The Good, the Bad, and the Debatable: A Survey on the Impacts of Data for In-Context Learning

Anonymous ACL submission

Abstract

In-context learning is an emergent learning paradigm that enables an LLM to learn an unseen task by seeing a number of demonstrations in the context window. The quality of the demonstrations is of paramount importance as 1) context window size limitations restrict the number of demonstrations that can be presented to the model, and 2) the model must identify the task and potentially learn new, unseen input-output mappings from the limited demonstration set. An increasing body of work has also shown the sensitivity of predictions to perturbations on the demonstration set. Given this importance, this work presents a survey on the current literature pertaining to the relationship between data and in-context learning. We present our survey in three parts: the "good" – qualities that are desirable when selecting demonstrations, the "bad" – qualities of demonstrations that can negatively impact the model, as well as issues that can arise in presenting demonstrations, and the "debatable" – qualities of demonstrations with mixed results or factors modulating data impacts.

1 Introduction

In-context learning (ICL) is an emergent capability of large language models (LLMs) that allows them to learn new tasks at inference time without any parameter updates (Wei et al., 2022a). By providing a few examples (demonstrations) within the context window (as illustrated in Figure 2), LLMs can effectively "learn" in context and generalize to unseen tasks (Brown et al., 2020). This is different from traditional fine-tuning, which requires updating the model's parameters to learn a specific task. ICL, on the other hand, can infer from demonstrations directly during prediction and leave model parameters unchanged.

In ICL, performance depends on two key factors: 1) the base LLM and its prompt formatting capabilities, and 2) the provided demonstrations in-context.

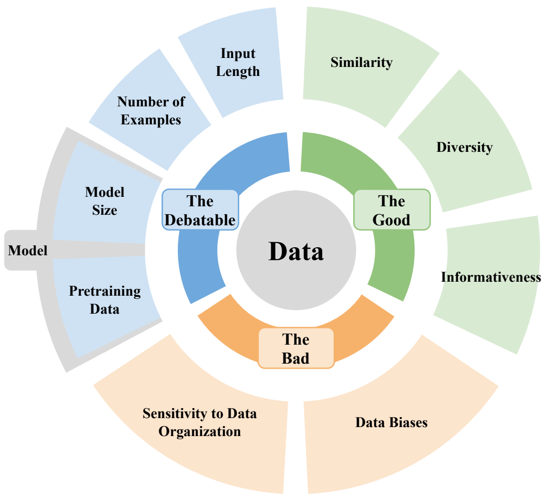

Figure 1: The data-centric view of the topics covered in this survey.

While the importance of the base model is well-established, a systematic analysis of ICL from the perspective of demonstration data has been largely overlooked.

However, the data used in ICL is crucial for both its performance and robustness, making it essential to study. For example, different selected examples can cause instability in performance, thereby causing a robustness issue dependent on the selected examples (Rubin et al., 2022; Liu et al., 2022; Wu et al., 2023; Zhao et al., 2021). Therefore, while previous work has given a broad overview of the ICL literature (Dong et al., 2024) and focused on theoretical interpretations of ICL (Zhou et al., 2024d), our work differs in that we take a data-centric angle to analyze the current work on ICL. Specifically, our work focuses on the impact of the demonstration data on ICL. As shown in Figure 1, we structure our survey in three parts: 1) the "good" qualities of ICL data (section 3), 2) the "bad" qualities of ICL data (section 4), and 3) the "debatable" qualities of ICL data (section 5), particularly as they relate to other components of

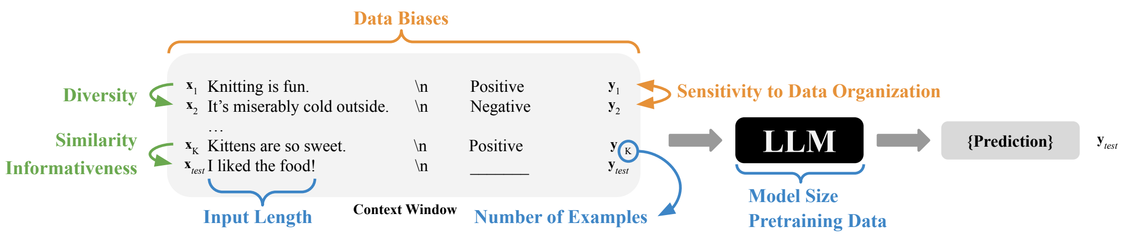

Figure 2: Overview of ICL using K input-output demonstrations concatenated to the test input $\{x_{test}, y_{test}\}$, overlaid with the topics covered in our survey (Good, Bad, Debatable).

the ICL paradigm.

2 Background

Brown et al. (2020) introduced in-context learning, where a model conditions on a few input-output pairings (demonstrations) concatenated to the target input in the context window. This enables the model to learn to perform a given task at inference, without any gradient updates. Formally, given a test example x_{test}, in-context learning concatenates K demonstrations to the task instruction I, where $S = \{x_i, y_i\}_{i=1}^{K}$ denotes the example set. The full context window of the model is provided as $C = \{I, S, x_{\text{test}}\}$. Brown et al. (2020) further identified few-shot ($K = n$), one-shot ($K = 1$), and zero-shot ($K = 0$) settings in in-context learning.

While "in-context learning" is the most common and descriptive term, other names have been used, sometimes interchangeably. For example, few-shot prompting (Wei et al., 2022a) has been used to refer to few-shot ICL (and sometimes even used synonymously with ICL in general (Lu et al., 2022; Ma et al., 2023)). Priming-based few-shot learning (Kumar and Talukdar, 2021) is another alternative. ICL can be considered a subcategory of prompt learning, as it incorporates demonstrations within the prompt. It is also related to traditional few-shot learning, which encompasses techniques like few-shot prompt-based fine-tuning or, simply, few-shot prompting (Köksal et al., 2023). Despite the variations, "in-context learning" remains the predominant term for the collection of methods described above and will be used in the rest of this survey.

3 The Good: Desirable Data Qualities for ICL

In this section, we address the question of what data qualities improve ICL performance by surveying demonstration selection methods. We identify and structure our discussion around three key aspects: similarity, diversity, and informativeness.

3.1 Similarity

Similarity focuses on the relationship between a test input and a candidate demonstration, typically computed using distance metrics to measure the similarity of embeddings. One approach is to use off-the-shelf embeddings (e.g. SBERT (Reimers and Gurevych, 2019)) in-conjunction with unsupervised similarity metrics. Liu et al. (2022) propose a k-nearest neighbor based retriever that selects the k semantically-similar candidates in embedding space for each test sample using cosine similarity or negative Euclidean distance. This method has been extended to cross-lingual settings (Tanwar et al., 2023). Shin et al. (2021) propose to instead directly use GPT-3 to select similar examples for few-shot semantic parsing, where the relevance of a training example $\{u_i, t_i\}$ to a test input u is computed using $p(u|u_i)$.

Rather than using off-the-shelf embeddings or directly using LLMs, other works aim to train a prompt retriever. Rubin et al. (2022) propose a method to learn embeddings for similarity-based retrieval, EPR. It first retrieves candidate examples using an unsupervised retriever (e.g. BM25 (Robertson et al., 2009)) and then uses these to train a dense retriever with contrastive learning. Finally, the trained retriever uses the example embeddings to select the top-k examples based on inner product similarity. Li et al. (2023) extend this to a unified, multi-task setting, and Hu et al. (2022) propose a similar method of two-stage learned embeddings for dialogue state tracking. Liu et al. (2024b) find that the previous methods learning similarity measurements work because they integrate task-agnostic similarities at different levels and incorporate task-specific similarity, and they propose two selection methods that address these factors.

While similarity considers the relationship between the test inputs and exemplars, considering the relationship between exemplars (i.e. diversity) is also effective, as discussed in the following section. Notably, most methods that utilize the diversity of examples also incorporate similarity.

3.2 Diversity

Diversity focuses on the relationship between candidate exemplars. Some methods incorporate diversity-enhancing components into learned retrievers, either at training or inference. Ye et al. (2023a) retrieve example sets using maximum a posteriori inference with a learned determinantal point process (DPP) module, where the DPP kernel is defined to incorporate both diversity and relevance. Liu et al. (2024a) propose a sequential example selection method that leverages LLM feedback to score candidate example sequences for training, then constructs diverse example sequences at inference using beam search.

Other works enhance diversity through iterative selection with penalty terms on similarity. Ye et al. (2023b) propose to iteratively select examples using maximum marginal relevance, incorporating a penalty term on similarity to already selected examples. Hongjin et al. (2022) iteratively select examples to annotate in a "select-then-annotate" paradigm, where candidate scores are discounted based on their graph-based similarity to previously selected examples. They further define a bucketing procedure to annotate examples across diverse model confidence scores, and finally select k examples from the annotated set using cosine similarity.

Similar to enhancing diversity through bucketing (Hongjin et al., 2022), other methods use intervals or clusters to select diverse examples. Zhang et al. (2023) use k-means clustering to select diverse exemplars. Yao et al. (2024) use intervals to select candidates across a diverse range of input-candidate similarity scores, which are then used in different prompts followed by a majority vote.

Finally, selecting diverse examples by diversifying the embedded representations of inputs has proven effective. Specifically, Qin et al. (2023) select the top-k examples based on the cosine similarity between each candidate exemplar and the zero-shot reasoning path on the test input, use the selected examples to generate a new reasoning path on the test input, iterate n times (selecting new examples with the updated reasoning paths each time), and perform majority voting. Notably, they argue that iterating on the reasoning path can enhance diversity by potentially selecting different examples in each iteration.

3.3 Informativeness

Informativeness of examples relates to the contribution of examples to the test input and has been defined both at the individual and set level. At the level of individual examples, Li and Qiu (2023) use LLM feedback to measure how informative an example is for the model to correctly classify the test input, and subsequently apply a diversity-guided search of permutations. Nguyen and Wong (2023) use the influence function (Koh and Liang, 2017) to select examples that have a positive impact on performance.

Beyond the level of individual example informativeness, notions of coverage have been used to select informative and diverse sets of examples. This includes syntactic and lexical coverage for machine translation (Tang et al., 2024) and substructure coverage for compositional generalization in semantic parsing (Levy et al., 2023). Gupta et al. (2023b) extend the notion of coverage to diverse tasks by selecting demonstration sets that are maximally informative for the salient aspects of the test input (e.g. reasoning patterns) using BERTScore-Recall (BSR). Related to information contained in the examples, Shi et al. (2023a) show that including examples with irrelevant information (i.e. distractors) can teach LLMs to ignore irrelevant context and help mitigate distractability on reasoning tasks.

3.4 Discussion

Similarity vs. Diversity: Task-Dependent Trade-offs. Several works point to a task- and dataset-dependence on the importance of similarity vs. diversity in selecting examples. When proposing in-context sampling (ICS), Yao et al. (2024) explored different sampling strategies: similarity (top-k based on cosine similarity of embeddings), diversity (k at different intervals based on cosine similarity, to capture more of the input space), and hybrid ($\frac{k}{2}$ from each). They found that no single strategy performed best across all datasets. Qin et al. (2023) found similar results when comparing random sampling (diversity setting) with similarity sampling. Other works that have shown impressive performance have directly acknowledged and accounted for this trade-off (Ye et al., 2023a,b).

Pre-Processed Input Representations & Other Information Sources. While many selection strategies directly utilize the embedded representations of test inputs and candidate exemplars, other works pre-process the inputs prior to embedding and subsequent selection, or otherwise incorporate richer information sources such as explanations. Qin et al. (2023) perform selection using the cosine similarity between candidate exemplars and iterative representations of the LLM's reasoning path on a test input. An et al. (2023) use an LLM to rewrite each candidate and test example using skill-based descriptions, and then using the cosine similarity between descriptions to select demonstrations. Other works incorporate the use of explanations Ye et al. (2023b) and chain-of-thought reasoning (Wei et al., 2022b) to enhance ICL performance. Expanding on the prior discussion on similarity and diversity, these factors are beneficial when using pre-processed representations and explanations as well (Ye et al., 2023b; Qin et al., 2023).

4 The Bad: Data Issues in ICL

In this section, we address the question of what qualities of data for ICL are undesirable, and what can go wrong when there are issues with the selected data. We center our discussion around: 1) sensitivity to data organization, and 2) data biases.

4.1 Sensitivity to Data Organization

LLMs are sensitive to the choice of selected examples (Zhao et al., 2021; Liu et al., 2022) as well as their order (Zhang et al., 2022; Chen et al., 2023b). Both organization factors are data and model dependent (Peng et al., 2024; Pecher et al., 2024). For example, the performance of example permutations cannot generalize across models, yet models of all sizes exhibit order sensitivity (Lu et al., 2022). Recent works have also shown a sensitivity to the position of relevant information in the context. Specifically, models are biased towards information at the beginning and end of the prompt in long-contexts (Liu et al., 2024c), shortcut triggers at the end of prompts (Tang et al., 2023), and labels that are proximal to the test input (Zhao et al., 2021; Li et al., 2024a; Nguyen and Wong, 2023) (covered in more detail in subsection 4.2). Another factor of data organization, the number of examples, is covered in section 5. Additionally, as we focus on the demonstrations themselves, the impact of prompt template is outside of the scope of our discussion. In the following subsection, we discuss mitigation strategies for sensitivity to example organization, with a particular focus on ordering.

4.1.1 Mitigating Ordering Sensitivity

Approaches to mitigating sensitivity to ordering can be categorized as: 1) identifying a good order of selected examples, 2) selecting examples simultaneously with their order, and 3) selecting examples with lower variance across permutations.

Select-*then*-Organize: Identifying an Effective Ordering. When selecting examples based on their similarity to the test input, one practice is to sort the examples in ascending order of similarity, with the most similar example the most proximal to the test input (Ye et al., 2023a; Rubin et al., 2022). Complexity, as measured by LLM perplexity, is also effective for ordering similar examples to the test input, from least to most complex in a curriculum learning framework (Liu et al., 2024d) Alternatively, Kumar and Talukdar (2021) use a genetic algorithm to search for a good permutation of demonstrations.

Concepts from information theory have also been effective to find optimal example orderings. Lu et al. (2022) propose local and global entropy metrics for demonstration reordering. Wu et al. (2023) propose an information-theory-driven ranking algorithm and find the best subset organization based on the codelength to compress and transmit label y given test input x and organization c. Guo et al. (2024) first filter candidate orderings using a content-free (Zhao et al., 2021) entropy metric, then select an order that maximizes the output influence of each test instance.

Select-*and*-Organize: Selecting Examples with Their Order. Approaches that focus on reordering examples may fail depending on the selected examples. Zhang et al. (2022) demonstrate that on TREC (Voorhees and Tice, 2000), even the best performing permutation of $k = 4$ examples ($4! = 24$ permutations) performs below a random baseline on 9 out of 30 selected example sets.

Sequential example selection can identify a good selection and permutation of examples. Ma et al. (2023) sequentially select a permutation of examples using entropy as a measure of predictive bias over labels, where higher entropy correlates with higher accuracy. Zhang et al. (2022) propose active example selection and use reinforcement learning to optimize a policy for sequential data selection and annotation. Liu et al. (2024a) sequentially select examples and score candidate example sequences using LLM feedback. These methods also increase stability across permutations (Liu et al., 2024a) and different unlabeled example pools (Zhang et al., 2022).

Selecting Stable Subsets. Rather than select-*and*-organize or select-*then*-organize paradigms, an alternative approach is to identify data subsets to sample from that are more robust to different orderings. Chang and Jia (2023) focus specifically on identifying stable data subsets to sample from, where stability is defined as having higher average and worst-case accuracy compared to sampling from the full training set. They propose two methods to find stable subsets: scoring each example by the average validation accuracy when combined with random examples (inspired by Data Shapley (Ghorbani and Zou, 2019)) and scoring each example based on the associated weights of a linear regression model fit to predict the LLM's output based on which example is present at each index in the prompt.

Zhao et al. (2021) suggested that instability and sensitivity to data organization arises from biases in models towards predicting certain answers. Interestingly, however, balanced labels do not consistently lead to greater performance or less variance across permutations than unbalanced labels (Zhang et al., 2022). We cover data biases, including label biases, in more detail in the following section.

4.2 Data Biases

In this section, we address two questions: 1) how do data biases impact the robustness and performance of ICL, and 2) how can negative impacts from data biases be mitigated?

4.2.1 Types of Data Biases

Based on the current literature, we identify and discuss two categories of data biases: shortcut learning and label biases.

Shortcut learning. Features learned by LLMs may be semantically meaningful (i.e. robust) or related to biases and spuriously correlated label mappings (non-robust) (Du et al., 2023). The learning of these features has been termed "shortcut learning" as it pertains to the model learning semantically irrelevant features that may not relate to the underlying task. While most previous studies look at settings with weight updates, recent works have demonstrated that LLMs can also learn shortcut features in the context window.

Token-level shortcut features learnable from demonstrations include letters, symbols, common words, rare words, and sentences (i.e. sequences of tokens) (Tang et al., 2023). At a higher level, features such as length (Schoch and Ji, 2025), text styles (Tang et al., 2023), and concepts (e.g. the concept "food" being spuriously correlated with a specific label) (Zhou et al., 2024c) have also been shown to be learnable from demonstrations. Tang et al. (2023) show there is a positional component in shortcut learning, where LLMs are particularly biased towards shortcuts placed at the end of prompts.

In addition to learning shortcut features from demonstrations, LLMs can exhibit shortcut behaviors on in-context demonstrations. Sun et al. (2024) show that LLMs can utilize reasoning shortcuts such as negation and word overlap in in-context settings. LLMs can also exhibit a tendency to instead copy answers from the exemplars, termed *copy bias*, rather than learning an underlying pattern in tasks that require novel responses (e.g. counting vowels) (Ali et al., 2024). Si et al. (2023) use underspecified demonstrations (where two features such as sentiment and topic are equally predictive of the label) to show that LLMs can exhibit *feature bias*, where the model is biased towards using one feature over the other. Jang et al. (2024) identified *demonstration bias* as the reliance of LLMs on semantic priors rather than learning new input-label relationships (discussed in more detail in section 5).

Label biases. In its simplest form, label bias refers to an undesirable behavior where a LLM predicts certain labels over others. Reif and Schwartz (2024) defined two measures to quantify label bias: relative standard deviation of class-wise accuracy (Croce et al., 2021; Benz et al., 2021), which is defined as the standard deviation of class-wise accuracy divided by the mean overall accuracy, and BiasScore, which is defined as the total variation distance between the estimated model output distribution and the uniform distribution over labels.

LLMs can acquire label biases through pretraining data and in-context demonstrations. Label bias acquired during pretraining has been termed *vanilla label bias* (Fei et al., 2023) and *common token bias* (Zhao et al., 2021). It can be thought of as the uncontextual preference of the model to predicting certain labels or answers, and may relate to the pretraining term frequencies (Fei et al., 2023). On multiple choice datasets, LLMs can also exhibit *selection bias* where the LLM exhibits a preference to select specific option IDs as answers (Zheng et al., 2024). Fei et al. (2023) also identify a further form of label bias that can be acquired during pretraining, *domain-label bias*, where the model relies on prior knowledge of the task when making predictions, based on learned associations between words and labels in pretraining.

The label bias acquired from demonstrations has been termed *context-label bias* (Fei et al., 2023). Both the distribution and position of labels in the demonstration set can bias outputs in ICL (Zhao et al., 2021). *Majority label bias* refers to the tendency of LLMs to predict labels that are seen frequently in the in-context examples, i.e. the distribution of in-context labels is skewed (Zhao et al., 2021; Gupta et al., 2023a). *Recency bias* occurs when the LLM is biased towards predicting labels seen at the end of the prompt (Zhao et al., 2021). Nguyen and Wong (2023) used influence to confirm recency bias, and Li et al. (2024a) demonstrated label recency bias in long-context LLMs. Notably, label recency bias has some connection to Tang et al. (2023) who found that LLMs were biased towards shortcut trigger placed at the end of prompts. While many of these works focus on classification tasks, Gao et al. (2024) extend the discussion to generation tasks, finding that label noise in demonstrations degrades ICL performance on generation tasks (i.e. noisy annotations on text generation tasks hurts performance).

While biases are generally problematic for performance and generalization, the presence of biases may also relate to observable robustness issues across different ICL configurations. (Zhao et al., 2021) suggested that label biases can cause high performance variance (i.e. instability) across different training examples, permutations, and prompt formats. Label bias also obscures sensitivity in ICL, yet sensitivity is important to quantify as predictions sensitive to perturbation are less likely to be correct (Chen et al., 2023b). In the next section, we discuss techniques to mitigate various data biases.

4.2.2 Mitigating Data Biases

In this section, we discuss methods that have been used to mitigate data biases. Notably, as data biases can lead to sensitivity to data organization, mitigation methods that address label biases often further address sensitivity to data organization.

One of the primary methods of mitigating label biases lies in calibrating the model's output distribution (i.e. shifting the decision boundary) using an estimated bias prior $\hat{\mathbf{p}} = \mathbf{p}(y \mid C)$, where $y \in \mathcal{Y}$ denotes the label set and C denotes the context. Zhao et al. (2021) propose to estimate this prior using a content-free input. Using $\hat{\mathbf{p}} = \mathbf{p}(y \mid [N/A], C)$, they define a calibration matrix $\mathbf{W} = \mathrm{diag}(\hat{\mathbf{p}})^{-1}$ and transform uncalibrated scores using $\mathbf{W}\mathbf{p}(y \mid x, C)$. This effectively shifts the output distribution so there is a uniform distribution over labels when using a content-free input. Fei et al. (2023) suggest that this cannot address "domain-label" biases arising from word-label associations of the task learned during pretraining. They propose to use random in-domain words rather than content-free inputs and averaging over M times, $\hat{\mathbf{p}} = \frac{1}{M}\sum_{j=1}^{M}\mathbf{p}(y \mid [\text{random}_{i.d.}]_j, C)$. They shift the output distribution by dividing by the prior,

$$\hat{y}_i = \text{argmax}_{y\in\mathcal{Y}} \frac{\mathbf{p}(y \mid x_i, C)}{\hat{\mathbf{p}}}. \quad (1)$$

Several works have suggested that methods using heuristics such as content-free or random in-domain inputs are too simplistic and may introduce new bias, and propose alternatives using the test inputs (Zhou et al., 2024a), generated sequences (Jiang et al., 2023), and in-context demonstrations (Reif and Schwartz, 2024). Zhou et al. (2024a) propose to directly use batches of M unlabeled test data, $\hat{\mathbf{p}} = \mathbf{p}(y \mid C)_j = \mathbb{E}_{x\sim P(x)}\Big[\mathbf{p}(y = y_j \mid x, C)\Big] \approx \frac{1}{M}\sum_{i=1}^{M}\mathbf{p}(y = y_j \mid x^{(i)}, C) \forall y_j \in \mathcal{Y}$ and calibrate the output probability with Equation 1. This is essentially shifting the decision boundary by the mean for each class and effectively aligns the score distribution to the estimated class mean to reduce any impact of label biases. Jiang et al. (2023) use the generative capabilities of LLMs to estimate the in-context label marginal using Monte Carlo sampling of generated sequences with $\hat{\mathbf{p}} = \frac{1}{L}\sum_{l=1}^{L}\mathbf{p}_{\text{LM}}\Big(\mathscr{T}(y) \mid \mathscr{D}(D_t^{\pi}) \oplus \mathscr{T}(x^l)\Big)$, where x^l is a generated sequence sampled from $\mathbf{p}_{\text{LM}}\Big(\mathscr{T}(y) \mid \mathscr{D}(D_t^{\pi})\Big)$. This value is then plugged back into Equation 1. Reif and Schwartz (2024) obtain output probabilities $p^i(y)$ for each in-context example using a leave-one-out method. They then average the output probabilities for each label and obtain $\hat{\mathbf{p}}$ using the mean of the intra-label averages $\hat{\mathbf{p}}(y) = \frac{1}{Y}\sum_{l\in Y}\left(\frac{1}{|D_l|}\sum_{y^i\in D_l} p^i(y)\right)$, where $D_l = \{p^i \mid y^i = l\}$. Calibration parameters are then computed as in (Zhao et al., 2021). Jang et al. (2024) similarly estimate the semantic prior on labels using a leave-one-out method on the demonstrations that additionally incorporates an estimate of the word-by-word semantic distribution using random shuffling (and use Equation 1). Estimation of bias priors has also shown effective for mitigating selection bias for option IDs in multiple choice datasets (Zheng et al., 2024).

Alternatively, ome calibration methods adopt statistical models to calibrate the output distribution. Han et al. (2023b) use a Gaussian Mixture Model to learn a robust decision boundary, and Nie et al. (2022) augment predictions with a k-nearest-neighbor classifier over a datastore.

Rather than calibrating the model output distribution externally, other works aim to calibrate the internal mechanisms of the model. Zhao et al. (2024) add noise to the model parameters to minimize the impact of pretrained token and label biases. To calibrate the model's prediction bias, they perturb model parameters using random noise sampled from a normal distribution $\mathcal{N}(0, \sigma^2)$ with intensity hyperparameter λ. This allows interpolation between each parameter θ_i and the noise matrix using $\theta_i' = (1-\lambda)\theta_i + \lambda\mathcal{N}(0, \sigma^2)$. Other works aim to identify and mitigate components responsible for the bias. Zhou et al. (2024b) showed that label biases can stem from biased behaviors of attention heads and feed-forward network vectors and mitigated their impact via masking. Ali et al. (2024) use Integrated Gradients (Sundararajan et al., 2017) to identify neurons responsible for copy bias and mitigate their impact via pruning. The pruned models perform better and also lead to better task vectors (Hendel et al., 2023), indicating that bias neurons can interfere with the model's ability to learn the underlying task.

The design of in-context demonstrations and prompts can also be used to mitigate shortcut behaviors, such as designing prompts to reduce reliance on negation and overlap on reasoning tasks (Sun et al., 2024), using in-context demonstrations to mitigate length biases from fine-tuned models (Schoch and Ji, 2025), and using semantically-relevant labels to mitigate feature biases (Si et al., 2023). On generation tasks, noisy annotations can be identified and replaced with their nearest neighbors that are likely to be clean, using a perplexity-based method (Gao et al., 2024).

5 The Debatable: Open Questions in ICL

In this section, we discuss data qualities in ICL that have mixed results (ground truth labels, input length, number of examples) as well as the relationship between ICL demonstrations and the underlying model (model size, pretraining data). Within this discussion, we include some open questions.

Ground Truth Labels. Some work has suggested that correct input-label pairings have minimal impact on ICL performance (Min et al., 2022). However, other works have suggested that the importance of ground truth labels is dependent on the task and task difficulty (Madaan and Yazdanbakhsh, 2023; Yoo et al., 2022), experimental configuration (Yoo et al., 2022), and model size (Pan et al., 2023; Wei et al., 2024). While some work has begun to analyze the mechanisms responsible for how LLMs utilize label information (Wang et al., 2023) and the influence of semantic priors (Pan et al., 2023), the role of ground truth labels (and underlying mechanisms) in in-context learning remains an open research area.

Model Size. Increasing the size of models can increase the potential performance gains from in-context learning (Milios et al., 2023; Lu et al., 2022). However, it can also increase the potential for robustness issues stemming from the in-context demonstrations. This includes vulnerability to shortcut features (Tang et al., 2023; Schoch and Ji, 2025), input noise (Shi et al., 2023b), and label noise (Pan et al., 2023; Wei et al., 2024; Shi et al., 2023b) in the demonstrations. This underscores an important direction in accounting for potential trade-offs between performance and robustness under ICL settings with respect to model size. Some works posit that the vulnerability to noise may arise from the fact that larger models cover more hidden features whereas smaller models emphasize more hidden features (Shi et al., 2023b), or from the ability of larger models to override their pretrained priors in comparison to smaller models (Pan et al., 2023; Wei et al., 2024). Other works, however, have shown promise for smaller models to override semantic priors and learn new input-label mappings (Kossen et al., 2024; Jang et al., 2024).

Input Length. The impact of input length on ICL performance is not currently well-understood.

Chang and Jia (2023) did not find a correlation between good examples selected by their method and sequence length, other than a small negative correlation when sequence length is very long. Length information, however, can be learned by the model in-context (Schoch and Ji, 2025). Some other studies have incorporated length into their methods of analysis and label bias mitigation. Fei et al. (2023) calibrate output distributions using random in-domain word sequences of the average input text length. Min et al. (2022) selected examples with similar lengths to the test inputs in their analysis of ICL. However, it is unclear whether similar length to test inputs is important given the absence of results with dissimilar or otherwise varied lengths.

Number of Examples. There are currently a number of conflicting results regarding the number of examples to use for ICL. Some works have suggested that learning with few demonstrations outperforms zero-shot settings (Min et al., 2022), yet other work has shown this may not generalize to all datasets and models (Brown et al., 2020; Xie et al., 2022; Lin and Lee, 2024). Further, some works show conflicting results on performance plateaus. Wang et al. (2024) found performance plateaus at $k = 4$ under their LLM-R framework, whereas Min et al. (2022) found performance plateaus occurring at $k \geq 8$. They further suggested that aspects important for ICL such as the input distribution, label space, and input-output mapping format are easily recoverable from few examples, whereas larger amounts of data (such as in fine-tuning settings) are required to supervise input-label correspondence (Min et al., 2022).

The performance plateaus at $k \geq 8$ (Min et al., 2022), however, may be dependent on the specific organization (selection and order) of examples. Wu et al. (2023) observed similar plateaus at $k = 8$ when using a random baseline, but under their self-adaptive method for selecting a good organization of demonstrations, performance consistently increased from $k = \{0, 1, ..., 32\}$. Lu et al. (2022) similarly observed performance increases using $k = \{1, 2, ..., 32\}$, and further underscored the importance of ordering by noting that increasing the number of examples does not decrease the variance across permutations. Beyond sensitivity to ordering, Schoch and Ji (2025) demonstrated that increasing the number of examples can increase the sensitivity of the model to data biases in the demonstrations.

There are also task-specific considerations in the benefit or risk of increasing the number of examples. On reasoning tasks, Chen et al. (2023a) also showed that one example can outperform settings with more examples due to interference and spurious correlations that can arise between examples. On text generation tasks, Gao et al. (2024) showed that increasing the number of examples in the presence of noisy annotations can degrade performance, even when using selection methods such as top-k.

Pretraining Data. The pretraining data distribution is impactful on ICL learnability (Wies et al., 2023). Properties that have been identified as beneficial for the emergence of ICL include burstiness, a large number of rarely occurring classes (Chan et al., 2022), and diverse tasks (Kirsch et al., 2022; Yadlowsky et al., 2023; Raventós et al., 2024). While task diversity is important, in few-shot ICL settings pretraining data does not necessarily require domain relevance to the downstream task (Han et al., 2023a; Shin et al., 2022).

The pretraining data distribution can also impact the model's performance on different test data in-context. Pretraining label and token term frequencies can introduce bias into the model's output distribution (Zhao et al., 2021). Other work has demonstrated positive correlations between term frequencies and ICL performance on numerical reasoning tasks (Razeghi et al., 2022) and QA tasks (Kandpal et al., 2023). For models where the pretraining data is unknown, this can make the evaluation of ICL performance difficult to interpret (Razeghi et al., 2022).

6 Discussion & Conclusion

In this survey, we gave an overview on the relationship between data and in-context learning. Beyond the open issues raised in section 5, there are several important directions for data-centric ICL research. Notably, much of the current work on understanding data impacts in ICL are on reasoning and classification tasks. Extending our understanding on generation tasks (Gao et al., 2024), low-resource tasks (Patel et al., 2022), and long-context settings (Li et al., 2024b; Liu et al., 2024c; Bertsch et al., 2024) would greatly enrich the discussion. Additionally, a number of different theoretical interpretations of ICL have been proposed (Xie et al., 2022; Dai et al., 2023), and understanding ICL data through these lenses could serve as an interesting future direction.

7 Limitations

In this work, we aimed to provide a comprehensive, data-centric overview of the ICL literature. While we made every effort to include all of the relevant works, we may have overlooked some valuable contributions given the extensive and rapidly progressing state of ICL research. Additionally, to realistically constrain the scope of our survey, we did not include works on prompt template design. However, we acknowledge that the prompt template is an important design component that interacts with the ICL demonstrations. We leave a survey on prompt template design to future work.

References

Ameen Ali, Lior Wolf, and Ivan Titov. 2024. Mitigating copy bias in in-context learning through neuron pruning. *arXiv preprint arXiv:2410.01288*.

Shengnan An, Bo Zhou, Zeqi Lin, Qiang Fu, Bei Chen, Nanning Zheng, Weizhu Chen, and Jian-Guang Lou. 2023. Skill-based few-shot selection for in-context learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13472–13492, Singapore. Association for Computational Linguistics.

Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. 2021. Robustness may be at odds with fairness: An empirical study on class-wise accuracy. In *NeurIPS 2020 Workshop on pre-registration in machine learning*, pages 325–342. PMLR.

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. *arXiv preprint arXiv:2405.00200*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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*, volume 33, pages 1877–1901. Curran Associates, Inc.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 35:18878–18891.

Ting-Yun Chang and Robin Jia. 2023. Data curation alone can stabilize in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8123–8144, Toronto, Canada. Association for Computational Linguistics.

Jiuhai Chen, Lichang Chen, Chen Zhu, and Tianyi Zhou. 2023a. How many demonstrations do you need for in-context learning? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11149–11159, Singapore. Association for Computational Linguistics.

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. 2023b. On the relation between sensitivity and accuracy in in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 155–167, Singapore. Association for Computational Linguistics.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. 2021. Robustbench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. 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*, pages 4005–4019, Toronto, Canada. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2023. Shortcut learning of large language models in natural language understanding. *Communications of the ACM*, 67(1):110–120.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. 2023. Mitigating label biases for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14014–14031, Toronto, Canada. Association for Computational Linguistics.

Hongfu Gao, Feipeng Zhang, Wenyu Jiang, Jun Shu, Feng Zheng, and Hongxin Wei. 2024. On the noise robustness of in-context learning for text generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In

Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2242–2251. PMLR.

Qi Guo, Leiyu Wang, Yidong Wang, Wei Ye, and Shikun Zhang. 2024. What makes a good order of examples in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14892–14904, Bangkok, Thailand. Association for Computational Linguistics.

Karan Gupta, Sumegh Roychowdhury, Siva Rajesh Kasa, Santhosh Kumar Kasa, Anish Bhanushali, Nikhil Pattisapu, and Prasanna Srinivasa Murthy. 2023a. How robust are llms to in-context majority label bias? *arXiv preprint arXiv:2312.16549*.

Shivanshu Gupta, Matt Gardner, and Sameer Singh. 2023b. Coverage-based example selection for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13924–13950, Singapore. Association for Computational Linguistics.

Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. 2023a. 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)*, pages 12660–12673, Toronto, Canada. Association for Computational Linguistics.

Zhixiong Han, Yaru Hao, Li Dong, Yutao Sun, and Furu Wei. 2023b. Prototypical calibration for few-shot learning of language models. In *The Eleventh International Conference on Learning Representations*.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

SU Hongjin, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. 2022. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations*.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A. Smith, and Mari Ostendorf. 2022. In-context learning for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2627–2643, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joonwon Jang, Sanghwan Jang, Wonbin Kweon, Minjin Jeon, and Hwanjo Yu. 2024. Rectifying demonstration shortcut in in-context learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4294–4321, Mexico City, Mexico. Association for Computational Linguistics.

Zhongtao Jiang, Yuanzhe Zhang, Cao Liu, Jun Zhao, and Kang Liu. 2023. Generative calibration for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2312–2333, Singapore. Association for Computational Linguistics.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Louis Kirsch, James Harrison, Jascha Sohl-Dickstein, and Luke Metz. 2022. General-purpose in-context learning by meta-learning transformers. *arXiv preprint arXiv:2212.04458*.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR.

Abdullatif Köksal, Timo Schick, and Hinrich Schuetze. 2023. MEAL: Stable and active learning for few-shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 506–517, Singapore. Association for Computational Linguistics.

Jannik Kossen, Yarin Gal, and Tom Rainforth. 2024. In-context learning learns label relationships but is not conventional learning. In *The Twelfth International Conference on Learning Representations*.

Sawan Kumar and Partha Talukdar. 2021. Reordering examples helps during priming-based few-shot learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4507–4518, Online. Association for Computational Linguistics.

Itay Levy, Ben Bogin, and Jonathan Berant. 2023. Diverse demonstrations improve in-context compositional generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1401–1422, Toronto, Canada. Association for Computational Linguistics.

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

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

Xiaonan Li, Kai Lv, Hang Yan, Tianyang Lin, Wei Zhu, Yuan Ni, Guotong Xie, Xiaoling Wang, and Xipeng Qiu. 2023. Unified demonstration retriever for in-context learning. In *Proceedings of the 61st Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4644–4668, Toronto, Canada. Association for Computational Linguistics.

Xiaonan Li and Xipeng Qiu. 2023. Finding support examples for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6219–6235, Singapore. Association for Computational Linguistics.

Ziqian Lin and Kangwook Lee. 2024. Dual operating modes of in-context learning. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Haoyu Liu, Jianfeng Liu, Shaohan Huang, Yuefeng Zhan, Hao Sun, Weiwei Deng, Furu Wei, and Qi Zhang. 2024a. se^2: Sequential example selection for in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5262–5284, Bangkok, Thailand. Association for Computational Linguistics.

Hui Liu, Wenya Wang, Hao Sun, Chris Xing Tian, Chenqi Kong, Xin Dong, and Haoliang Li. 2024b. Unraveling the mechanics of learning-based demonstration selection for in-context learning. *arXiv preprint arXiv:2406.11890*.

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 (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024c. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

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

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)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided few-shot prompting for large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Aman Madaan and Amir Yazdanbakhsh. 2023. Text and patterns: For effective chain of thought it takes two to tango.

Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. 2023. In-context learning for text classification with many labels. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pages 173–184, Singapore. Association for Computational Linguistics.

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

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

Feng Nie, Meixi Chen, Zhirui Zhang, and Xu Cheng. 2022. Improving few-shot performance of language models via nearest neighbor calibration. *arXiv preprint arXiv:2212.02216*.

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

Ajay Patel, Nicholas Andrews, and Chris Callison-Burch. 2022. Low-resource authorship style transfer with in-context learning. *CoRR*.

Branislav Pecher, Ivan Srba, and Maria Bielikova. 2024. On sensitivity of learning with limited labelled data to the effects of randomness: Impact of interactions and systematic choices. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 522–556, Miami, Florida, USA. Association for Computational Linguistics.

Keqin Peng, Liang Ding, Yancheng Yuan, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2024. Revisiting demonstration selection strategies in in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9090–9101, Bangkok, Thailand. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Chen Chen, Anirudh Dagar, and Wenming Ye. 2023. In-context learning with iterative demonstration selection. *arXiv preprint arXiv:2310.09881*.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. 2024. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *Advances in Neural Information Processing Systems*, 36.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuval Reif and Roy Schwartz. 2024. Beyond performance: Quantifying and mitigating label bias in LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6784–6798, Mexico City, Mexico. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

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.

Stephanie Schoch and Yangfeng Ji. 2025. In-context learning (and unlearning) of length biases. *Preprint*, arXiv:2502.06653.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023a. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Zhenmei Shi, Junyi Wei, Zhuoyan Xu, and Yingyu Liang. 2023b. Why larger language models do in-context learning differently? In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Richard Shin, Christopher Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7699–7715, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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)*, pages 11289–11310, Toronto, Canada. Association for Computational Linguistics.

Zechen Sun, Yisheng Xiao, Juntao Li, Yixin Ji, Wenliang Chen, and Min Zhang. 2024. Exploring and mitigating shortcut learning for generative large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6883–6893, Torino, Italia. ELRA and ICCL.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Chenming Tang, Zhixiang Wang, and Yunfang Wu. 2024. SCOI: Syntax-augmented coverage-based in-context example selection for machine translation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9956–9971, Miami, Florida, USA. Association for Computational Linguistics.

Ruixiang Tang, Dehan Kong, Longtao Huang, and Hui Xue. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4645–4657, Toronto, Canada. 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)*, pages 6292–6307, Toronto, Canada. Association for Computational Linguistics.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, page 200–207, New York, NY, USA. Association for Computing Machinery.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. 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*, pages 9840–9855, Singapore. Association for Computational Linguistics.

Liang Wang, Nan Yang, and Furu Wei. 2024. Learning to retrieve in-context examples for large language models. *Accepted to EACL 2024*.

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. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2024. Larger language models do in-context learning differently.

Noam Wies, Yoav Levine, and Amnon Shashua. 2023. The learnability of in-context learning. *Advances in Neural Information Processing Systems*, 36:36637–36651.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. 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)*, pages 1423–1436, Toronto, Canada. 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 *International Conference on Learning Representations*.

Steve Yadlowsky, Lyric Doshi, and Nilesh Tripuraneni. 2023. Pretraining data mixtures enable narrow model selection capabilities in transformer models. *arXiv preprint arXiv:2311.00871*.

Bingsheng Yao, Guiming Chen, Ruishi Zou, Yuxuan Lu, Jiachen Li, Shao Zhang, Yisi Sang, Sijia Liu, James Hendler, and Dakuo Wang. 2024. More samples or more prompts? exploring effective few-shot in-context learning for LLMs with in-context sampling. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1772–1790, Mexico City, Mexico. Association for Computational Linguistics.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. 2023a. Compositional exemplars for in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 39818–39833. PMLR.

Xi Ye, Srinivasan Iyer, Asli Celikyilmaz, Veselin Stoyanov, Greg Durrett, and Ramakanth Pasunuru. 2023b. Complementary explanations for effective in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4469–4484, Toronto, Canada. Association for Computational Linguistics.

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*, pages 2422–2437, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

Yufeng Zhao, Yoshihiro Sakai, and Naoya Inoue. 2024. Noisyicl: A little noise in model parameters calibrates in-context learning. *arXiv preprint arXiv:2402.05515*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine A Heller, and Subhrajit Roy. 2024a. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *The Twelfth International Conference on Learning Representations*.

Hanzhang Zhou, Zijian Feng, Zixiao Zhu, Junlang Qian, and Kezhi Mao. 2024b. Unibias: Unveiling and mitigating LLM bias through internal attention and FFN manipulation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yuhang Zhou, Paiheng Xu, Xiaoyu Liu, Bang An, Wei Ai, and Furong Huang. 2024c. Explore spurious correlations at the concept level in language models for text classification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 478–492, Bangkok, Thailand. Association for Computational Linguistics.

Yuxiang Zhou, Jiazheng Li, Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. 2024d. The mystery of in-context learning: A comprehensive survey on interpretation and analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14365–14378, Miami, Florida, USA. Association for Computational Linguistics.