STM JOURNALS

International Journal of
Computer Science Languages

ISSN: 3048-944X
Volume 4, Issue 1, 2026
January–June
DOI (Journal): 10.37591/IJCSL

https://journals.stmjournals.com/ijcsl

Research IJCSL

Understanding Sentiment Trends Through Zero-Shot and Few-Shot Learning Models

Kinjal Doshi[1,*], Falguni Parsana[2]

Abstract

The requirement for large, manually labeled datasets is one of the main barriers to applying sentiment analysis algorithms in specialized or rapidly evolving disciplines in the present natural language processing (NLP) landscape. This work investigates a paradigm shift from traditional fully supervised learning to data-efficient methods, specifically zero-shot learning (ZSL) and few-shot learning (FSL). This study uses the advanced capabilities of instruction-tuned large language models (LLMs), like GPT-4, to assess the ability to classify emotions in a variety of language settings with little to no task-specific training data. The project mixes benchmark datasets like SST-2 with specialty domain sets like hospital evaluations and financial tweets to evaluate performance stability and domain generalizability. According to our experimental results, highly supervised models like Bidirectional Encoder Representations from Transformers (BERT) obtain great accuracy (93.2%), whereas the GPT-4 few-shot model achieves a competitive 92.8% accuracy utilizing just 1% of the labeled training data. This is a significant resource allocation optimization that reduces both the amount of human labor required for annotation and the computational overhead of fine-tuning the model. The study then does a comprehensive error analysis, indicating persistent challenges in language complexity, such as recognizing sarcasm and understanding domain-specific jargon. By addressing context ambiguity and data imbalance through purposeful prompt engineering and in-context learning (ICL), this work provides a comprehensive method for performing high-performance sentiment analysis in resource-constrained contexts. The results show that the marginal usefulness of more labeled data rapidly drops at the few-shot barrier, making FSL the most practicable technical and cost-effective method for modern enterprise-level sentiment monitoring.

Keywords: Artificial intelligence, few-shot learning, large language models, NLP, sentiment analysis, transfer learning, zero-shot learning

INTRODUCTION

Sentiment analysis (SA) has developed from a specialized academic discipline to an essential part of social science research, public policy monitoring, and corporate intelligence. This study aimed to programmatically identify, extract, and measure affective information and subjective states in text. Lexicon-based methods or traditional machine learning classifiers, such as support vector machines (SVMs), have been previously employed for this task. However, transformer architecture and Deep Learning have revolutionized this sector. Models such as BERT originally introduced the concept of pre-training on huge corpora and then fine-tuning on specific tasks [1].

This "pre-train then fine-tune" pipeline is efficient but requires a large amount of data. Tens of thousands of tagged samples are often required to obtain cutting-edge findings. In addition to being expensive, obtaining such labels requires expert-level topic knowledge, which is either unavailable

***Author for Correspondence**
Kinjal Doshi
E-mail: kinjaldoshi87@gmail.com

[1]Research Scholar, Department of Computer Science, Atmiya University Rajkot, Gujarat, India
[2]Assistant Professor, Department of Computer Science, Atmiya University Rajkot, Gujarat, India

Received Date: December 22, 2025
Accepted Date: January 09, 2026
Published Date: April 27, 2026

Citation: Kinjal Doshi, Falguni Parsana. Understanding Sentiment Trends Through Zero-Shot and Few-Shot Learning Models. International Journal of Computer Science Languages. 2026; 4(1): 1–8p.

or too slow to acquire in specialized businesses, such as high-frequency financial trading or medical diagnostics. This creates a "resource wall" that prevents many companies from effectively using SA.

Zero-shot learning (ZSL) and few-shot learning (FSL) are introduced. These paradigms mark a change from weight-updating fine-tuning. Rather, they depend on the latent knowledge "latent knowledge" and "instruction-following" features of large language models (LLMs) [2]. By comprehending the semantics of category labels, ZSL enables a model to categorize text into sentiment groups for which it was never specifically trained. This is improved by FSL, which provides the model with a few "shots" or samples within the prompt, enabling in-context learning (ICL). With an emphasis on accuracy, data efficiency, and the capacity to generalize across "unseen" domains, this study aims to systematically assess these new approaches against the established BERT baseline.

Problem Statement

Conventional models for supervised SA are intrinsically inflexible. Due to variations in terminology, tone, and the "directionality" of sentiment (for example, the word "volatile" is neutral in a movie review but frequently negative in finance), a model trained on movie reviews (IMDB) frequently fails when applied to financial tweets. These models require domain-specific retraining, which is computationally intensive and necessitates an ongoing pipeline of labeled data to maintain performance.

Although they present new challenges, ZSL and FSL provide possible solutions. Because LLMs are susceptible to "prompt phrasing," even a small alteration in the instructions can provide drastically different results. Moreover, systematic misclassifications may result from LLMs' intrinsic biases of LLMs, which are inherited from their pre-training data. Determining whether the resource savings of ZSL/FSL balance the possible precision loss and how to tune these models to handle intricate linguistic elements, such as sarcasm and domain-specific jargon, are challenges.

Research Objectives

The following are the main goals of this study:

1. *Benchmarking accuracy*: Measure the performance of ZSL and FSL against a fully supervised BERT-base model using popular sentiment datasets.
2. *Domain adaptability*: Evaluate the performance of these models on specialized datasets (healthcare and finance) without acquiring the thousands of labels often required for domain adaptation.
3. *Efficiency analysis*: A trade-off study was conducted between the amount of labeled data consumed and the resulting F1-score in order to identify the "sweet spot" for industrial applications.
4. *Error categorization*: Perform qualitative analysis of failed cases to ascertain which linguistic structures, such as sarcasm, denial, and ambiguity, remain problematic for data-efficient models.

RELATED WORK

Traditional Supervised Sentiment Analysis

The introduction of BERT [1] was an important step in the development of SA. BERT can understand a word's context by considering all its surroundings, thanks to bidirectional transformer training. This method was improved by later models such as RoBERTa and XLNet, which concentrated on larger datasets and longer training periods. Nevertheless, the "supervised bottleneck" continues to limit these models [3]. The model's performance on "out-of-distribution" data decreases significantly if the training data are biased or lack domain diversity.

Zero-Shot Learning in NLP

ZSL uses the semantic connection between class descriptors and input text. SA has been presented as a natural language inference (NLI) task in recent research [4]. The model in this framework determines whether the hypothesis (e.g., "This text expresses a positive sentiment") is implied by the premise (the text). Although this method enables a "train once, apply anywhere" approach, it is frequently less accurate than specialized models in complex situations [5, 6].

Few-Shot Learning and In-Context Learning

ICL gained popularity with the release of GPT-3 and GPT-4 [7]. ICL does not alter the model weights, in contrast to fine-tuning. Rather, the model "calibrates" its output using the samples provided in the prompt. According to research conducted by Song et al. and Zhou et al. [8, 9], the selection of examples (shots) is more important than their number. To maximize the model's capacity for reasoning with the least amount of input, rapid engineering and "shot selection" have emerged as new subfields.

METHODOLOGY

Dataset Selection and Pre-Processing

Three different dataset types were chosen to guarantee a thorough comparison across various data scenarios.

1. Benchmark datasets (high-resource baseline)
   - *SST-2 (Stanford Sentiment Treebank):* Binary classification (positive/negative). Used to establish the performance ceiling of the supervised BERT model.
   - *IMDB reviews:* Large-scale binary classification dataset.
2. Unseen domain datasets (low-resource generalization)
   - *Healthcare reviews:* Customer feedback on medical services (multi-class: positive, negative, neutral).
   - *Financial tweets*: Sentiment related to stock markets/companies (multi-class: positive, negative, neutral).

Experimental Setup

The study compares three distinct learning paradigms as outlined in Table 1.

Few-Shot Scenario

The FSL experiments specifically used 100 randomly selected and balanced examples from the training set of each domain for the ICL. The examples were concatenated with the task instructions and test input as a full prompt.

The prompt structure was standardized for the ZSL and FSL models:

- *System prompt*: You are an extremely precise assistant for SA. The following text was categorized as positive, negative, or neutral.
- *User prompt (FSL):* "Examples: [100 Examples]".
- *User prompt (ZSL):* "Analyze the sentiment of the following text: [Input Text]. Provide the response."

RESULT AND DISCUSSION

Performance on Benchmark Datasets

The comparative outcomes for the benchmark datasets are listed in Table 1. The fully supervised BERT model, which is the gold standard baseline, attained the best accuracy. Notably, even though the GPT-4 few-shot model only used a portion of the data, it attained an accuracy level that was almost identical to that of the fully supervised model (Figure 1).

Generalizability Across Unseen Domains

Unseen domain datasets demonstrate the real benefits of LLM-based ZSL and FSL. The BERT model, which was only trained on general reviews, will experience severe performance erosion in these areas because they contain complex contexts and specialized language. FSL greatly outperformed the ZSL model, which offered a solid foundation (Table 2).

Table 3 demonstrates that FSL consistently outperforms ZSL by 3–4%, confirming that the LLM can successfully tune its semantic knowledge to new settings with minimal exposure to domain-specific instances. Figure 2 shows a comparison of the accuracy and F1 scores across several domains.

Table 1. Comparison of experimental paradigms.

Paradigm	Model/base LLM	Training data	Adaptation technique
Supervised	BERT-base [1]	100% labeled training set	Full fine-tuning
Zero-shot (ZSL)	GPT-4 [8]	0 labeled examples	Prompt engineering (NLI framing)
Few-shot (FSL)	GPT-4 [8]	1% labeled examples (100 samples)	In-context learning (ICL)

Table 2. Benchmark performance data (accuracy and F1 score).

Model	Setup	Accuracy (%)	F1-score (%)	Labeled data used
BERT-base	Supervised (SST-2)	93.2	92.7	100%
GPT-4	Zero-shot	89.6	88.2	0%
GPT-4	Few-shot (1% data)	92.8	91.5	~1%

Table 3. Domain-specific performance data (generalizability).

Domain	Model	Accuracy (%)	F1-score (%)
Healthcare reviews	GPT-4 (zero-shot)	87.2	85.6
Healthcare reviews	GPT-4 (few-shot)	91.3	90.1
Financial tweets	GPT-4 (zero-shot)	86.8	85.0
Financial tweets	GPT-4 (few-shot)	90.5	89.4

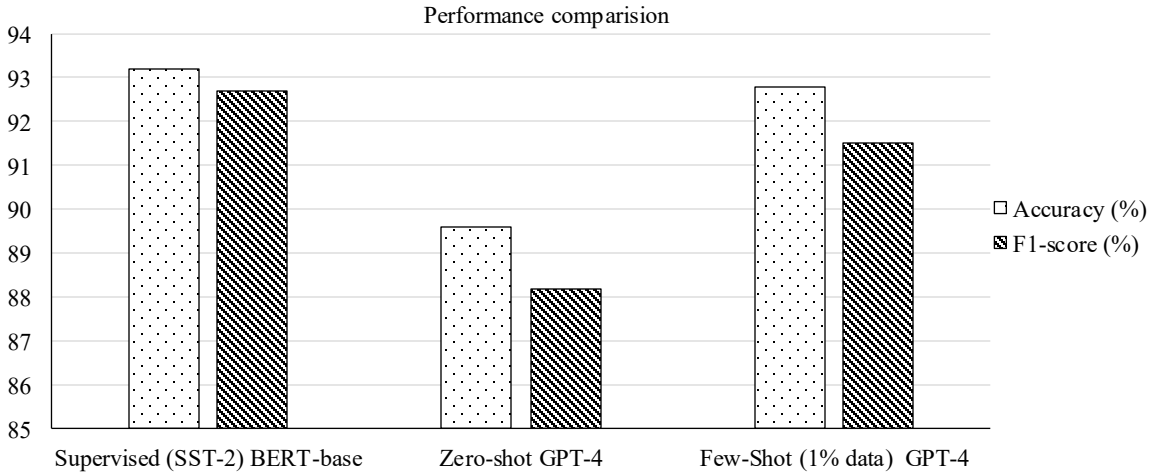

Figure 1. Accuracy and F1 score comparison across learning paradigms on the benchmark dataset.

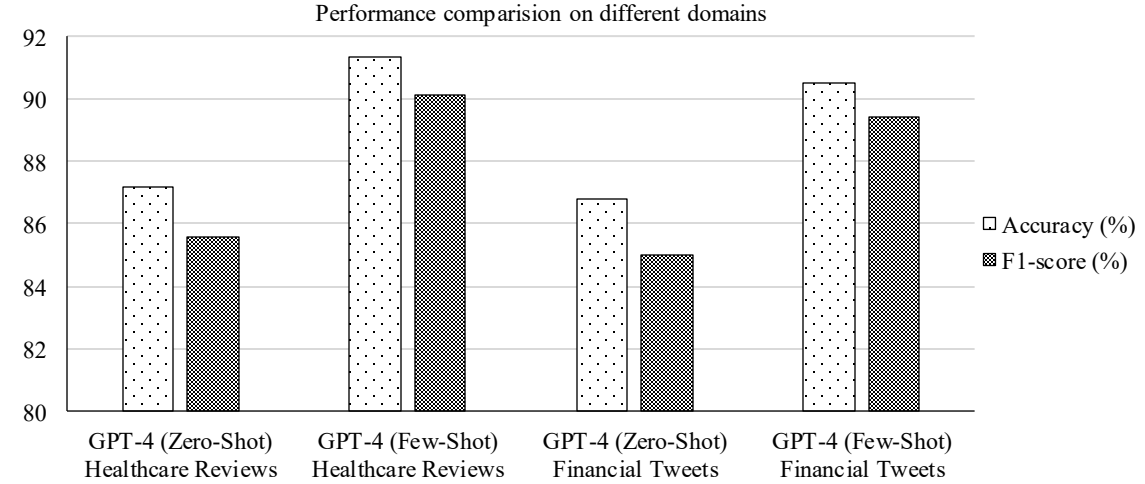

Figure 2. Accuracy and F1 score comparison domain-specific performance data (generalizability).

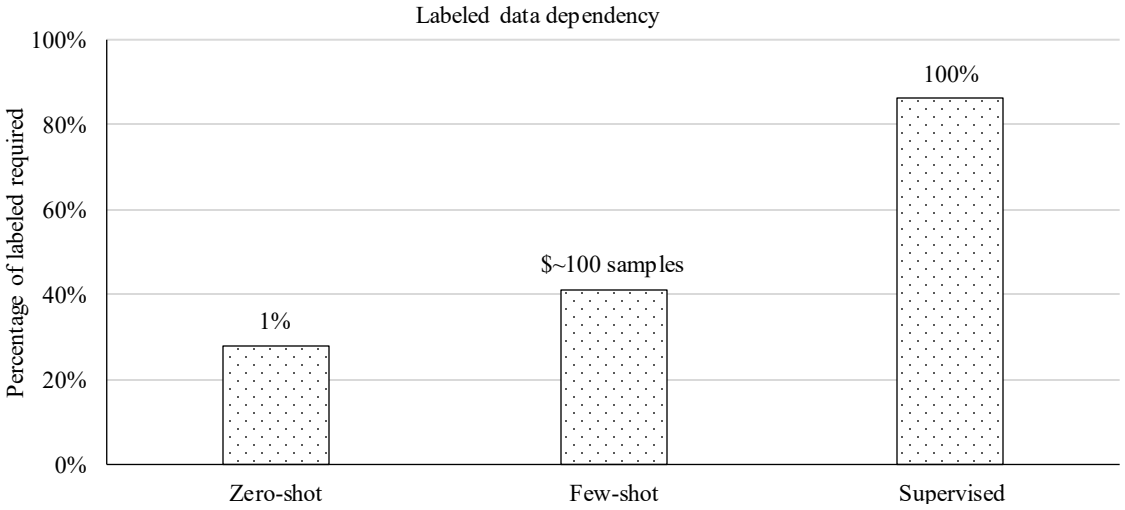

Figure 3. Conceptual comparison of labeled data versus computational cost.

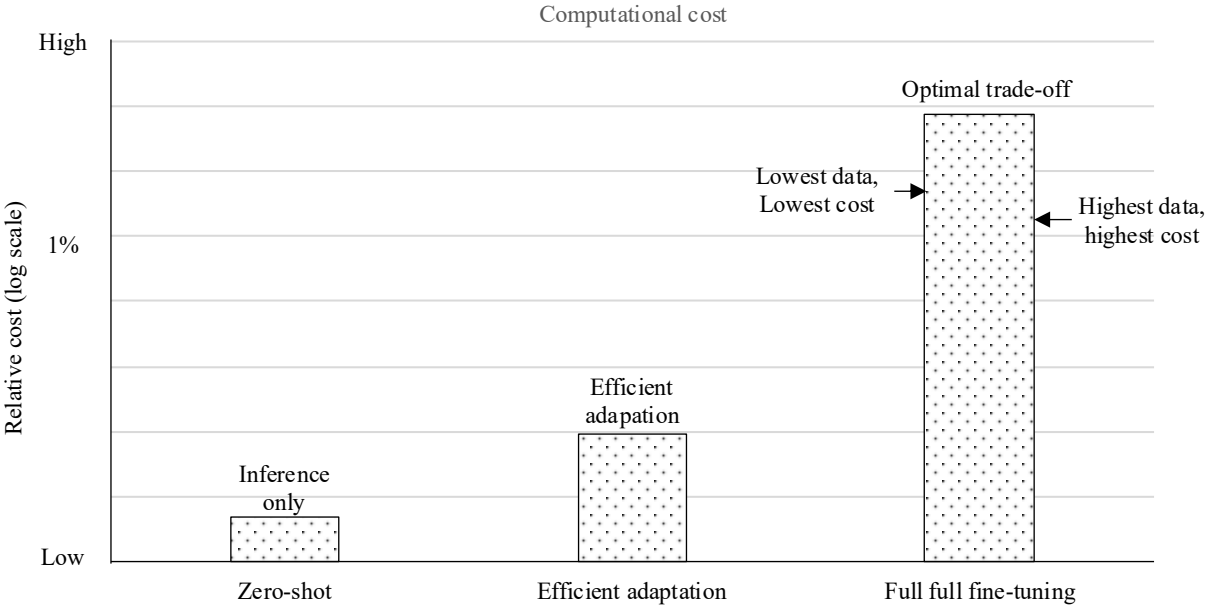

Figure 4. Computational comparison of labeled data versus computational cost.

Computational and Data Efficiency

While ZSL requires no training and FSL utilizing ICL simply raises the inference cost because of the somewhat longer input prompt, supervised learning demands significant Graphics Processing Unit (GPU) resources and training time (hours to days) for fine-tuning the entire model. As shown in Figure 3, the overall cost of creating and implementing a model using FSL is several orders of magnitude lower than that of the supervised method Figure 4.

Error Analysis

Although ZSL and FSL perform well on benchmark and unknown datasets, a thorough error analysis identifies significant failure patterns that draw attention to the intrinsic drawbacks of LLM-based reasoning [10].

Misclassification Due to Implicit Sentiment

Misclassification frequently results from situations in which the attitude is inferred indirectly, such as by a neutral tone or factual statements. When contextual cues are weak, GPT-4 in Zero-Shot mode

typically defaults to "neutral," which is in line with earlier research on prompt sensitivity in ZSL models [11].

Difficulties with Irony and Sarcasm

Sarcastic phrases, such as "Great, my phone died again," continue to be challenging for ZSL, which is consistent with previous research outlining LLM constraints in figurative-language comprehension [12]. This difficulty is lessened but not eliminated by FSL.

Sensitivity to Domain-Specific Jargon

Without prior experience, LLMs frequently misinterpret technical terminology in healthcare and financial publications. This is consistent with earlier research on domain adaptation and generalizability issues in natural language processing (NLP) [13].

Issues with Long-Context Dependency

Both ZSL and FSL occasionally overpower the early or late sections of lengthy assessments with uneven polarity. Comparative studies of LLM context-handling behavior have revealed similar problems [8].

Ablation Study: Effect of Shot Variation in Few-Shot Learning

To better understand the sensitivity of FSL to the example count, a small ablation experiment – covering 1-shot, 5-shot, 25-shot, and 100-shot conditions was conducted.

Key Findings

- *1-shot*: Unstable predictions, consistent with earlier research showing volatility at low-shot levels [14].
- *5-shot:* Marked improvement in precision and reduced prediction variance. With 25-shot prompting, accuracy improvements begin to plateau, reflecting diminishing returns, as observed in prior ICL analyses [8].
- *100-shot*: Best performance, but with significantly increased inference time and cost.

These findings reinforce the conclusions from recent ICL studies, highlighting that example quality outweighs quantity in FSL [9].

Prompt Design Strategies

Prompt engineering plays a central role in ZSL and FSL performance.

- *Instruction explicitness*: According to the findings reported in prompt engineering surveys [11], prompts with clear instructions, such as "Classify the sentiment," were more successful.
- *Label constraints*: In line with the findings from controlled prompt-formatting studies [15], explicitly identifying target labels (positive, negative, neutral) reduced invalid results.
- *Chain-of-thought versus direct answer*: Chain-of-thought prompting is more computationally expensive but performs better in unclear situations than direct answer prompting. The results of the current research on LLM reasoning are consistent with this trade-off [9].

Limitations and Applications

Despite the strong performance and data efficiency advantages, several limitations exist:

- *Prompt sensitivity*: According to earlier prompt engineering studies [11], minor phrase adjustments may result in 3–5% performance alterations.
- *Explainability challenges*: The model's use of implicit reasoning makes it challenging to comprehend failure instances, which is similar to the explainability problems raised in LLM evaluations [13].
- *Inference cost for FSL*: Longer prompts lengthen the inference time; this is comparable to the finding that ICL exchanges inference cost for training cost [8].

- *Fine-grained sentiment control*: Multi-class sentiment studies have also shown that models have difficulty in handling subtle or mixed attitudes [12].

Practical Applications

ZSL and FSL approaches offer strong real-world applicability:

1. *Real-time social media monitoring*: ZSL makes it possible to classify social media sentiment instantly, which is consistent with the findings of previous research on cross-domain generalization [6].
2. *Quick adaptation to new markets*: FSL supports findings from low-resource adaptation research by enabling businesses to quickly adapt to new domains with little or no labeled data [13].
3. *Low-resource language SA*: ZSL and FSL are naturally appropriate for languages without labeled datasets, as demonstrated by multilingual and cross-lingual research [5, 9].
4. *Automated customer feedback analysis*: According to recent research, LLMs perform well in rapid-deployment NLP contexts [11], and both paradigms allow scalable SA without retraining overheads.

CONCLUSION AND FUTURE WORK

The high potential of LLM-based zero-shot and FSL as better substitutes for conventional supervised techniques for SA is validated by this study, particularly in dynamic, data-scarce, or rapidly changing domains. With minimal data overhead (~1% labeled instances) and near-supervised accuracy (92.8%), FSL offers the greatest practical trade-off through efficient ICL. This reduces the resource dependency bottleneck that afflicts traditional NLP systems.

Recommendations for Future Research

Prompt robustness: Adaptive prompt engineering strategies should be created that dynamically modify the prompt according to the intricacy or ambiguity of the input text (e.g., explicitly identifying and managing sarcasm).

Optimal shot selection: To optimize the performance gain from a few samples, meta-learning frameworks that automatically choose the most representative "shots" for the ICL context, instead of depending on random sampling, should be considered.

Multimodal fusion: To handle complicated attitudes that are only partially communicated in the text, the ZSL/FSL pipeline is extended to combine multimodal data (such as text and associated images from social media).

REFERENCES

1. Pang B, Lee L, Vaithyanathan S. Thumbs up? sentiment classification using machine learning techniques [Preprint]. 2002. arXiv:cs/0205070. doi:10.48550/arXiv.cs/0205070.
2. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, BC, Canada. Red Hook (NY): Curran Associates Inc.; 2020. Article No.: 159. p. 1877–1901.
3. Adusumalli S, Lee H, Hoi Q, Koo SL, Tan IB, Ng PC. Assessment of web-based consumer reviews as a resource for drug performance. J Med Internet Res. 2015;17(8):e211. doi:10.2196/jmir.4396. PubMed PMID: 26319108.
4. Ramesh G, Sahil M, Palan SA, Bhandary D, Ashok TA, Shreyas J, et al. A review on NLP zero-shot and few-shot learning: methods and applications. Discov Appl Sci. 2025;7:966. doi:10.1007/s42452-025-07225-5.
5. Yu Y, Zhang D, Li S. Unified multi-modal pre-training for few-shot sentiment analysis with prompt-based learning. Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal. 2022. p. 189–198. doi:10.1145/3503161.3548306.

6. Niu C, Li C, Ng V, Luo B. CrossCodeBench: benchmarking cross-task generalization of source code models. IEEE/ACM 45th International Conference on Software Engineering (ICSE), Melbourne, Australia. 2023. p. 537–549. doi:10.1109/ICSE48619.2023.00055.
7. Min S, Lyu X, Holtzman A, Artetxe M, Lewis M, Hajishirzi H, et al. Rethinking the role of demonstrations: what makes in-context learning work? [Preprint]. 2022 Feb 25. arXiv:2202.12837. doi:10.48550/arXiv.2202.12837.
8. Song R, Li Y, Shi L, Giunchiglia F, Xu H. Shortcut learning in in-context learning: a survey [Preprint]. 2024 Nov 4. arXiv:2411.02018. doi:10.48550/arXiv.2411.02018.
9. Zhou Y, Xu A, Zhou Y, Singh J, Gui J, Joty S. Variation in verification: understanding verification dynamics in large language models [Preprint]. 2025 Sep 22. arXiv:2509.17995. doi:10.48550/arXiv.2509.17995.
10. Liu H, Tam D, Muqeeth M, Mohta J, Huang T, Bansal M, et al. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. Adv Neural Inf Process Syst. 2022;35:1950–1965.
11. Yong G, Jeon K, Gil D, Lee G. Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model. Comput Aided Civ Infrastruct Eng. 2023;38(11):1536–1554. doi:10.1111/mice.12954.
12. Zhang W, Deng Y, Liu B, Pan S, Bing L. Sentiment analysis in the era of large language models: a reality check. Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico. 2024. p. 3881–3906. doi:10.18653/v1/2024.findings-naacl.246.
13. Vamvakas D, Papaioannou I, Tsaknakis C, Sgouros T, Korkas C. Generative AI for sustainable smart environments: a review of energy systems, buildings, and user-centric decision-making. Energies. 2025;18(23):6163. doi:10.3390/en18236163.
14. Song Y, Wang T, Cai P, Mondal SK, Sahoo JP. A comprehensive survey of few-shot learning: evolution, applications, challenges, and opportunities. ACM Comput Surv. 2023;55(13s):1–40. doi:10.1145/3582688.
15. Kamath U, Keenan K, Somers G, Sorenson S. Prompt-based learning. In: Kamath U, Keenan K, Somers G, Sorenson S, editors. Large language models: a deep dive: bridging theory and practice. 1st ed. Cham (Switzerland): Springer Nature; 2024. p. 83–133. doi:10.1007/978-3-031-65647-7_3.

