et al., 2022; Wang et al., 2023e; Jiang, 2023), Gradient Descent (Dai et al., 2023a; Irie et al., 2022; Mahankali et al., 2023), Others (Garg et al., 2022; Akyürek et al., 2023; Li et al., 2023e; Pan et al., 2023b)

Figure 2: Taxonomy of in-context learning.

further investigation in ICL. Although a range of vanilla GPT models show excellent ICL capability, several studies have found that this capability can be significantly improved through adaptation during pretraining (Min et al., 2022b; Li et al., 2024c). Moreover, the performance of ICL is sensitive to specific settings, including the prompt template, the selection and order of demonstration examples, and other factors (Wang et al., 2023e; Liu et al., 2024b). Additionally, optimizing the conciseness of demonstration examples and improving the computational efficiency of ICL are critical areas of ongoing research (Liu et al., 2024a). Furthermore, despite preliminary explanations (Dai et al., 2023a; Jiang, 2023), the underlying working mechanism of ICL remains unclear and requires further investigation.

With the rapid growth of studies in ICL, our survey aims to sensitize the community toward the current progress. In the following sections, we delve into an in-depth discussion of related studies, and we summarize the taxonomy in Figure 2 and the key findings in Appendix A. We highlight the challenges and potential directions and hope our work provide a useful roadmap for beginners interested in this area and shed light on future research.

2 Definition and Formulation

Following Brown et al. (2020), we here provide a formal definition of in-context learning:

> *In-context learning is a paradigm that allows language models to learn tasks given only a few examples in the form of demonstration.*

Formally, given a query input text x and a set of candidate answers $Y = \{y_1, \ldots, y_m\}$, a pretrained language model $\mathcal{M}$ takes the candidate answer with the maximum score as the prediction,[1] conditioned a demonstration set C. C contains an optional task instruction I and k demonstration examples, thus $C = \{I, s(x_1, y_1), \ldots, s(x_k, y_k)\}$ or $C = \{s'(x_1, y_1, I), \ldots, s'(x_k, y_k, I)\}$, where $s'(x_i, y_i, I)$ is an example written in natural language according to the task. Depending on whether k and the demonstration examples belong to the same task, it can be categorized as task-specific ICL and cross-task ICL. In the latter, different examples have their own instructions. The likelihood of a candidate answer y_j comes from a scoring function f on the whole input sequence:

$$P(y_j \mid x) \triangleq f_{\mathcal{M}}(y_j, C, x) \tag{1}$$

The final predicted label $\hat{y}$ is the candidate answer with the highest probability:

$$\hat{y} = \arg\max_{y_j \in Y} P(y_j \mid x). \tag{2}$$

According to the definition, we can see that ICL differs from related concepts as follows: (1) *Prompt Learning*: prompts can be discrete templates or soft parameters that encourage the model to predict the desired output. ICL can be regarded as a subclass of prompt tuning where the demonstration examples are part of the prompt. Liu et al. (2023c) made a thorough survey on prompt learning, but ICL was not included in their study. (2) *Few-shot Learning*: few-shot learning is a general machine learning approach that involves adapting model parameters to perform a task with a limited number of supervised examples (Wang and Yao, 2019). In contrast, ICL does not require parameter updates and is directly performed on pretrained LLMs.

3 Model Training

Although LLMs have demonstrated promising ICL capability directly, many studies revealed that these ICL capabilities can be further enhanced through specialized training before inference (Chen et al., 2022; Gu et al., 2023; Shi et al., 2024).

3.1 Pretraining

One straightforward direction to boost the ICL capability of LLMs is through pretraining or continual pretraining. For instance, Gu et al. (2023) and Shi et al. (2024) proposed to reorganize pretraining corpora by aggregating related contexts, making models learn to reason across prior demonstrations. Differently, Li et al. (2024c) introduced a meta-distillation pretraining process, which allows LLMs to reason with distilled demonstration vectors, thereby enhancing ICL efficiency without compromising its effectiveness.

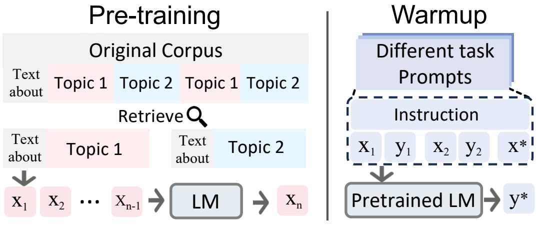

Figure 3: Illustration of model training methods to enhance ICL capabilities through two different stages: pretraining and warmup.

3.2 Warmup

Another way to enhance ICL ability is adding a continual training stage between pretraining and ICL inference, which we call model warmup for short. Warmup is an optional procedure for ICL, which adjusts LLMs before inference by modifying or adding parameters.

As most pretraining data are not tailored for ICL (Chen et al., 2022), researchers have introduced various warmup strategies to bridge the gap between pretraining and ICL inference. Both Min et al. (2022b) and Wang et al. (2022b) proposed to continually finetune LLMs on a broad range of tasks with multiple demonstration examples, which boosts ICL abilities. To encourage the model to learn input-label mappings from the context, Wei et al. (2023a) proposed symbol tuning, which substitutes natural language labels (e.g., "positive/negative sentiment") with arbitrary symbols (e.g., "foo/bar"). Chen et al. (2022) proposed a self-supervised method to align raw text with ICL formats in downstream tasks. Besides, multiple studies have indicated the potential value of instructions (Mishra et al., 2021; Wei et al., 2022a). Tuning the 137B LaMDA-PT (Thoppilan et al., 2022) on over 60 datasets verbalized via natural language instruction templates, FLAN (Wei et al., 2022a) improves the ability of LLMs to follow instructions, boosting both the zero-shot and few-shot ICL performance. Chung et al. (2022) and Wang et al. (2022b) proposed to further scale up instruction tuning with more than 1000+ task instructions.

[1] Y could be class labels or a set of free-text phrases.

4 Prompt Designing

In this section, we focus on the principles of ICL during inference, including demonstration organization (§4.1) and instruction formatting (§4.2) .

4.1 Demonstration Organization

Many studies have shown that the performance of ICL strongly relies on the demonstration surface, including the selection, formatting, and ordering of demonstration examples (Zhao et al., 2021; Lu et al., 2022). In this subsection, we survey demonstration organization strategies and classify them into three categories, as shown in Table 1.

4.1.1 Demonstration Selection

Demonstrations selection aims to answer a fundamental question: *Which samples are good examples for ICL?* We categorize the related studies into two approaches: unsupervised methods based on predefined metrics and supervised methods.

Unsupervised Method A straightforward approach to selecting ICL examples is to choose the nearest neighbors of input instances based on their similarities (Liu et al., 2022; Tanwar et al., 2023; Qin et al., 2023). Distance metrics, such as L2 distance or cosine similarity based on sentence embeddings, are commonly used for this purpose. For example, Liu et al. (2022) proposed KATE, the first kNN-based unsupervised retriever for selecting in-context examples. Similarly, k-NN cross-lingual demonstrations can be retrieved for multi-lingual ICL to strengthen source-target language alignment (Tanwar et al., 2023). Su et al. (2023) proposed to combine graphs and confidence scores to select diverse and representative examples. In addition to distance metrics, mutual information (Sorensen et al., 2022) and perplexity (Gonen et al., 2023) have proven valuable for prompt selection without labeled examples or specific LLMs. Furthermore, using output scores of LLMs as unsupervised metrics has shown effectiveness in demonstration selection (Wu et al., 2023b; Nguyen and Wong, 2023; Li and Qiu, 2023). Particularly, Wu et al. (2023b) selected the best subset permutation of kNN examples based on the code length for data transmission to compress label y given x and C. Li and Qiu (2023) used infoscore, i.e., the average of $P(y|x_i, y_i, x)P(y|x)$ for all (x, y) pairs in a validation set with a diversity regularization.

Supervised Method Though off-the-shelf retrievers offer convenient services for extensive NLP tasks, they are heuristic and sub-optimal due to the lack of task-specific supervision. To address this issue, numerous supervised methods have been developed (Rubin et al., 2022; Ye et al., 2023; Wang et al., 2023e; Zhang et al., 2022a). EPR (Rubin et al., 2022) introduced a two-stage method to train a dense retriever for demonstration selection. For a specific input, it first utilized unsupervised methods (e.g., BM25) to recall similar examples as candidates and then used this data to build a supervised dense retriever. Following EPR, Li et al. (2023d) adopted a unified demonstration retriever to select demonstrations across different tasks. Unlike prior work that retrieves individual demonstrations, Ye et al. (2023) proposed retrieving entire demonstration sets to model inter-relationships between examples. Additionally, Mavromatis et al. (2023) introduced AdaICL, a model-adaptive method that employs LLM to predict the unlabeled data set, generating an uncertainty score for each instance.

Based on prompt tuning, Wang et al. (2023e) viewed LLMs as topic models that can infer concepts θ from a few demonstrations and generate tokens based on these concepts. They represent latent concepts with task-related concept tokens, which are learned to maximize $P(y|x, \theta)$. Demonstrations are selected based on their likelihood to infer the concept variable using $P(\theta|x, y)$. Additionally, reinforcement learning was introduced by Zhang et al. (2022a) for example selection. They formulated demonstration selection as a Markov decision process (Bellman, 1957) and selected demonstrations via Q-learning. The action is choosing an example, and the reward is defined as the accuracy of a labeled validation set.

In order to have a more intuitive comparison of the performance of several unsupervised methods, we select topk (Liu et al., 2022), votek (Su et al., 2023), mdl (Wu et al., 2023b) to conduct experiments. The result is shown in Table 2. The details of the experiment can be found in Appendix B.

4.1.2 Demonstration Reformatting

In addition to directly selecting examples from training data, another research trend involves utilizing LLMs to reformat the representation of existing demonstrations (Kim et al., 2022; Yang et al., 2023a; Hao et al., 2022b; Yang et al., 2023b; Liu et al., 2024a; Li et al., 2024a). For instance, Kim et al. (2022) proposed generating demonstrations directly from LLMs to reduce the reliance on external demonstration data. Structured Prompting (Hao

Category	Methods	Demonstration Acquisition	LLMs	Features
Demonstration Selection	KATE (Liu et al., 2022)	Human design	GPT-3	KNN Selection
	MI (Sorensen et al., 2022)	Human design	GPT-3	Mutual Information
	EPR (Rubin et al., 2022)	Human design	GPT-{J, 3}/CodeX	Score-based Retrieval
	IDS (Qin et al., 2023)	Human design	GPT-3.5	Iterative Selection
	AdaICL (Mavromatis et al., 2023)	Human design	GPT-{J, Neo}	Selective Demonstration
	UDR (Li et al., 2023d)	Human design	GPT-Neo-2.7B	Unified Retrieval
Demonstration Reformatting	SG-ICL (Kim et al., 2022)	LM generated	GPT-J	Auto Demonstration Generation
	AutoICL (Yang et al., 2023a)	LM generated	GPT-3.5-Turbo-0301	Reasoning Path Generation
	MSP (Yang et al., 2023b)	Human design	GPT series	Adjusting Demonstration Weight
	ICV (Liu et al., 2024a)	Human design	Falcon-7b / Llama-7b	Demonstration Embedding
Demonstration Ordering	GlobalE & LocalE (Lu et al., 2022)	Human design	GPT-{2, 3}	Best Order Selection
	ICCL (Liu et al., 2024b)	Human design	Llama2/Mixtral/Qwen	Ordering from Simple to Complex

Table 1: Summary of representative demonstration designing methods.

Model	Method	SST5	SST2	CQA	SNLI	News	Avg
GPT2	topk	40.1	74.9	30.2	39.7	62.7	49.5
	votek	32.4	51.0	29.8	35.8	25.5	34.9
	mdl	**43.3**	**86.7**	**32.7**	**41.4**	**68.0**	**54.4**
GPT-J	topk	**46.9**	84.6	58.4	**60.7**	**69.1**	**63.9**
	votek	33.8	87.3	63.4	43.1	25.3	50.6
	mdl	37.6	**87.9**	**64.1**	59.8	68.2	63.5
Qwen2	topk	54.1	83.3	76.3	**68.2**	64.9	**69.4**
	votek	**55.3**	**86.9**	76.1	51.6	**65.3**	67.0
	mdl	54.6	86.1	**77.1**	65.0	63.2	69.2
Llama3	topk	53.0	**90.3**	76.1	**64.0**	74.0	**71.5**
	votek	54.9	88.9	72.6	57.7	**78.3**	70.5
	mdl	**54.4**	89.1	**76.5**	59.9	74.6	70.9

Table 2: Fair comparison of demonstration selection methods. CQA and News are abbreviations of Commonsense QA and AG News, respectively. The best results are **bolded**. Our experiments on topk (Liu et al., 2022), votek (Su et al., 2023), mdl (Wu et al., 2023b) show that the effectiveness of ICL example selection methods are model-dependent. On GPT-2, the mdl method performs the best, while on the other three models, topk performs the best.

et al., 2022b) proposed to encode demonstration examples separately with special positional embeddings, which are then provided to the test examples using a rescaled attention mechanism. Diverging from these methods, other approaches focus on modifying the latent representation of demonstrations (Liu et al., 2024a; Li et al., 2024a). Specifically, Liu et al. (2024a) developed In-Context Vectors (ICVs) derived from the latent embeddings of demonstration examples in LLMs. These ICVs are used during inference to adjust the latent states of the LLM, thereby enhancing the model's ability to follow the demonstrations more effectively.

4.1.3 Demonstration Ordering

Ordering the selected demonstration examples is also an important aspect of demonstration organization. Lu et al. (2022) have proven that order sensitivity is a common problem and always exists for various models. To handle this problem, previous studies have proposed several training-free methods for sorting demonstration examples. Particularly, Liu et al. (2022) arranged examples based on their proximity to the input, positioning the closest example as the rightmost demonstration. Lu et al. (2022) introduced global and local entropy metrics, finding a positive correlation between these metrics and the ICL performance. Consequently, they utilized the entropy metric to determine the optimal demonstration ordering. Additionally, ICCL (Liu et al., 2024b) suggested ranking demonstrations from simple to complex, thereby gradually increasing the complexity of demonstration examples during the inference process.

4.2 Instruction Formatting

A common way to format demonstrations is concatenating examples $(x_1, y_1), \ldots, (x_k, y_k)$ with a template $\mathcal{T}$ directly. However, in some tasks that need complex reasoning (e.g., math word problems and commonsense reasoning), it is not easy to learn the mapping from x_i to y_i with only k demonstrations. Although template engineering has been studied in prompting (Liu et al., 2023c), some researchers aim to design a better format of demonstrations for ICL by describing tasks with the instruction I. Honovich et al. (2023) found that given several demonstration examples, LLMs can generate task instructions themselves. Considering the generation abilities of LLMs, Zhou et al. (2023c) proposed an Automatic Prompt Engineer for automatic instruction generation and selection.

Method	Target	Efficiency	Coverage	Stability
Direct	$\mathcal{M}(y_j \mid C, x)$	+++	+	+
PPL	PPL(S_j)	+	+++	+
Channel	$\mathcal{M}(x \mid C, y_j)$	+	+	++

Table 3: Summary of different scoring functions. Coverage refers to task coverage. The qualitative results for 'Efficiency' and 'Stability' metrics are elaborated in Table 4 and Table 5, respectively.

To further improve the quality of the automatically generated instructions, several strategies have proposed using LLMs to bootstrap off its own generations (Wang et al., 2023f; Chen et al., 2024). Additionally, chain-of-thought (CoT) (Wei et al., 2022c) introduces intermediate reasoning steps between inputs and outputs to enhance problem-solving and comprehension. Recent advancements have also emphasized the process of enhancing step-by-step reasoning in models (Zhang et al., 2023c; Wang et al., 2022a; Zhou et al., 2023a).

4.3 Scoring Function

The scoring function determines how to transform the predictions of a language model into an estimation of the likelihood of a specific answer. The Direct method uses the conditional probability of candidate answers represented by tokens in the model's vocabulary (Brown et al., 2020). The answer with the highest probability is selected as the final answer, but this method restricts template design by requiring answer tokens to be at the end of input sequences. Perplexity (PPL) is another commonly used metric that computes the sentence perplexity of the entire input sequence $S_j = \{C, s(x, y_j, I)\}$, which includes tokens from demonstration examples C, the input query x, and the candidate label y_j. PPL evaluates the probability of the sentence, eliminating token position limitations but requiring additional computation time. Min et al. (2022a) proposed using channel models (Channel) to compute the conditional probability in reverse, estimating the likelihood of the input query given the label. This approach requires language models to generate every token in the input, potentially boosting performance under imbalanced training data. We summarize all three scoring functions in Table 3. Note that in Table 3, 'Efficiency' refers to the language model inference latency; 'Coverage' reflects whether the method utilizes the output probability of the local or all token positions in the input sequence; and 'Stability' indicates whether the in-context learning ability is easily affected by changes in the demonstration examples.

5 Analysis

To understand ICL, recent studies attempt to investigate what influence ICL performance (Shin et al., 2022; Yoo et al., 2022; Kossen et al., 2023) and why ICL works (Dai et al., 2023a; Irie et al., 2022). In this section, we present a detailed elaboration of influencing factors (§5.1) and learning mechanisms (§5.2) of ICL, as illustrated in Figure 4.

5.1 Influencing Factors

We discuss relevant research addressing *what influences ICL performance*, including factors both in the pretraining stage and in the inference stage.

5.1.1 Pretraining Stage

We first introduce factors that influence the pretraining stage. The diversity of pretraining corpora significantly impacts ICL performance (Shin et al., 2022; Yadlowsky et al., 2023; Raventós et al., 2023). In particular, Shin et al. (2022) found that the source domain is more important than the corpus size, suggesting that combining multiple corpora may lead to the emergence of ICL ability. Similarly, Raventós et al. (2023) empirically identified a task diversity threshold beyond which LLMs exhibit strong ICL capabilities in unseen tasks. Another line of research investigates the impact of data distribution on ICL (Chan et al., 2022; Wies et al., 2023). For instance, Chan et al. (2022) demonstrated that ICL capability emerges when the training data exhibits specific distributional properties, such as burstiness, wherein items appear in clusters rather than being uniformly distributed over time.

Beyond these works, several studies have investigated the impact of model architecture and training process on ICL performance (Wei et al., 2022b; Brown et al., 2020; Ding et al., 2024). Wei et al. (2022b) investigated the emergent abilities of many large-scale models on multiple tasks. They suggested that a pretrained model acquires some emergent ICL abilities when it reaches a large scale of pretraining steps or model parameters. Ding et al. (2024) pointed out that the in-context samples should attend to each other during inference, indicating that current causal LLMs may lead to suboptimal ICL performance.

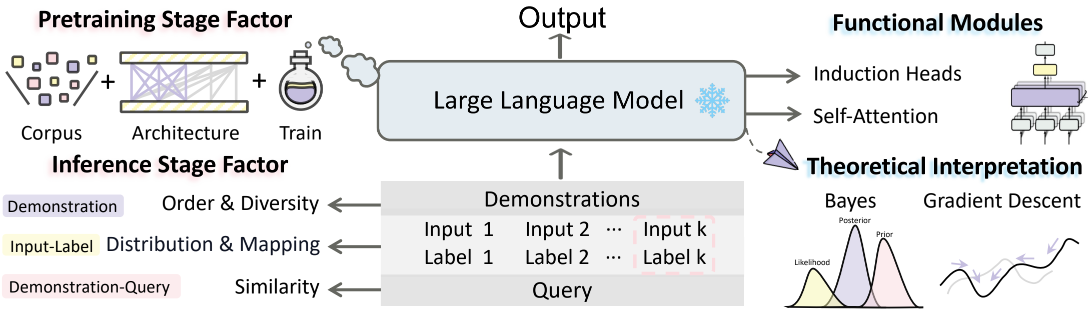

Figure 4: Summary of factors that have a relatively strong correlation to ICL performance and different perspectives to explain why ICL works.

5.1.2 Inference Stage

During inference, there are also multiple properties of demonstration examples that influence ICL performance. Min et al. (2022c) proved that input-label settings such as the pairing format, the exposure of label space, and the input distribution contribute substantially to ICL performance. However, contrary to the conclusion in Min et al. (2022c) that input-label mapping matters little to ICL, latter studies showed that the accurate mapping influence ICL performance significantly (Yoo et al., 2022; Pan et al., 2023a; Tang et al., 2023a). Wei et al. (2023b) further pointed that flipped or semantically-unrelated input-label mapping also can be learned.

From the perspective of demonstration construction, recent literature focuses on the diversity and simplicity of demonstrations (An et al., 2023), the order of samples (Lu et al., 2022; Zhang et al., 2022b; Liu et al., 2023b), and the similarity between demonstrations and queries (Liu et al., 2022). For example, Liu et al. (2022) found that demonstration samples with embeddings closer to those of the query samples typically yield better performance than those with more distant embeddings. Notably, despite efforts to refine demonstrations to optimize the performance, there still remain clear feature biases during ICL inference (Si et al., 2023). Overcoming strong prior biases and ensuring the model gives equal weight to all contextual information remain challenges (Kossen et al., 2023).

5.2 Learning Mechanism

From a learning mechanism perspective, we delve into the research addressing why ICL is effective.

5.2.1 Functional Modules

The ICL capability is intimately connected to specific functional modules within Transformers. As one of the core components, the attention module is a focal point in the study of ICL mechanism (Olsson et al., 2022; Bietti et al., 2023; Dai et al., 2023a; Irie et al., 2022; Li et al., 2023c; Gao et al., 2023; Zhang et al., 2023b). Particularly, Olsson et al. (2022) identified specific attention heads, referred to as “induction heads”, that can replicate previous patterns for next-token prediction, thus progressively developing ICL capabilities. Additionally, Wang et al. (2023b) focused on the information flow in Transformers and found that during the ICL process, demonstration label words serve as anchors, which aggregate and distribute key information for the final prediction.

5.2.2 Theoretical Interpretation

In this subsection, we introduce the theoretical interpretations of ICL from different views.

Bayesian View In the Bayesian framework, ICL is explained as implicit Bayesian inference, where models perform ICL by identifying a shared latent concept among examples (Xie et al., 2022; Wies et al., 2023; Ahuja et al., 2023; Jiang, 2023; Wang et al., 2023e). Additional perspectives suggest that LLMs encode the Bayesian Model Averaging algorithm via the attention mechanism (Zhang et al., 2023b). As the number of in-context examples increases, implicit Bayesian inference becomes analogous to kernel regression (Han et al., 2023a).

Gradient Descent View Gradient descent offers another valuable lens for understanding ICL. Dai et al. (2023a) identified a dual form between Transformer attention and gradient descent, finding that GPT-based ICL behaves similarly to explicit fine-tuning from multiple perspectives. Other studies have attempted to establish connections between ICL and gradient descent in simplified regression settings (von Oswald et al., 2023; Ahn et al., 2023; Mahankali et al., 2023; Li et al., 2023c). For instance, von Oswald et al. (2023) showed that linear attention-only Transformers with manually constructed parameters are closely related to models learned by gradient descent. Li et al. (2023c) found that self-attention-only Transformers exhibit similarities with models trained via gradient descent. However, the simplified settings used in these studies have led to debates about the direct applicability of these connections in real-world contexts (Shen et al., 2024). Fu et al. (2023) argued that Transformers perform ICL on linear regression using higher-order optimization techniques rather than gradient descent.

Other Views Beyond connecting ICL with a single algorithm, researchers have analyzed it from various perspectives, including ability decoupling, algorithmic learning, and information theory. Pan et al. (2023b) decoupled ICL capabilities into task recognition ability and task learning ability, each manifesting under different conditions. Another typical theory abstracts ICL as an algorithmic learning problem (Akyürek et al., 2023; Garg et al., 2022; Li et al., 2023e; Bai et al., 2023b), where Transformers dynamically select algorithms, such as gradient descent and ridge regression, tailored to different ICL instances. Moreover, Hahn and Goyal (2023) utilized information theory to show an error bound for ICL under linguistically motivated assumptions, explaining how next-token prediction can bring about the ICL ability.

These analytical studies have taken an essential step to explain ICL. However, most of them focused on simple tasks and small models. Extending analysis on extensive tasks and large models may be the next step to be considered.

6 Application

Given its user-friendly interface and lightweight prompting method, ICL has broad applications on traditional NLP tasks (Kim et al., 2022; Min et al., 2022b; Zhu et al., 2023b). Particularly, by using demonstrations that explicitly guide the reasoning process, ICL manifests remarkable effects on tasks requiring complex reasoning (Wei et al., 2022c; Li et al., 2023b; Zhou et al., 2022) and compositional generalization (Zhou et al., 2023a).

We explore several emerging and prevalent applications of ICL, including data engineering, model augmentation, and knowledge updating. **1) Data Engineering:** Unlike traditional methods such as human annotation and noisy automatic annotation, ICL generates relatively high-quality data at a lower cost, leading to improved performance. (Wang et al., 2021; Khorashadizadeh et al., 2023; Ding et al., 2023). **2) Model Augmentation:** The context-flexible nature of ICL shows promise in model augmentation. It can enhance retrieval-augmented methods by prepending grounding documents to the input (Ram et al., 2023). Additionally, ICL for retrieval demonstrates potential in steering models toward safer outputs (Panda et al., 2023; Meade et al., 2023). **3) Knowledge Updating:** LLMs often contain outdated or incorrect knowledge (Dong et al., 2023). ICL has demonstrated efficacy in revising such knowledge through carefully crafted demonstrations, yielding higher success rates compared to gradient-based methods (De Cao et al., 2021).

As mentioned above, ICL has yielded significant benefits on both traditional and emergent NLP applications. The tremendous success of ICL in NLP has inspired researchers to explore its potential in various modalities beyond text (elaborated in Appendix D), including vision (Bar et al., 2022; Wang et al., 2023c), vision-language (Tsimpoukelli et al., 2021; Alayrac et al., 2022), as well as speech applications (Wang et al., 2023a; Zhang et al., 2023d).

7 Challenges and Future Directions

In this section, we review existing challenges and discuss future directions for ICL.

Efficiency and Scalability The use of demonstrations in ICL introduces two challenges: (1) higher computational costs with an increasing number of demonstrations (*efficiency*), and (2) fewer learnable samples due to the maximum input length of LLMs (*scalability*). Prior research has attempted to mitigate these issues by distilling lengthy demonstrations into compact vectors (Li et al., 2024d,c) or expediting LLM inference times (Liu et al., 2023d). However, these methods often involve a trade-off in performance or necessitate access to model parameters, which is impractical for closed-source models like ChatGPT and Claude (Zhou et al., 2023b). Thus, enhancing the scalability and efficiency of ICL with more demonstrations remains a significant challenge.

Generalization ICL heavily relies on high-quality demonstrations selected from annotated examples, which are often scarce in low-resource languages and tasks. This scarcity poses a challenge to the generalization ability of ICL (He et al., 2024). Given that there is a substantial discrepancy in the availability of annotated high-resource data and low-resource data, the potential to leverage high-resource data to address low-resource tasks is highly appealing (Chatterjee et al., 2024; Tanwar et al., 2023).

Long-context ICL Recent advances in context-extended LLMs have spurred research into the impact of ICL when using an increasing number of demonstration examples (Agarwal et al., 2024; Bertsch et al., 2024). However, researchers have found that increasing the number of demonstrations does not necessarily enhance performance and may even be detrimental. These performance declines indicate a need for further investigation. Additionally, Li et al. (2024b) developed LongICLBench, which includes diverse extreme-label classification tasks, revealing further weaknesses of LLMs in comprehending extended demonstrations.

8 Conclusion

In this paper, we comprehensively review the existing literature on ICL, examining advanced techniques, conducting analytical studies, discussing relevant applications, and identifying critical challenges and potential directions for future research. To our knowledge, this is the first comprehensive survey dedicated to ICL. We aim to highlight the current state of research in ICL and provide insights to guide future work in this promising area.

Limitations

This paper offers a comprehensive examination and summary of current methodologies and analyses in the area of In-Context Learning (ICL). However, given the extensive body of related work, particularly in demonstration design and the principle analysis of ICL, we may have overlooked some equally valuable contributions. Additionally, we outline several future directions for research in ICL, including long-context ICL, efficiency and scalability in ICL, etc. We plan to leave these aspects for future work. Furthermore, many papers covered by this survey did not utilize the most up-to-date models while running experiments. We advocate for more thorough and up-to-date research to provide actionable insights for practitioners.

References

Rishabh Agarwal, Avi Singh, Lei M. Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, John D. Co-Reyes, Eric Chu, Feryal Behbahani, Aleksandra Faust, and Hugo Larochelle. 2024. Many-shot in-context learning. *Preprint*, arXiv:2404.11018.

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. 2023. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Kabir Ahuja, Madhur Panwar, and Navin Goyal. 2023. In-context learning through the bayesian prism. *CoRR*, abs/2306.04891.

AI@Meta. 2024. Llama 3 model card. *Technical report, Meta*.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Shengnan An, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Jian-Guang Lou, and Dongmei Zhang. 2023. How do in-context examples affect compositional generalization? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11027–11052. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. 2023b. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information*

Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023.

Amir Bar, Yossi Gandelsman, Trevor Darrell, Amir Globerson, and Alexei Efros. 2022. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems*, 35:25005–25017.

Richard Bellman. 1957. A markovian decision process. *Journal of mathematics and mechanics*, pages 679–684.

Amanda Bertsch, Maor Ivgi, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. 2024. In-context learning with long-context models: An in-depth exploration. *CoRR*, abs/2405.00200.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Hervé Jégou, and Léon Bottou. 2023. Birth of a transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *ArXiv*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

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 *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Marc-Etienne Brunet, Ashton Anderson, and Richard S. Zemel. 2023. ICL markup: Structuring in-context learning using soft-token tags. *CoRR*, abs/2312.07405.

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya K. Singh, Pierre H. Richemond, James L. McClelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Anwoy Chatterjee, Eshaan Tanwar, Subhabrata Dutta, and Tanmoy Chakraborty. 2024. Language models can exploit cross-task in-context learning for data-scarce novel tasks. *CoRR*, abs/2405.10548.

Ding Chen, Shichao Song, Qingchen Yu, Zhiyu Li, Wenjin Wang, Feiyu Xiong, and Bo Tang. 2024. Grimoire is all you need for enhancing large language models. *CoRR*, abs/2401.03385.

Mingda Chen, Jingfei Du, Ramakanth Pasunuru, Todor Mihaylov, Srini Iyer, Veselin Stoyanov, and Zornitsa Kozareva. 2022. Improving in-context few-shot learning via self-supervised training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3558–3573, Seattle, United States. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben

Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Timothy Chu, Zhao Song, and Chiwun Yang. 2023. Fine-tune language models to approximate unbiased in-context learning. *CoRR*, abs/2310.03331.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023a. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4005–4019. Association for Computational Linguistics.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023b. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Editing factual knowledge in language models. In *Proc. of EMNLP*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11173–11195. Association for Computational Linguistics.

Nan Ding, Tomer Levinboim, Jialin Wu, Sebastian Goodman, and Radu Soricut. 2024. CausalLM is not optimal for in-context learning. In *The Twelfth International Conference on Learning Representations*.

Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Zhifang Sui, and Lei Li. 2023. Statistical knowledge assessment for large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 29812–29830. Curran Associates, Inc.

Deqing Fu, Tian-Qi Chen, Robin Jia, and Vatsal Sharan. 2023. Transformers learn higher-order optimization methods for in-context learning: A study with linear models. *CoRR*, abs/2310.17086.

Yeqi Gao, Zhao Song, and Shenghao Xie. 2023. In-context learning for attention scheme: from single softmax regression to multiple softmax regression via a tensor trick. *CoRR*, abs/2307.02419.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. 2022. What can transformers learn in-context? A case study of simple function classes. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Hila Gonen, Srini Iyer, Terra Blevins, Noah A. Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 10136–10148. Association for Computational Linguistics.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Pre-training to learn in context. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4849–4870. Association for Computational Linguistics.

Michael Hahn and Navin Goyal. 2023. A theory of emergent in-context learning as implicit structure induction. *CoRR*, abs/2303.07971.

Chi Han, Ziqi Wang, Han Zhao, and Heng Ji. 2023a. Explaining emergent in-context learning as kernel regression. *Preprint*, arXiv:2305.12766.

Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. 2023b. Understanding in-context learning via supportive pretraining data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12660–12673. Association for Computational Linguistics.

Yaru Hao, Haoyu Song, Li Dong, Shaohan Huang, Zewen Chi, Wenhui Wang, Shuming Ma, and Furu Wei. 2022a. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*.

Yaru Hao, Yutao Sun, Li Dong, Zhixiong Han, Yuxian Gu, and Furu Wei. 2022b. Structured prompting: Scaling in-context learning to 1,000 examples. *ArXiv preprint*, abs/2212.06713.

Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, and Heng Tao Shen. 2023. ICL-D3IE: in-context learning with diverse demonstrations updating for document information extraction. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 19428–19437. IEEE.

Wei He, Shichun Liu, Jun Zhao, Yiwen Ding, Yi Lu, Zhiheng Xi, Tao Gui, Qi Zhang, and Xuanjing Huang. 2024. Self-demos: Eliciting out-of-demonstration generalizability in large language models. *CoRR*, abs/2404.00884.

Clyde Highmore. 2024. In-context learning in large language models: A comprehensive survey.

Or Honovich, Uri Shaham, Samuel R. Bowman, and Omer Levy. 2023. Instruction induction: From few examples to natural language task descriptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1935–1952. Association for Computational Linguistics.

Qian Huang, Hongyu Ren, Peng Chen, Gregor Krzmanc, Daniel Zeng, Percy Liang, and Jure Leskovec. 2023a. PRODIGY: enabling in-context learning over graphs. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Johan Bertil Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023b. Language is not all you need: Aligning perception with language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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 *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 9639–9659. PMLR.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization.

Hui Jiang. 2023. A latent space theory for emergent abilities in large language models. *CoRR*, abs/2304.09960.

Hanieh Khorashadizadeh, Nandana Mihindukulasooriya, Sanju Tiwari, Jinghua Groppe, and Sven Groppe. 2023. Exploring in-context learning capabilities of foundation models for generating knowledge graphs from text. *arXiv preprint arXiv:2305.08804*.

Hyuhng Joon Kim, Hyunsoo Cho, Junyeob Kim, Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. 2022. Self-generated in-context learning: Leveraging autoregressive language models as a demonstration generator. *ArXiv preprint*, abs/2206.08082.

Jannik Kossen, Tom Rainforth, and Yarin Gal. 2023. In-context learning in large language models learns label relationships but is not conventional learning. *CoRR*, abs/2307.12375.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Jia Li, Yunfei Zhao, Yongmin Li, Ge Li, and Zhi Jin. 2023b. Towards enhancing in-context learning for code generation. *arXiv preprint arXiv:2303.17780*.

Jiahao Li, Quan Wang, Licheng Zhang, Guoqing Jin, and Zhendong Mao. 2024a. Feature-adaptive and data-scalable in-context learning. *Preprint*, arXiv:2405.10738.

Shuai Li, Zhao Song, Yu Xia, Tong Yu, and Tianyi Zhou. 2023c. The closeness of in-context learning and weight shifting for softmax regression. *CoRR*, abs/2304.13276.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024b. Long-context llms struggle with long in-context learning. *ArXiv*, abs/2404.02060.

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023d. Unified demonstration retriever for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4644–4668. Association for Computational Linguistics.

Xiaonan Li and Xipeng Qiu. 2023. Finding supporting examples for in-context learning. *CoRR*, abs/2302.13539.

Yichuan Li, Xiyao Ma, Sixing Lu, Kyumin Lee, Xiaohu Liu, and Chenlei Guo. 2024c. MEND: meta demonstration distillation for efficient and effective in-context learning. *CoRR*, abs/2403.06914.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. 2023e. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19565–19594. PMLR.

Yinheng Li. 2023. A practical survey on zero-shot prompt design for in-context learning. *arXiv preprint arXiv:2309.13205*.

Zhuowei Li, Zihao Xu, Ligong Han, Yunhe Gao, Song Wen, Di Liu, Hao Wang, and Dimitris N. Metaxas. 2024d. Implicit in-context learning. *Preprint*, arXiv:2405.14660.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out: The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, DeeLIO@ACL 2022, Dublin, Ireland and Online, May 27, 2022*, pages 100–114. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023b. Lost in the middle: How language models use long contexts. *CoRR*, abs/2307.03172.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023c. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):195:1–195:35.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024a. In-context vectors: Making in context learning more effective and controllable through latent space steering. *Preprint*, arXiv:2311.06668.

Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, and Wei Lu. 2024b. Let's learn step by step: Enhancing in-context learning ability with curriculum learning. *Preprint*, arXiv:2402.10738.

Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Ré, and Beidi Chen. 2023d. Deja vu: Contextual sparsity for efficient llms at inference time. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 22137–22176. PMLR.

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 *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8086–8098. Association for Computational Linguistics.

Arvind Mahankali, Tatsunori B. Hashimoto, and Tengyu Ma. 2023. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *CoRR*, abs/2307.03576.

Costas Mavromatis, Balasubramaniam Srinivasan, Zhengyuan Shen, Jiani Zhang, Huzefa Rangwala, Christos Faloutsos, and George Karypis. 2023. Which examples to annotate for in-context learning? towards effective and efficient selection. *CoRR*, abs/2310.20046.

Nicholas Meade, Spandana Gella, Devamanyu Hazarika, Prakhar Gupta, Di Jin, Siva Reddy, Yang Liu, and Dilek Hakkani-Tur. 2023. Using in-context learning to improve dialogue safety. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 11882–11910. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022a. Noisy channel language model prompting for few-shot text classification. In *Proc. of ACL*, pages 5316–5330, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022b. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022c. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11048–11064. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions. *arXiv preprint arXiv:2104.08773*.

Tai Nguyen and Eric Wong. 2023. In-context example selection with influences. *arXiv preprint arXiv:2302.11042*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads. *CoRR*, abs/2209.11895.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023a. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Annual Meeting of the Association for Computational Linguistics*.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. 2023b. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8298–8319. Association for Computational Linguistics.

Ashwinee Panda, Tong Wu, Jiachen T. Wang, and Prateek Mittal. 2023. Differentially private in-context learning. *CoRR*, abs/2305.01639.

Chengwei Qin, Aston Zhang, Anirudh Dagar, and Wenming Ye. 2023. In-context learning with iterative demonstration selection. *CoRR*, abs/2310.09881.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *Technical report, OpenAi*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 784–789. Association for Computational Linguistics.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *CoRR*, abs/2302.00083.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. 2023. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. 2022. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2655–2671, Seattle, United States. Association for Computational Linguistics.

Abulhair Saparov and He He. 2023. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Lingfeng Shen, Aayush Mishra, and Daniel Khashabi. 2024. Do pretrained transformers learn in-context by gradient descent? *Preprint*, arXiv:2310.08540.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *ArXiv preprint*, abs/2210.03057.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Xi Victoria Lin, Noah A. Smith, Luke Zettlemoyer, Wen tau Yih, and Mike Lewis. 2024. In-context pretraining: Language modeling beyond document boundaries. In *The Twelfth International Conference on Learning Representations*.

Seongjin Shin, Sang-Woo Lee, Hwijeen Ahn, Sungdong Kim, HyoungSeok Kim, Boseop Kim, Kyunghyun Cho, Gichang Lee, Woomyoung 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 *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5168–5186, Seattle, United States. Association for Computational Linguistics.

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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11289–11310. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013a. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013b. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL.

Taylor Sorensen, Joshua Robinson, Christopher Rytting, Alexander Shaw, Kyle Rogers, Alexia Delorey, Mahmoud Khalil, Nancy Fulda, and David Wingate. 2022. An information-theoretic approach to prompt engineering without ground truth labels. In *Proc. of ACL*, pages 819–862, Dublin, Ireland. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv preprint*, abs/2206.04615.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. 2023. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. 2022. Black-box tuning for language-model-as-a-service. *ArXiv preprint*, abs/2201.03514.

Yanpeng Sun, Qiang Chen, Jian Wang, Jingdong Wang, and Zechao Li. 2023. Exploring effective factors for improving visual in-context learning. *arXiv preprint arXiv:2304.04748*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13003–13051. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023a. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4645–4657. Association for Computational Linguistics.

Yuting Tang, Ratish Puduppully, Zhengyuan Liu, and Nancy Chen. 2023b. In-context learning of large language models for controlled dialogue summarization: A holistic benchmark and empirical analysis. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 56–67, Singapore. Association for Computational Linguistics.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual llms are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6292–6307. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *ArXiv preprint*, abs/2201.08239.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 200–212.

Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. 2022. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). *ArXiv preprint*, abs/2206.10498.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. `https://github.com/kingoflolz/mesh-transformer-jax`.

Boshi Wang, Xiang Deng, and Huan Sun. 2022a. Iteratively prompt pre-trained language models for chain of thought. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2714–2730. Association for Computational Linguistics.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023a. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023b. Label words are anchors: An information flow perspective for understanding in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9840–9855. Association for Computational Linguistics.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 4195–4205. Association for Computational Linguistics.

Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. 2023c. Images speak in images: A generalist painter for in-context visual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6830–6839.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. 2023d. Seggpt: Towards segmenting everything in context. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 1130–1140. IEEE.

Xinyi Wang, Wanrong Zhu, and William Yang Wang. 2023e. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *arXiv preprint arXiv:2301.11916*.

Yaqing Wang and Quanming Yao. 2019. Few-shot learning: A survey. *CoRR*, abs/1904.05046.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023f. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13484–13508. Association for Computational Linguistics.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022b. Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5085–5109. Association for Computational Linguistics.

Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang (Atlas) Wang, and Mingyuan Zhou. 2023g. In-context learning unlocked for diffusion models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022c. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Jerry W. Wei, Le Hou, Andrew K. Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc V. Le. 2023a. Symbol tuning improves in-context learning in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 968–979. Association for Computational Linguistics.

Jerry W. Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023b. Larger language models do in-context learning differently. *CoRR*, abs/2303.03846.

Noam Wies, Yoav Levine, and Amnon Shashua. 2023. The learnability of in-context learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Patrick H Winston. 1980. Learning and reasoning by analogy. *Communications of the ACM*, 23(12):689–703.

Zhenyu Wu, YaoXiang Wang, Jiacheng Ye, Jiangtao Feng, Jingjing Xu, Yu Qiao, and Zhiyong Wu. 2023a. Openicl: An open-source framework for in-context learning. *CoRR*, abs/2303.02913.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023b. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1423–1436. Association for Computational Linguistics.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Benfeng Xu, Quan Wang, Zhendong Mao, Yajuan Lyu, Qiaoqiao She, and Yongdong Zhang. 2023a. k nn prompting: Learning beyond the context with nearest neighbor inference. In *International Conference on Learning Representations*.

Xin Xu, Yue Liu, Panupong Pasupat, Mehran Kazemi, et al. 2024. In-context learning with retrieved demonstrations for language models: A survey. *arXiv preprint arXiv:2401.11624*.

Zhiyang Xu, Ying Shen, and Lifu Huang. 2023b. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11445–11465. Association for Computational Linguistics.

Steve Yadlowsky, Lyric Doshi, and Nilesh Tripuraneni. 2023. Pretraining data mixtures enable narrow model selection capabilities in transformer models. *CoRR*, abs/2311.00871.

Jinghan Yang, Shuming Ma, and Furu Wei. 2023a. Auto-icl: In-context learning without human supervision. *CoRR*, abs/2311.09263.

Zhe Yang, Damai Dai, Peiyi Wang, and Zhifang Sui. 2023b. Not all demonstration examples are equally beneficial: Reweighting demonstration examples for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13209–13221. Association for Computational Linguistics.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023. Compositional exemplars for in-context learning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 39818–39833. PMLR.

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 *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2422–2437. Association for Computational Linguistics.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022a. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical*

Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, pages 9134–9148. Association for Computational Linguistics.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022b. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9134–9148. Association for Computational Linguistics.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. 2023a. What makes good examples for visual in-context learning? In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. 2023b. What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *CoRR*, abs/2305.19420.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023c. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023d. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*.

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*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc V. Le, and Ed H. Chi. 2023a. Least-to-most prompting enables complex reasoning in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Hattie Zhou, Azade Nova, Hugo Larochelle, Aaron C. Courville, Behnam Neyshabur, and Hanie Sedghi. 2022. Teaching algorithmic reasoning via in-context learning. *CoRR*, abs/2211.09066.

Wangchunshu Zhou, Yuchen Eleanor Jiang, Ryan Cotterell, and Mrinmaya Sachan. 2023b. Efficient prompting via dynamic in-context learning. *CoRR*, abs/2305.11170.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023c. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yuxiang Zhou, Jiazheng Li, Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. 2023d. The mystery and fascination of llms: A comprehensive survey on the interpretation and analysis of emergent abilities. *arXiv preprint arXiv:2311.00237*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023b. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A Takeaway

Through a comprehensive literature review of ICL, we have discovered takeaways across several domains. These include training, demonstration design, scoring functions, analysis, and ICL applications that go beyond text.

A.1 Training

To further enhanced ICL capabilities, methods propose to train the LLMs in the stage of pre-training and warmup before ICL inference.

◇ **Takeaway**: (1) The key idea of training before inference is to bridge the gap between pretraining and downstream ICL formats by introducing objectives close to in-context learning. Warmup is optional for ICL as many pretrained LLMs have manifested the ICL ability. (2) Compared to in-context finetuning involving demonstration, instruction finetuning without a few examples as demonstration is simpler and more popular. All these warmup methods improve the ICL capability by updating the model parameters, which implies that the ICL capability of the original LLMs has great potential for improvement. Therefore, although ICL does not strictly require model warmup, we recommend adding a warmup stage before ICL inference. (3) The performance advancement made by warmup encounters a plateau when increasingly scaling up the training data, indicating that LLMs only need a small amount of data to adapt to learn from the context during warmup.

A.2 Demonstration Organization

The performance of ICL strongly relies on the demonstration surface, including the selection, formatting, and ordering of demonstration examples.

◇ **Takeaway**: (1) Demonstration selection strategies improve the ICL performance, but most of them are instance level. Since ICL is mainly evaluated under few-shot settings, the corpus-level selection strategy is more important yet underexplored. (2) The output score or probability distribution of LLMs plays an important role in instance selecting. (3) For k demonstrations, the size of search space of permutations is k!. How to find the best orders efficiently or how to approximate the optimal ranking better is also a challenging question. (4) Adding chain-of-thoughts can effectively decompose complex reasoning tasks into intermediate reasoning steps. During inference, multi-stage demonstration designing strategies are applied to generate CoTs better. How to improve the CoT prompting ability of LLMs is also worth exploring. (5) In addition to human-written demonstrations, the generative nature of LLMs can be utilized in demonstration designing. LLMs can generate instructions, demonstrations, probing sets, chain-of-thoughts, and so on. By using LLM-generated demonstrations, ICL can largely get rid of human efforts on writing templates.

A.3 Scoring Function

The scoring function determines how to transform the predictions of a language model into an estimation of the likelihood of a specific answer. The answer with the highest probability is selected as the final answer.

◇ **Takeaway**: (1) Although directly adopting the conditional probability of candidate answers is efficient, this method still poses some restrictions on the template design. Perplexity is also a simple and widely scoring function. This method has universal applications, including both classification tasks and generation tasks. However, both methods are still sensitive to demonstration surface, while Channel is a remedy that especially works under imbalanced data regimes. (2) Existing scoring functions all compute a score straightforwardly from the conditional probability of LLMs. There is limited research on calibrating the bias or mitigating the sensitivity via scoring strategies.

A.4 Analysis

Numerous analytical studies investigate influencing factors of ICL during both the pretraining and inference stages, and attempt to figure out the learning mechanisms of ICL from the perspective of functional modules and theoretical interpretation.

◇ **Takeaway**: (1) Knowing and considering why ICL works and what factors may influence can help us improve the ICL performance. (2) Although some analytical studies have taken a preliminary step to explain ICL, most of them are limited to simple tasks and small models. Extending analysis on extensive tasks and large models may be the next step to be considered. (3) Among existing work, explaining ICL with gradient descent seems to be a reasonable, general, and promising direction for future research. If we build clear connections between ICL and gradient-descent-based learning, we can borrow ideas from the history of traditional deep learning to improve ICL.

A.5 In-context Learning Beyond Text

The tremendous success of ICL in NLP has inspired researchers to explore in-context learning in different modalities beyond natural language with promising results.

◇ **Takeaway**: (1) Properly formatted data (e.g., interleaved image-text datasets for vision-language tasks) and architecture designs are key factors for activating the potential of in-context learning. Exploring it in a more complex structured space such as for graph data is challenging and promising (Huang et al., 2023a). (2) Findings in textual in-context learning demonstration design and selection cannot be trivially transferred to other modalities. Domain-specific investigation is required to fully leverage the potential of in-context learning in various modalities.

B Experimental Detail

In the experiment, we utilize 8 demonstrations and test on gpt2 (Radford et al., 2019), gptj (Wang and Komatsuzaki, 2021), LLaMA3-8B-Instruct(AI@Meta, 2024) and Qwen2-7B-Instruct (Bai et al., 2023a). All experiments are executed on a single NVIDIA A100 (80G). For datasets we choose sst2 (Socher et al., 2013a), sst5 (Socher et al., 2013b), commonsense_qa (Talmor et al., 2019), ag_news (Zhang et al., 2015) and snli (Bowman et al., 2015). For the last two datasets, we only select 1000 data from the train-

Model	Direct	PPL	Channel
GPT2	44.13(1.00)	114.02(2.58)	157.70(3.57)
GPT-J	611.04(1.00)	1766.82(2.89)	1793.27(2.93)
Qwen2	745.89(1.00)	1886.63(2.53)	1957.97(2.63)
Llama3	790.46(1.00)	1935.04(2.45)	1956.21(2.47)
AVG	**1.00**	**2.61**	**2.90**

Table 4: The qualitative results of the Efficiency metric in Table 3 which record the language model inference latency (including the time for scoring with different scoring functions, with input data containing 8 in-context examples). The unit is milliseconds (ms). Each cell's parentheses contain the ratio of the latency for the current column model using the current row scoring function to the latency using direct inference. The final calculated average is the average of these ratios.

Model	Direct	PPL	Channel
GPT2	1.12	0.85	3.18
GPT-J	1.00	0.77	4.06
Qwen2	0.72	0.70	2.43
Llama3	0.89	0.78	2.43
AVG	**0.93**	**0.78**	**3.03**

Table 5: The qualitative results of the Stability metric in Table 3 which reflect whether the in-context learning ability is easily affected by changes in demonstration examples. We conducted experiments using a test set of size 10k and set up 5 different random seeds. Each time, 8 examples were randomly selected from 5k training examples for the experiments. The table records the variance of performance.

ing set for retrieval and the first 1000 data from the test set for testing. During the inference phase, a PPL-based approach is employed. The entire code framework is built upon OpenICL (Wu et al., 2023a), for which we extend our gratitude to the authors.

Table 4 and Table 5 show the quantitative results on the efficiency and stability metrics for different scoring functions in Table 3.

C Evaluation and Resources

C.1 Traditional Tasks

As a general learning paradigm, ICL can be examined on various traditional datasets and benchmarks, e.g., SuperGLUE (Wang et al., 2019), SQuAD (Rajpurkar et al., 2018). Implementing ICL with 32 randomly sampled examples on SuperGLUE, Brown et al. (2020) found that GPT-3 can achieve results comparable to state-of-the-art (SOTA) finetuning performance on COPA and ReCoRD, but still falls behind finetuning on most NLU tasks. Hao et al. (2022b) showed the potential of scaling up the number of demonstration examples. However, the improvement brought by scaling is very limited. At present, compared to finetuning, there still remains some room for ICL to reach on traditional NLP tasks.

Benchmark	Tasks	#Tasks
BIG-Bench (Srivastava et al., 2022)	Mixed tasks	204
BBH (Suzgun et al., 2023)	Unsolved problems	23
PRONTOQA (Saparov and He, 2023)	Question answering	1
MGSM (Shi et al., 2022)	Math problems	1
LLMAS (Valmeekam et al., 2022)	Plan and reasoning tasks	8
OPT-IML Bench (Iyer et al., 2022)	Mixed tasks	2000

Table 6: New challenging evaluation benchmarks for ICL. For short, we use LLMAS to represent LLM Assessment Suite (Valmeekam et al., 2022).

C.2 New Challenging Tasks

In the era of large language models with in-context learning capabilities, researchers are more interested in evaluating the intrinsic capabilities of large language models without downstream task finetuning (Bommasani et al., 2021).

To explore the capability limitations of LLM on various tasks, Srivastava et al. (2022) proposed the BIG-Bench (Srivastava et al., 2022), a large benchmark covering a large range of tasks, including linguistics, chemistry, biology, social behavior, and beyond. The best models have already outperformed the average reported human-rater results on 65% of the BIG-Bench tasks through ICL (Suzgun et al., 2023). To further explore tasks actually unsolvable by current language models, Suzgun et al. (2023) proposed a more challenging ICL benchmark, BIG-Bench Hard (BBH). BBH includes 23 unsolved tasks, constructed by selecting challenging tasks where the state-of-art model performances are far below the human performances. Besides, researchers are searching for inverse scaling tasks,[2] that is, tasks where model performance reduces when scaling up the model size. Such tasks also highlight potential issues with the cur-

[2] https://github.com/inverse-scaling/prize

rent paradigm of ICL. To further probe the model generalization ability, Iyer et al. (2022) proposed OPT-IML Bench, consisting of 2000 NLP tasks from 8 existing benchmarks, especially benchmark for ICL on held-out categories.

Specifically, a series of studies focus on exploring the reasoning ability of ICL. Saparov and He (2023) generated an example from a synthetic world model represented in first-order logic and parsed the ICL generations into symbolic proofs for formal analysis. They found that LLMs can make correct individual deduction steps via ICL. Shi et al. (2022) constructed the MGSM benchmark to evaluate the chain-of-thought reasoning abilities of LLMs in multilingual settings, finding that LLMs manifest complex reasoning across multiple languages. To further probe more sophisticated planning and reasoning abilities of LLMs, Valmeekam et al. (2022) provided multiple test cases for evaluating various reasoning abilities on actions and change, where existing ICL methods on LLMs show poor performance.

In addition, Tang et al. (2023b) proposed a benchmark called SAMSum, which is a human-annotated dataset specifically designed for multi-turn dialogue summarization, to evaluate the quality of dialogue summaries generated by LLMs via ICL.

C.3 Open-source Tools

Noticing that ICL methods are often implemented differently and evaluated using different LLMs and tasks, Wu et al. (2023a) developed OpenICL, an open-source toolkit enabling flexible and unified ICL assessment. With its adaptable architecture, OpenICL facilitates the combination of distinct components and offers state-of-the-art retrieval and inference techniques to accelerate the integration of ICL into advanced research.

D In-Context Learning Beyond Text

The tremendous success of ICL in NLP has inspired researchers to explore its potential in different modalities, including visual, vision+language and speech tasks as well.

D.1 Visual In-Context Learning

Employing masked auto-encoders (MAE) for image patch infilling, the model trained by Bar et al. (2022) generates consistent output images at inference, demonstrating robust ICL capabilities for tasks like image segmentation. This method is expanded in Painter (Wang et al., 2023c), which incorporates multiple tasks to develop a generalist model with competitive performance. SegGPT (Wang et al., 2023d) further builds on this by integrating diverse segmentation tasks and exploring ensemble techniques to enhance example quality. Additionally, Wang et al. (2023g) introduce the Prompt Diffusion model, the first diffusion-based model with ICL abilities, guided by an extra text prompt for more precise image generation, as illustrated in Figure 5.

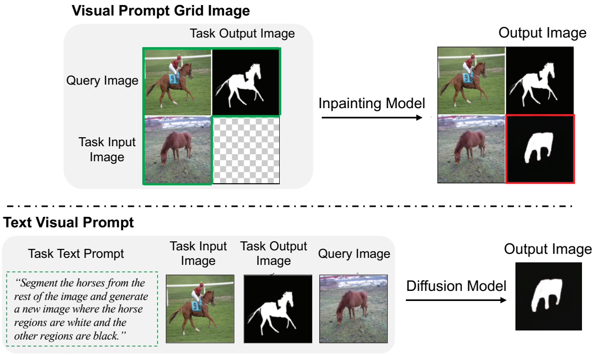

Figure 5: Image-only and textual augmented prompting for visual in-context learning.

Similar to ICL in NLP, the effectiveness of visual in-context learning greatly depends on the choice of demonstration images, as shown in research by (Zhang et al., 2023a) and (Sun et al., 2023). To optimize this, Zhang et al. (2023a) examine two strategies: using an unsupervised retriever to select the nearest samples with an existing model, and a supervised approach to train a specialized retriever to boost ICL performance. These approaches improve results by ensuring semantic similarity and better alignment in viewpoint, background, and appearance. Beyond retrieval, Sun et al. (2023) also investigate a prompt fusion technique to further enhance outcomes.

D.2 Multi-Modal In-Context Learning

In the vision-language domain, a vision encoder paired with a frozen language model demonstrates multi-modal few-shot learning capabilities after training on image-caption datasets, as shown by the Frozen model (Tsimpoukelli et al., 2021). Extending this, Flamingo integrates a vision encoder with large language models (LLMs) for enhanced in-context learning across multi-modal tasks, leveraging large-scale web corpora (Alayrac et al., 2022). Similarly, Kosmos-1 exhibits zero-shot, few-shot, and multi-modal chain-of-thought prompting abilities (Huang et al., 2023b). METALM introduces a semi-causal language modeling objective to achieve strong ICL performance across vision-language tasks (Hao et al., 2022a). The ICL-D3IE approach employs a novel in-context learning framework that iteratively updates diverse demonstrations—including hard, layout-aware, and formatting demonstrations to train large language models (LLMs) for enhanced document information extraction (DIE)(He et al., 2023). Recent advancements include creating instruction tuning datasets from existing vision-language tasks or with advanced LLMs like GPT-4, connecting LLMs with powerful vision foundational models like BLIP-2 for multi-modal learning (Xu et al., 2023b; Li et al., 2023a; Liu et al., 2023a; Zhu et al., 2023a; Dai et al., 2023b).

D.3 Speech In-Context Learning

In the speech area, Wang et al. (2023a) treated text-to-speech synthesis as a language modeling task. They use audio codec codes as an intermediate representation and propose the first TTS framework with strong in-context learning capability. Subsequently, VALLE-X (Zhang et al., 2023d) extend the idea to multi-lingual scenarios, demonstrating superior performance in zero-shot cross-lingual text-to-speech synthesis and zero-shot speech-to-speech translation tasks.

D.4 Comparison with other survey papers

Our survey was drafted and posted on the Arxiv at the end of 2022, which is, to the best of our knowledge, the very first to review in-context learning in the field. We also regularly update this survey in a timely manner, with four major revisions.

Starting from 2023, we notice the emerge of several related survey in the field of in-context learning. Xu et al. (2024) made a comprehensive review on the choices for models, training procedures and inference algorithms to retrieve demonstrative examples of in-context learning. Li (2023) provided practical suggestions on prompt engineering for in-context learning. Zhou et al. (2023d) and Highmore (2024) focused on the theoretical interpretation and analysis of ICL, which corresponds to Section 5 in this survey. All the above-mentioned survey papers differ with ours in terms of scope and topics. This survey focused on the general development of ICL, including the formal definition of ICL, training strategies, prompt designing strategies, analysis and applications.