Journal of Technological Information, Management & Engineering Sciences (2023),4(1), 14-26

Generalization Capabilities, Limitations, and Emerging Approaches to Zero-Shot and Few-Shot Learning in Foundation Models

[1] Dr. Farhan Ali
[2] Zainab Iqbal

How to cite this article:

Ali, F., & Iqbal, Z. (2023). Generalization Capabilities, Limitations, and Emerging Approaches to Zero-Shot and Few-Shot Learning in Foundation Models. *Journal of Technological Information, Management & Engineering Sciences, 4(1), 14-26.*

Received	Accepted	Published
8 January 2023	16 March 2023	28 May 2023

Abstract

By enabling broad task transfer through large-scale pre-training, prompting, in-context learning, and parameter-efficient adaptation, foundation models have revolutionized machine learning. In few-shot learning, a few demonstrations or examples are provided to help the adaptation, while in zero-shot learning the model needs to solve a task without any labelled example. This paper summarizes the state of the art in zero-shot and few-shot learning in language, vision, and multimodal foundation models, and discusses their generalization potential, their limitations, and novel methods. It reveals that foundation models can attain a high zero-shot and few-shot performance when the target tasks are semantically close to their pre-training distribution and when prompts, demonstrations, or instructions specify the target task well enough. It is important to note, however, that performance is still sensitive to prompt design, type and ordering of the demonstrations, label bias, domain shift, context length, and task recognition vs. real task learning. In-context learning is simple to compute as no parameters need to be updated, but it can be unstable and could be taking advantage of a surface-level correlation, not the intended task rule. There are complementary methods that can enhance data efficiency and robustness, such as parameter-efficient fine-tuning, retrieval-augmented prompting, instruction tuning, and adaptive prompting. It suggests standardized protocols for future evaluations of zero-shot and few-shot learning, which clearly distinguish between task recognition and task learning, include information on variability across prompt permutations, measure calibration and robustness, and explicitly test low-resource and cross-domain scenarios.

Keywords: *foundation models, zero-shot learning, few-shot learning, in-context learning, prompting, parameter-efficient fine-tuning, generalization, large language models, vision-language models, low-resource learning.*

Introduction

[1] Department of Computer Science, University of Malakand, Chakdara, Pakistan.
[2] Department of Computer Science, University of Swat, Swat, Pakistan.

Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

A foundation model is a large-scale model trained on a large and diverse dataset that can be adapted to many downstream tasks via prompting, fine-tuning or other light-weight adaptation techniques. They share some key traits: a large, diverse body of pre-training data, self-supervised learning objectives, and the ability to be used across tasks and modalities (Bommasani et al., 2021). In recent years, Large language models, vision-language models, and multimodal foundation models have shown that a single, pre-trained model can handle classification, generation, reasoning, retrieval and perception tasks, without requiring separate training for each application.

Zero-shot learning is the ability to do a task without having specific training examples for the task. In a vision-language model, it could be a matching of image embedding with textual class descriptions; in a language model, it could be answering a question with an instruction being provided as an input. Few-shot learning is the ability to adapt to a small number of labelled examples, typically either provided as demonstrations in a prompt or to provide a few-shot update to a limited number of parameters (Brown et al., 2020; Radford et al., 2021).

These paradigms are appealing because... They can lower the expenses of data collection and annotation, shorten deployment time on particular applications in targeted domains, and allow for quick adaptation to new tasks. But it is not a reliable generalization to know the answer when you have no training examples at hand or only a few. A model can provide plausible outputs but may be based on superficial patterns, memorized knowledge or biases from the pre-training data. Furthermore, the degree of performance may be quite different depending on the wording of the prompt, the order in which the demonstration is executed, the format of the task and the shift in its distribution.

This paper reviews the potential and challenges for few-shot and zero-shot learning in foundation models. It examines the key ways in which foundation models generalize, explores widespread failure patterns, and considers new methods and strategies aimed at making models more efficient, robust, and trustworthy. The paper provides a structured synthesis and an evaluation methodology for future research.

[1] Department of Computer Science, University of Malakand, Chakdara, Pakistan.
[2] Department of Computer Science, University of Swat, Swat, Pakistan.

Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Technological Information, Management & Engineering Sciences (2023),4(1), 14-26

Literature Review

Foundation Models and Transfer

Transferred representations are learned using large and diverse corpora via self-supervised objectives by foundation models. These representations may represent semantic, syntactic, visual, and cross-modal regularities that facilitate subsequent adaptation. Bommasani et al. (2021) define foundation models as models trained on large amounts of data, which are generalizable and can be fine-tuned for a variety of downstream tasks, while noting the generality and potential harms of these models.

The transferability of foundation models is a product of data size, model capability, design, and training goals. Large language models are trained from text corpora to learn linguistic and world knowledge, and vision-language models, like CLIP, learn a shared embedding space from image-text pairs. This shared representation allows for zero-shot classification by comparing visual features to textual descriptions of candidate category (Radford et al., 2021).

Lu et al. (2025) propose that the internal representation of a foundation model has two properties: representation potential – it can extract task-specific information in one modality, and align and unify information across different modalities. This view is helpful in understanding why a model that is pre-trained on a large distribution is sometimes also able to perform tasks in similar, but unseen, contexts.

Zero-Shot Learning

The key to zero-shot learning is the model's previous knowledge, and its ability to understand natural-language instructions or semantic descriptions. The Zero-Shot Prompting in language models usually involves a prompt along with a question without any examples. In vision-language models, the embeddings of an image are compared to embeddings derived from textual class prompts and the class with the highest similarity is chosen.

The best performance is seen when the target task is similar to the tasks that were seen during pre-training or instruction tuning. Model training on a set of natural-language templates of various tasks can be used to improve the

[1] Department of Computer Science, University of Malakand, Chakdara, Pakistan.
[2] Department of Computer Science, University of Swat, Swat, Pakistan.

Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

performance on a different set of zero-shot tasks. Wei et al. (2022) demonstrate that by adding instruction tuning, language models will follow instructions better and they will also gain in zero-shot and few-shot performance.

However, the ambiguity and the distribution shift are issues that the zero-shot learning approach is susceptible to. A prompt might not clearly indicate the purpose of the task, the target label space may not be known, or the input could have a domain-specific knowledge that is missing from the pre-training corpus. In such situations, the model can generate realistic but false results, especially in the case of a high degree of confidence when a prediction is not informed by the task specification.

In-Context Learning and Few-Shot Learning.

Few-shot learning is a technique that allows a model to learn from a few examples. This adaptation is commonly done in foundation models by in-context learning (ICL): demonstrations are inserted into the prompt, and the model predicts without changing its parameters (Brown et al., 2020). ICL lowers adaptation cost, and allows specification of a task at an inference time.

The effectiveness of ICL relies on the quality, relevance, diversity and organization of demonstrations. Schoch and Ji (2025) shows that the selection and ordering of demonstrations can have a significant impact on performance, and models might be vulnerable to label bias and prompt formatting. The more demonstrations the more the performance does not always increase, and performance can level off or even deteriorate if the demonstrations are too loud, repetitive, and unstructured.

There are various interpretations for the meaning of ICL. Bayesian explanations treat the model as solving tasks implicitly, and gradient-descent analogies propose that transformer attention can solve optimization tasks implicitly. However, Dong et al. (2024) point out that many of the theoretical studies are based on simplified tasks and small models, making them only applicable to larger scale real-world systems.

P**arameter-Efficient Adaptation**

In addition to in-context learning, another approach to achieving parameter efficiency is fine-tuning a small number of model parameters, known as

[1] Department of Computer Science, University of Malakand, Chakdara, Pakistan.
[2] Department of Computer Science, University of Swat, Swat, Pakistan.

Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

parameter-efficient fine-tuning (PEFT). The methods are adapters, prefix tuning, prompt tuning, and low rank adapters (LoRA). These techniques provide memory and training savings and task specialization.

In fine-grained few-shot classification, Adaptive Prompt Tuning (APT) (Brouwer et al., 2025) suggests an approach to fine-tune textual prompts for visual input through cross-attention. Unlike static prompt-tuning approaches, APT continuously improves text embeddings with visual cues to avoid the risk of getting the learned prompt too specific to a very limited support set.

It is important to note that PEFT outperforms ICL, as long as there are even a few target-task examples available, because it does not only rely on contextual inference but also updates its parameters. But it needs access to the model parameters and carries the dangers of overfitting, catastrophic forgetting and overgeneralization if the adaptation is too strong.

Multimodal Foundation Models

Multimodal foundation models have been developed to go beyond few-shot and zero-shot learning from text to connect visual, linguistic and sometimes auditory representations. Zero-shot classification benefits from image-text similarity methods and instruction-tuned multimodal models enable visual reasoning, captioning and question answering tasks, when given natural language instructions.

Few-shot adaptation of generative visual models also becomes a new area of interest in emerging research. By redefining visual tasks as input-output video transitions and training compact low-ranks adapters, Acuaviva et al. (2025) demonstrate the ability of video diffusion models to be adapted to new visual tasks with a few examples. The method indicates that generative foundation models can be made to have latent visual knowledge that can be extracted with minimal supervision.

But multimodal generalization still faces problems such as modality gaps, noisy or incomplete annotations and the difficulty of defining the visual concepts using the language. But the one example may not be enough to differentiate a general idea

[1] Department of Computer Science, University of Malakand, Chakdara, Pakistan.
[2] Department of Computer Science, University of Swat, Swat, Pakistan.
Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

from a specific category, so the model may be based on surface similarities or quick features.

Research Gap

While there has been rapid progress, four gaps prevent understanding and application of zero-shot and few-shot learning. First, there is no consistency in the separation of task recognition and task learning in the field. The reported few-shot gains could be illusory if there is a familiar task pattern that is recognized from a prompt but not actually learned from a few demonstrations.

Second, assessment systems are disorganized. The number of studies varies in prompt templates, the selection of demonstration, number of examples, model versions, and scoring methods, as well as train-test splits. This heterogeneity makes it hard to distinguish between differences in model capabilities, prompt engineering, dataset characteristics or evaluation methodology.

Third, robustness to distribution shift is poorly assessed. A large number of benchmarks are based on tasks which are semantically similar to the data used for pre-training or feature significant domain gaps which are easily separable. But in the real world, there are more nuanced shifts between domains, less-resourced languages, special terms, and noisy labels, which can lead to very poor zero-shot and few-shot results.

Fourth, the topics of efficiency, safety, and uncertainty are under-explored. In-context learning can be highly time-consuming due to the context length, and inference cost, and PEFT might need parameter access and careful regularization. Furthermore, foundation models have a tendency to hallucinate, to come up with unstable outputs, and to show inappropriate confidence, especially in high-stakes situations.

Objectives

This study aims at achieving the following objectives:

Review the mechanisms by which foundation models can enable generalization for zero-and few shots.

To compare prompting, in-context learning, instruction tuning, retrieval augmentation and parameter-efficient fine-tuning as adaptation approaches.

[1] Department of Computer Science, University of Malakand, Ꮯhakdara, Pakistan.
[2] Department of Computer Science, University of Swat, Swat, Pakistan.

Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

3. Understand the main problems of zero-shot learning and few-shot learning: prompt sensitivity, label bias, domain shift and context constraints.

4. To suggest a future evaluation paradigm for evaluating generalization, robustness, calibration, and efficiency in the context of foundation-model research.

Methodology

Review Design

This paper presents a structured narrative review of recent research on zero-shot learning, few-shot learning, in-context learning and foundation-model adaptation. The focus of the review is on influential methods, empirical studies from 2020 to 2025, and peer-reviewed surveys. The sources were chosen for methodological transparency, relevance for generalization, and coverage of language, vision and multimodal setting.

The review categorizes the methods into five classes: zero-shot prompting, few-shot in-context learning, instruction tuning, retrieval-augmented prompting and parameter-efficient fine-tuning. The assessment of each category is based on data needs, adaptation cost, generalization, strengths and limitations.

Comparative Criteria

Zero-shot is evaluated by accuracy, exact-match score, F1-score, or task-specific generation metrics, without task-specific labelled examples. The few-shot ability is measured by a small number of examples per class/task, typically one to sixteen. The number of demonstrations should be used as a measure of performance as gains from more examples are not always monotonic.

Base-to-new class evaluation, cross domain transfer, and low resource task evaluation are used to assess generalization. Base-to-new evaluation: A model is adapted to base classes and is evaluated on both base and unseen classes; the harmonic mean of the base and new-class performance will give a balanced evaluation of specialization and generalization.

Evaluation Protocol

Future research should investigate the models in four settings: zero-shot prompting, few-shot prompting, parameter-efficient fine-tuning, and full fine-

[1] Department of Computer Science, University of Malakand, Chakdara, Pakistan.
[2] Department of Computer Science, University of Swat, Swat, Pakistan.
Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Technological Information, Management & Engineering Sciences (2023),4(1), 14-26

tuning when possible. They should state the average and standard deviation for various prompt templates, demonstration orders and random seeds. This is because the performance of the ICL may be very sensitive to the choice of demonstration (and ordering).

In the protocol, the separation of task recognition from task learning is recommended. The following can be used to test task recognition: demonstrations with shuffled or wrong labels can be used if the model continues to perform well, despite this, it might be using prior knowledge about the tasks it was able to generalize rather than learning the demonstrated mapping. The key to task learning is to measure the improvement in the ability to perform a task when the model is presented with valid models on a task that has never been performed before.

The protocol also suggests to report calibration metrics (such as expected calibration error) in addition to accuracy. An overconfident model that is accurate is not the optimal model to deploy in high-stakes situations, since it fails to reliably communicate uncertainty. The computational costs include prompt length, latency, memory requirement, and number of trainable parameters.

Integrated Adaptation Framework

To address these requirements, the paper introduces an integrated adaptation framework that chooses the appropriate adaptation approach based on the availability of the data, risk of the task and deployment constraints. If no labelled examples are available, then using zero-shot prompting with clear instructions and limited output formats is suitable. If there is a small labelled support set, retrieval-augmented ICL or lightweight PEFT can be used to enhance task specification.

The framework suggests few-shot adaptation alongside uncertainty estimation and OOD detection as well as human review for high-risk applications. When a model is not very confident, if it detects input data that is out of distribution, or if it does not validate output data, it should abstain or escalate. This design does not consider zero-shot and few-shot learning as independent algorithms, but rather parts of a reliable AI system.

[1] Department of Computer Science, University of Malakand, Chakdara, Pakistan.
[2] Department of Computer Science, University of Swat, Swat, Pakistan.
Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Journal of Technological Information, Management & Engineering Sciences (2023),4(1), 14-26

Results

The reviewed evidence suggests that foundation models have good zero-shot and few-shot performance, though this is highly dependent on the context. The overall performance is best when the task is semantically similar to the pre-training data and when the task instructions explicitly state the desired output. This alignment is enhanced by instruction tuning to introduce models to a variety of tasks represented as natural language templates.

Few-shot ICL tends to perform better than zero-shot prompting, although this is not always the case. There are conflicting reports on the benefits of increasing the number of demonstrations: some studies show an improvement in the number of demonstrations, and others show no improvement or even degradation.Additional demonstrations may have positive, negative, or neutral impacts on performance, with some studies reporting gains with demonstrations and others reporting stagnation or performance losses in additional demonstrations. The evidence that has been reviewed indicates that the quality, relevance, diversity, and ordering of demonstration examples can be as important as the number of examples.

When labelled examples are present, parameter-efficient fine-tuning can offer more consistent task adaptation than ICL. They claim their cross-attention-based APT method achieved 47% accuracy on FGVC-Aircraft, which includes 16 shots, whereas the zero-shot baseline accuracy was 17%. The improvement shows the benefit of using prompts that are not just text-based, but adapted to the visual evidence as well.

However, cross-domain few-shot learning is still a difficult task. Zhou et al. (2025) point out that current fine-tuning methods may fail when there is a significant domain shift between the pre-training dataset and the task of interest, particularly when the target samples are labeled few. An expert model is trained on the target data and is then used to regularize foundation-model adaptation to solve this problem, as proposed by their absorption adaptation framework.

The results also indicate that the zero-shot multimodal inference is not generally accurate. General Purpose Vision-Language Models did not achieve reliable

[1] Department of Computer Science, University of Malakand, Chakdara, Pakistan.
[2] Department of Computer Science, University of Swat, Swat, Pakistan.

Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

detection on their own at an APCER of 1% under strict operating conditions in a security oriented identity-document presentation-attack detection benchmark, with all the tested methods achieving a BPCER of more than 70%. This is an example of how pre-training across a wide range of tasks does not necessarily lead to reliable performance in more niche, high-risk situations.

Discussion

The primary conclusion is that zero-shot and few-shot learning are strong but limited types of generalization. To use ZSL, the semantic description of the task needs to be aligned with the pre-trained knowledge of the model. Few-shot learning introduces task-specific evidence, but relying on it hinges on whether the demonstrations are conveying the desired rule in a clear manner and whether the model can effectively process the evidence while avoiding distractions from unrelated patterns.

The difference between task recognition and task learning is significant. A model can look like learning from a demonstration, but is really identifying a known task structure and using a pre-trained solution. This behavior can lead to high benchmark accuracy but low robustness in the case of a truly new mapping, a novel label space, and/or a different reasoning procedure.

One of the major practical limitations is prompt sensitivity. Variations in how the instructions are worded, how they are demonstrated, how the labels are distributed or how the work is output can result in very different performances. This sensitivity is a barrier for the validation and deployment of zero-shot and few-shot systems in regulated or safety-critical context. Only the top one prompt being reported, however, may overestimate reliability in the real world.

New methods overcome some of these drawbacks in complementary ways. Instruction tuning enables task alignment, retrieval augmentation provides relevant task-specific context, adaptive prompting adds task-specific information to the input, and PEFT allows for stable specialization of a small set of task-specific examples. The best choice depends on the consequences of errors, computational resources and labelled data.

[1] Department of Computer Science, University of Malakand, Chakdara, Pakistan.
[2] Department of Computer Science, University of Swat, Swat, Pakistan.
Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Future systems ought to have the few-shot reliability and the zero-shot flexibility. In a real pipeline, the zero-shot inference may be the first step, retrieval or demonstrations would be used if there is a high level of uncertainty, PEFT could be introduced once a small validation set becomes available, and OOD detection and human monitoring could be kept throughout the process. A layered design offers a combination of adaptability, cost and trust.

There are a number of restrictions to note: This is an evaluation and not an empirical comparison, so reported results should not be compared directly between studies due to differences in data sets, prompts, models and evaluation procedures. Furthermore, many of the theoretical developments with ICL involve simplifications, and its correspondence with the larger-scale real world foundation models is unclear.

Conclusion

The ability to transfer large numbers of tasks without needing to train a model from scratch for each is a key promise of foundation models, and is termed zero and few-shot learning. Unlike few-shot learning, which requires only a few demonstrations and/or examples in order to guide adaptation, zero-shot learning uses pre-learned knowledge and natural-language task descriptions. These capabilities can be achieved with complementary mechanisms such as in-context learning, instruction tuning, retrieval augmentation, and parameter-efficient fine-tuning.

But good benchmarks performance is no guarantee of reliable generalization. Zero-shot & few-shot systems are still prompt sensitive, demo quality and ordering, label bias, domain shift, context length, and task ambiguity. Specifically, the models may be able to identify common task structures instead of really learning new input-output relationships from small amounts of data.

Future research should be conducted using consistent task evaluation protocols to isolate task recognition from task learning, measure task recognition variability across the presentation of prompts and demonstration order, evaluate calibration and robustness, and evaluate low-resource and cross domain settings. The most promising approach is towards an integrated adaptation framework that merges

[1] Department of Computer Science, University of Malakand, Chakdara, Pakistan.
[2] Department of Computer Science, University of Swat, Swat, Pakistan.
Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

the advantages of zero-shot flexibility, retrieval-augmented few-shot learning, parameter-efficient specialization, uncertainty estimation and human oversight. This type of systems can evolve from impressive one-shots and begin to deploy reliable and trusted foundation-models.

References

Acuaviva, P., Davtyan, A., et al. (2025). Emergent few-shot learning in video diffusion models. arXiv. https://doi.org/10.48550/arXiv.2506.07280

Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). Opportunities and risks of foundation models. arXiv. https://doi.org/10.48550/arXiv.2108.07258

Brouwer, E., van Woerden, J. E., Burghouts, G., Valdenegro-Toro, M. and Zullich, M. (2025). Cross attention in fine-grained few-shot learning. arXiv. https://doi.org/10.48550/arXiv.2412.14640

Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners! Advances in Neural Information Processing systems 33, 1877-1901.

Dong, Q., Li, L., Dai, D., et al. (2024). An online survey regarding In-Context Learning. In proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP). 1107–1128). Association for Computational Linguistics.

Lu, J., Wang, H., Xu, Y., Wang, Y., Yang, K., & Fu, Y. (2025). A survey of representation potentials of foundation models for multimodal alignment. In the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP). 16676–16681). Association for Computational Linguistics.

Radford, A., Kim, J. W., Hallacy, C., et al. (2021). Transferring visual models from natural language. In Proceedings of the 38th International Conference on Machine Learning, volume 2.In Proceedings of the 38th International Conference on Machine Learning, Volume 2. 8748–8763). PMLR.

Schoch, S., & Ji, Y. (2025). Good, bad, or debatable? A survey of the effects of data in in-context learning. On the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP), which was published at the same time.Published concurrently in Proceedings of the 2025 Conference on Empirical Methods in

[1] Department of Computer Science, University of Malakand, Chakdara, Pakistan.
[2] Department of Computer Science, University of Swat, Swat, Pakistan.

Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.

Natural Language Processing (EMNLP). 29797–29812). Association for Computational Linguistics.

Tapia, J. E., & Busch, C. (2025). Will the foundation models be capable of generalizing the ability to present attack detection in ID cards? arXiv. https://doi.org/10.48550/arXiv.2506.05263

Wei, J., Bosma, M., Zhao, V. Y., et al. (2022). Zero-shot learning are finetuned language models. In Proceedings of ICLR 2013.

Zhou, F., Wang, P., Zhang, L., Wei, W., Ding, C., Lin, G., & Zhang, Y. (2025). Towards effective Foundation Model Adaptation for Extreme Cross-Domain Few-Shot Learning. In IEEE/CVF International Conference on Computer Vision, 2006. 4582–4593). IEEE.

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Practicing and understanding how to prompt for vision-language models. International Journal of Computer Vision, 130(9), 2337–2348. https://doi.org/10.1007/s11263-022-01653-1

[1] Department of Computer Science, University of Malakand, Chakdara, Pakistan.
[2] Department of Computer Science, University of Swat, Swat, Pakistan.
Copyright (c) 2023 Author(s)
This work is licensed under a Creative Commons Attribution 4.0 International License.