GLOBUS PUBLICATION

Globus Journal of Progressive Education
A Refereed Research Journal
Vol 16 / No 2 / Jul-Dec 2026 P - ISSN: 2231-1335, E- ISSN: 2582-7200

FEW-SHOT AND ZERO-SHOT LEARNING: TEACHING MODELS TO LEARN FROM VERY FEW EXAMPLES

[*]AbdulBasir Momand
[**]AbdulGhafar Omerkhil
[***]Osama Zain
[****]Asmatullah Ahmadzai

Paper Received: 17.05.2026 / **Paper Revised**: 18.07.2026 / **Paper Accepted**: 09.08.2026 / **Paper Published**: 17.08.2026

Corresponding Author: AbdulBasir Momand; doi: 10.46360/globus.edu.220262004

Abstract

Despite the remarkable progress achieved by deep learning over the past decade, the majority of high-performing models continue to depend heavily on large quantities of high-quality labeled data. This reliance poses a substantial barrier in many real-world applications where obtaining labeled examples is costly, time-consuming, or practically infeasible. Domains such as medical imaging, specialized industrial inspection, rare disease diagnosis, and low-resource languages often lack sufficient annotated datasets to train conventional supervised models effectively. Few-shot learning focuses on scenarios where only a handful of labeled examples - sometimes only one - are available for each target class. Zero-shot learning, by contrast, addresses an even more challenging setting in which models must recognize categories for which no labeled examples are provided during training. Instead of direct supervision, ZSL systems rely on auxiliary semantic information - such as natural language descriptions, attribute vectors, or large language models - to bridge the gap between seen and unseen classes. This paper provides a comprehensive comparative study of modern few-shot and zero-shot learning techniques, focusing on their conceptual foundations, methodological distinctions, and practical applications.

Keywords: Few-Shot Learning, Zero-Shot Learning, Meta-Learning, Prototypical Networks, Prompt-Based Learning, Transfer Learning, Low-Data Regime, Deep Learning.

Introduction

Deep learning has achieved remarkable performance across numerous domains, including computer vision, natural language processing, speech recognition, robotics, biomedical analysis, and security. Much of this success, however, is closely tied to the availability of large-scale labeled datasets, which provide the supervision necessary for training high-capacity models. Systems such as convolutional neural networks, large transformer architectures, and multimodal foundation models often require tens of thousands to millions of annotated examples to reach generalizable performance. When substantial labeled data exist, deep learning systems can learn robust representations and outperform traditional machine learning techniques by a significant margin. Yet in many real-world settings - especially those involving rare phenomena, specialized expert knowledge, or rapidly evolving environments - the assumption of abundant labeled training data breaks down. Data scarcity remains one of the most critical barriers to deploying deep learning models widely, particularly in high-stakes or resource-constrained domains.

Such limitations are evident in fields like rare disease diagnosis, where patient data are inherently limited and labeling requires costly clinical expertise, often involving time-consuming manual review by specialists. Similarly, efforts to build natural language processing systems for low-resource languages face challenges due to the lack of annotated corpora, limited digitized text, and diverse linguistic structures that differ significantly from well-studied languages like English or Mandarin. In security and threat detection, the difficulty is compounded by the constantly evolving nature of adversarial behaviors, emerging malware variants, and unseen attack patterns. Collecting labeled examples of every possible threat is unrealistic, and delay in obtaining labels can render the training data obsolete as malicious actors develop new strategies. These scenarios illustrate the growing need for techniques that reduce dependence on extensive labeled datasets and instead leverage alternative forms of supervise.

*-***Rana University, Afghanistan.*

To address these challenges, the machine learning community has developed paradigms specifically designed to operate under extreme data scarcity. Two of the most influential are few-shot learning (FSL) and zero-shot learning (ZSL). Few-shot learning seeks to enable models to generalize from only a handful of labeled examples for each class - often as few as one or five. Instead of relying on large datasets, FSL models learn transferable priors or task-learning strategies that can be quickly adapted to novel tasks during evaluation. The fundamental assumption underlying few-shot learning is that meta-knowledge, extracted from a broad distribution of related tasks during training, can be reused to perform well on unseen tasks with minimal supervision. This perspective has led to a variety of learning approaches, ranging from metric-learning techniques that compare queries to learned prototypes, to optimization-based frameworks that fine-tune model parameters efficiently.

Zero-shot learning, conversely, eliminates the need for labeled examples entirely for unseen classes. In ZSL, models leverage auxiliary knowledge sources - such as natural language descriptions, semantic attribute vectors, ontological relationships, or large pre-trained language models - to build a connection between known and unknown categories. Zero-shot learning is particularly powerful in scenarios where class labels describe novel entities for which data may be scarce or entirely absent. Classical approaches used attribute-based representations, while modern methods heavily rely on multimodal pretraining that embeds text and images into shared semantic spaces. With the rise of giant foundation models trained on vast internet-scale corpora, ZSL has become increasingly viable, making it possible to perform tasks such as object classification, text classification, image retrieval, or even reasoning tasks without requiring explicit labeled examples for every new category.

The present paper offers a comprehensive review and comparative analysis of these two learning paradigms, examining their theoretical foundations, architectural choices, pretraining strategies, evaluation benchmarks, and real-world applications. We begin by surveying the fundamental concepts and motivations behind few-shot and zero-shot learning, discussing how these paradigms shift the traditional supervised learning framework toward more flexible and data-efficient inference. To provide a structured exploration, we categorize few-shot learning methods into metric-based approaches, optimization-based meta-learning, and data-augmentation or hallucination techniques. Metric-based methods, such as Siamese networks, Matching Networks, and Prototypical Networks, operate by embedding examples into a representation space where proximity reflects class similarity. These models are attractive due to their simplicity, stability, and interpretability, making them widely used in modern applications.

Similarly, we investigate zero-shot learning strategies, tracing their evolution from early attribute-based and semantic embedding models to more advanced transformer-based and prompt-driven frameworks. Early ZSL methods relied on human-defined attribute vectors or word embeddings to relate unseen classes to seen ones. However, modern approaches like CLIP, ALIGN, and PaLI - massive multimodal transformer models trained jointly on image–text pairs - have transformed zero-shot learning by aligning visual and linguistic representations into a unified embedding space. These models are capable of performing a wide variety of recognition and reasoning tasks with natural-language descriptions functioning as class labels or task classifiers. Recent developments in prompting, where textual instructions or class descriptions guide the model toward desired behavior, further enhance ZSL capabilities by enabling context-aware, instruction-following classification without requiring explicit retraining.

Beyond architectural considerations, pretraining strategies play a crucial role in determining the effectiveness of low-shot and zero-shot models. For few-shot learning, successful pretraining often involves meta-learning on broad task distributions or leveraging self-supervised learning to produce powerful general-purpose feature extractors. For example, large vision transformers trained using contrastive or masked image modeling techniques have been shown to significantly boost few-shot performance when combined with metric-based classifiers. Zero-shot learning heavily benefits from scale: the vast amount of data used to train multimodal or language-only foundation models enables them to internalize rich semantic relationships that generalize far beyond their original training objective. Understanding how pretraining influences downstream low-data, performance is essential both for developing efficient models and for deploying them in real-world environments.

To contextualize these methods empirically, we review widely used evaluation benchmarks for few-shot and zero-shot learning, including miniImageNet, tieredImageNet, Omniglot, CIFAR-FS, Fungi, SUN, AWA, and more recent benchmarks designed for large foundation models. We also discuss metrics for evaluation, challenges in fair comparison across studies, and concerns related to dataset bias, domain shift, and overly narrow benchmark designs. While many classical benchmarks focus on image classification, recent work has emphasized more diverse evaluation tasks - including natural language understanding,

medical diagnosis, anomaly detection, and multimodal reasoning - to better capture the real-world utility of data-efficient learning.

Practical applications form another major focus of the paper, illustrating how few-shot and zero-shot learning can be used to address pressing challenges in various domains. In medicine, few-shot learning is employed to diagnose rare conditions, classify unusual pathology findings, and adapt models to new imaging devices or protocols without requiring extensive annotated data. Zero-shot techniques allow models to interpret medical images by connecting natural-language prompts to visual features, enabling flexible diagnostic reasoning in settings where labeled data are limited or unavailable. In natural language processing, both FSL and ZSL support tasks such as intent detection in low-resource languages, domain adaptation for specialized text, and classification tasks where annotated corpora are scarce. Zero-shot and prompt-based models are particularly valuable for handling emergent categories or tasks specified only by textual instructions.

In addition to presenting existing approaches, the paper highlights key challenges and open research questions that remain unaddressed. Few-shot learning models often struggle when the distribution of new tasks differs significantly from the distribution encountered during training, a phenomenon known as task shift or domain shift. Zero-shot models, while powerful, may rely heavily on biases present in large-scale pretraining data, leading to unreliable predictions in specialized domains or when semantic descriptions are ambiguous. Both paradigms must also contend with issues of uncertainty estimation, interpretability, robustness to adversarial examples, and ethical considerations arising from model biases or misuse. Furthermore, despite rapid progress, integrating these paradigms into production systems remains nontrivial, requiring careful evaluation, model calibration, and sometimes hybrid approaches that combine FSL, ZSL, and active learning.

This paper aims to synthesize current knowledge on few-shot and zero-shot learning, offering a structured and comparative perspective on architectures, training strategies, benchmarks, and real-world applications. By highlighting both the strengths and limitations of existing approaches, we provide insights into how future work might develop more robust, generalizable, and context-aware systems capable of reasoning effectively under extreme data scarcity. As deep learning continues to evolve toward more adaptive and resource-efficient paradigms, understanding these data-scarce learning frameworks will be essential for building AI systems that are equitable., scalable, and broadly applicable.

Related Work

Problem Setting

Few-shot learning and zero-shot learning have emerged as two closely related yet distinct approaches for addressing one of the most persistent challenges in modern machine learning: the need for vast quantities of labeled data. Traditional deep learning models thrive on abundance. When millions of labeled images, sentences, or audio clips are available, modern neural architectures - especially large transformers and deep convolutional networks - can learn incredibly rich representations and achieve human-level or even superhuman performance. However, this success hides a difficult reality: in many real-world scenarios, collecting such large labeled datasets is either extremely expensive, painfully slow, or simply impossible.

This is where few-shot learning (FSL) and zero-shot learning (ZSL) become vital. At a high level, few-shot learning attempts to teach a model how to generalize from a very small number of labeled examples, typically just one to five per class. Instead of learning a task from scratch, the model is trained in a special way - often using a meta-learning setup - that helps it learn how to learn new tasks quickly. The guiding intuition is similar to how humans operate: if you show a person one or two examples of a new type of plant or tool, they can usually recognize it in the future, even if they've never encountered it before. The human brain is incredibly efficient at leveraging prior knowledge and transferring it to new situations. Few-shot learning attempts to give machine learning systems the same kind of adaptive capability.

Zero-shot learning, on the other hand, pushes this idea even further. Whereas few-shot learning still relies on a handful of labeled examples, zero-shot learning removes labeled examples entirely for the new classes. Instead, models rely on auxiliary semantic information - natural-language descriptions, human-defined attributes, large language models, ontologies, or text–image associations - to build an understanding of categories they have never encountered directly. For example, a zero-shot vision model may have never seen an image labeled "giant panda," but it may still recognize a panda by connecting textual descriptions ("a black and white bear found in central China") to patterns it has learned about bears, colors, shapes, and animals more broadly. This ability mirrors how humans' reason about unfamiliar concepts: even if you have never seen a fossa (a rare animal native to Madagascar), a verbal description can help you picture what it might look like.

Few-Shot Learning: Learning from Scarcity

Few-shot learning has grown largely as a response to the mismatch between the data-rich regimes required by today's deep networks and the data-poor conditions encountered in many practical fields. Consider a medical scenario where a researcher only has three labeled examples of a rare disease. Classical supervised learning collapses under such scarcity: the model overfits immediately, memorizing the examples rather than learning anything generalizable. Few-shot learning attempts to overcome this by embedding the learning process itself into the training pipeline.

Most few-shot learning methods fall into one of two broad categories: metric-based approaches and optimization-based approaches.

Metric-based approaches build an embedding space where visually or semantically similar items are nearby. During classification, a model simply compares a test example with the few available labeled examples - often referred to as support examples - to find which class prototype it is closest to. Prototypical Networks, Matching Networks, and Siamese Networks exemplify this strategy. The elegance of metric-learning lies in its simplicity: once the embedding space is well-learned, classification becomes a matter of similarity comparison.

Optimization-based approaches tackle the challenge differently. Instead of building an embedding space, they attempt to find model parameters that are easily fine-tuned when only a few examples are available. The classic example is Model-Agnostic Meta-Learning (MAML), which learns model parameters that are not necessarily optimal for one specific task but are close to optimal for many possible tasks. With this "meta-initialization," a few gradient updates on new task data lead to effective learning. This mimics the way humans adapt quickly to new contexts by leveraging prior experience.

Few-shot learning is especially powerful in applications such as:

- Rare disease classification, where only a handful of samples exist
- Wildlife conservation, where identifying endangered species requires limited observations
- Robotics, where robots must adapt to new tasks without extensive retraining
- Low-resource speech and language processing, where labeled data in certain languages or dialects may be scarce

In all of these cases, few-shot learning provides a path toward making AI systems more flexible, adaptable, and efficient.

Zero-Shot Learning: Understanding the Unseen

Zero-shot learning operates on an even more ambitious principle: instead of learning from a small number of examples, it learns from no examples at all. This is not magic - it relies on the model's ability to connect textual or semantic descriptions with learned patterns in data. Early zero-shot learning systems used manually defined attribute vectors (for example, describing animals with traits like "has stripes," "is large," "is carnivorous"). More recent systems rely on deep textual embeddings, especially those derived from large language models such as GPT-based architectures, BERT variants, or multimodal models like CLIP.

One of the key innovations driving modern zero-shot learning is the alignment of different modalities - particularly text and images - into a shared embedding space. Models like CLIP are trained on hundreds of millions of image–caption pairs, enabling them to map both text and images to comparable vectors. This means that, at inference time, classification can be performed simply by comparing the embedding of an image with the embedding of a text description such as "a photo of a fire truck" or "a type of flower with pink petals." If the model has never seen that specific category but understands related concepts through descriptions, it can still perform the classification.

This capability is especially valuable in domains where categories evolve quickly or where collecting labels is practically impossible. For example, in cybersecurity, new malware variants appear daily, and manually labeling each one would be infeasible. Zero-shot techniques allow a model to generalize to new categories by relying on textual descriptions of behavior, making early detection more viable. In natural language, processing, zero-shot classification enables models to perform sentiment analysis, topic classification, or intent detection without any labeled training data for the specific categories. The user simply provides text descriptions of the desired classes, and the model uses its underlying knowledge to perform the classification.

Even more exciting is the emergence of instruction-based zero-shot learning, where models follow natural-language instructions to perform tasks they were never explicitly trained for. Large language models have demonstrated that, with sufficient pretraining, models can internalize enough general knowledge to interpret new tasks through natural-language prompts alone. This trend is driving a shift toward more universal and flexible AI systems that do not rely on task-specific fine-tuning.

Why These Paradigms Matter

Few-shot and zero-shot learning are not just technical curiosities - they directly address one of the biggest bottlenecks in deploying AI responsibly and

equitably. Data scarcity often occurs in areas where AI has the most potential for social impact: healthcare, scientific research, rare language preservation, humanitarian response, and environmental monitoring. When AI systems depend on massive datasets, only institutions with substantial resources can build and deploy effective models. Few-shot and zero-shot learning democratize the process by reducing the dependence on data scale and instead leveraging generalization, transfer, and prior knowledge.

Furthermore, these approaches promote greater robustness and flexibility. A model that can adapt quickly to new tasks with minimal data is also more resilient to distribution shift, making it better suited for dynamic environments. In cases where classes or behaviors evolve - such as emerging diseases, new types of fraud, or unexpected security threats - zero-shot capabilities become essential.

From a research standpoint, these paradigms push machine learning toward a more human-like understanding of the world. Humans are remarkably data-efficient learners; we generalize from sparse examples and interpret new concepts through language. Few-shot and zero-shot learning methodologies attempt to capture this efficiency and adaptiveness in artificial systems.

A Convergence of Approaches

Despite their differences, few-shot and zero-shot learning are increasingly connected. Many modern models combine elements of both. For example, prompt-based models can operate in both few-shot and zero-shot regimes depending on whether a handful of examples (in-context learning) or purely instructions are provided. Vision–language models can classify unseen categories using descriptions but can also be fine-tuned with a few images to improve accuracy. This blending suggests that the boundary between the two paradigms may continue to blur as models become more capable, more multimodal, and more aligned with natural-language reasoning.

Metric-based and Episodic Methods

Metric learning, combined with episodic training, has become one of the most influential and intuitive approaches for tackling few-shot classification. The central idea behind metric learning is straightforward: instead of training a model to directly predict class labels, the model is trained to embed images (or other data types) into a representation space where similar examples cluster together and dissimilar examples lie far apart. When paired with episodic training-a training setup designed to mimic the structure of few-shot tasks encountered during testing-this strategy enables models to generalize remarkably well from only a handful of labeled examples. The elegance of this framework lies in its simplicity: if the model can learn a meaningful embedding space during training, classification under few-shot conditions becomes a matter of comparing distances between embeddings, rather than learning an entirely new classifier for each task.

The idea of episodic training was popularized by Matching Networks (Vinyals et al., 2016), one of the earliest and most influential works in modern few-shot learning. Matching Networks introduced the notion that the training procedure itself should mimic the test-time few-shot scenario. Instead of training on full datasets in a standard supervised manner, models are trained in "episodes," where each episode contains a small support set (with only a few labeled examples per class) and a query set. This episodic format encourages the model to practice the exact type of rapid generalization it will need during few-shot evaluation. Matching Networks also introduced an attention-based nearest-neighbor classifier, where the model predicts the label of a query example by comparing it to all support examples using learned attention weights. This mechanism allows the model to focus on the most relevant support examples and adapt dynamically to the task at hand. Matching Networks demonstrated for the first time that a combination of episodic training and distance-based classification could yield strong performance on challenging benchmarks like Omniglot and miniImageNet, setting the stage for a wave of subsequent research.
Building on the success of Matching Networks, Prototypical Networks (Snell et al., 2017) refined the metric-learning approach by introducing a more streamlined, conceptually clear framework. Instead of comparing queries to individual support examples, Prototypical Networks compute a prototype for each class-a single vector representing the mean embedding of all support examples for that class. Intuitively, this prototype acts as a central "reference point" for the class in the embedding space. Classification is then performed by computing the distance between a query embedding and each class prototype, typically using Euclidean distance. Despite the simplicity of this approach, Prototypical Networks achieved impressive results, demonstrating that well-structured embedding spaces allow for surprisingly effective few-shot classification. Moreover, the use of prototypes offers several practical benefits: improved robustness to noisy support examples, reduced computational cost compared to comparing against many individual embeddings, and increased interpretability of class representations. The method's conceptual clarity and strong empirical performance helped establish Prototypical Networks as a widely used baseline in few-shot learning research.

While Prototypical Networks rely on a fixed, distance-based comparison, Relation Networks

(Sung et al., 2018) introduced a more flexible, learned similarity function. Instead of assuming that simple metrics like Euclidean distance or cosine similarity are adequate for measuring relationships between support and query examples, Relation Networks use a neural network module to learn a non-linear similarity score. This allows the model to capture more complex relationships between examples that might not be well-represented in a purely geometric space. For instance, in fine-grained classification tasks where two classes may appear visually similar but differ in subtle ways, a learned relation module can highlight critical distinctions. Relation Networks also use episodic training to ensure the relation module encounters diverse tasks during training, allowing it to generalize effectively during evaluation. By learning a deeper and more adaptive notion of similarity, Relation Networks extend the metric-learning paradigm toward richer representations and more flexible decision boundaries.

Together, Matching Networks, Prototypical Networks, and Relation Networks illustrate the evolution of metric-based few-shot learning methods-from attention-driven nearest-neighbor models to prototype-based embedding structures and finally to fully learned similarity functions. Despite their differences, these methods share several unifying principles. First, they all rely on the assumption that a well-learned embedding space can support rapid generalization with very limited supervision. Second, episodic training acts as a powerful regularizer by forcing models to repeatedly adapt to new, small-scale classification tasks. Third, the classification process in all three approaches is conceptually simple: instead of relying on heavy fine-tuning or retraining, they classify examples by comparing embeddings. This makes metric-learning approaches both computationally efficient and highly practical in real-world low-data environments.

Beyond their historical significance, these methods form the backbone of many modern systems. Their concepts appear in more sophisticated architectures such as meta-learning algorithms, contrastive-learning–based few-shot models, and hybrid approaches that integrate metric learning with large pretrained models or transformer architectures. Even contemporary vision–language systems, like CLIP, rely on deeply learned embedding spaces and similarity comparisons, echoing the design principles developed in early few-shot research.

In summary, metric learning combined with episodic training provides a powerful and intuitive foundation for few-shot classification. Matching Networks demonstrated how episodic training and attention-based nearest-neighbor reasoning could enable rapid adaptation. Prototypical Networks showed that simple, elegant prototype representations can yield strong performance with minimal computational overhead. Relation Networks pushed the boundary further by learning flexible similarity functions capable of modeling complex relationships. Together, these approaches helped shape the trajectory of few-shot learning research and continue to influence the design of data-efficient machine learning systems today.

Optimization-based Meta-Learning

Optimization-based meta-learners represent another major direction in few-shot learning, grounded in the idea that models should not only learn from data but also learn how to learn. Instead of focusing on distances or similarity measures, as metric-learning approaches do, optimization-based methods aim to discover model parameters-or even entire update rules-that can be rapidly adapted to new tasks with minimal training data. The underlying intuition mirrors human learning: when we encounter a new concept, we rarely start from scratch. Rather, we rely on a lifetime of accumulated prior knowledge that helps us interpret and adapt to unfamiliar situations quickly. Optimization-based meta-learners try to encode this "prior experience" into the model itself. A landmark contribution in this space is Model-Agnostic Meta-Learning (MAML), introduced by Finn et al. (2017). MAML proposed a deceptively simple yet powerful idea: instead of training a model to perform well on a single task, train it to be *easy to fine-tune*. More concretely, MAML learns an initialization of the model's parameters such that only a few gradient updates on a small support set will lead to high performance on a new task. This means the model begins each task already close to a good solution, allowing it to adapt quickly and efficiently. The elegance of MAML lies in its generality-it makes no assumptions about the model architecture, the optimizer, or the task domain. As a result, it can be applied to classification, regression, reinforcement learning, and beyond.

The success of MAML triggered a wave of research exploring how to make meta-learning more stable, more efficient, and more expressive. Some variants, such as Reptile and First-Order MAML, simplify the optimization procedure by removing second-order derivatives, making training easier and more scalable. Other approaches, like Meta-SGD and LEO (Latent Embedding Optimization), go further by learning not only the initialization but also the update rules themselves, allowing the model to discover how quickly different parameters should adapt. These innovations highlight an important shift: rather than manually designing learning algorithms, researchers can use meta-learning to learn the learning process automatically.

Optimization-based methods are particularly appealing in domains where rapid adaptation is

critical. For example, in robotic control, a robot may need to adjust to new objects or environments without lengthy retraining. In medical AI, where patient data can vary widely across hospitals or devices, the ability to quickly fine-tune to new conditions is highly valuable. Even in natural language processing, meta-learning can help models generalize across different writing styles, topics, or languages.

Despite their strengths, optimization-based meta-learners do face challenges. They are often computationally intensive due to their bi-level optimization structure, and their performance can degrade when the distribution of new tasks differs significantly from those seen during meta-training. Nonetheless, the core idea introduced by MAML-that learning how to adapt is as important as learning to classify-continues to influence a wide range of research in few-shot learning and fast adaptation frameworks.

Transfer Learning and Large-Scale Pretraining

Pretrained visual and language backbones have become one of the most reliable ways to enhance few-shot learning performance, fundamentally changing how researchers think about low-data regimes. Rather than training models from scratch-an approach that typically requires large datasets and extensive compute-modern systems begin with representations learned from massive, diverse pretraining corpora. These pretrained backbones, whether they come from large-scale vision datasets like ImageNet or from multimodal sources such as image–text pairs, provide rich, general-purpose features that transfer remarkably well to new tasks. The intuition is simple: if a model already understands edges, shapes, textures, objects, or linguistic structure, it needs far less labeled data to adapt to a specific task.

A surprisingly effective technique that emerged from this insight is the use of linear probes. In a linear probe, the pretrained backbone is kept entirely frozen, and only a simple linear classifier is trained on top of the extracted features. Despite its simplicity, this method often produces strong baselines in few-shot scenarios, sometimes even outperforming more complex meta-learning algorithms. This phenomenon highlights the tremendous value of high-quality representations: when the features are good enough, even minimal adaptation can achieve competitive performance. Linear probing also has practical benefits-it is fast to train, requires minimal computation, and avoids the risk of overfitting that can occur when fine-tuning large models on tiny datasets.

The success of linear probes has motivated a range of hybrid strategies that combine the strengths of large-scale pretraining with lightweight forms of task adaptation. Instead of fully fine-tuning the backbone, these methods partially adapt the network through techniques like adapter modules, prompt tuning, or selective layer tuning. These strategies allow a model to adjust its behavior to a new task while preserving the broad generalization capabilities learned during pretraining. For instance, adding small trainable modules-known as adapters-between layers allows for task-specific customization without modifying the entire network. Similarly, prompt tuning in vision–language models adapts the behavior of the model through learnable inputs rather than through modifying the backbone weights directly.

Such hybrid approaches strike a balance between flexibility and stability. They retain the robustness and broad knowledge encoded in the pretrained backbone while granting the model just enough adaptability to specialize for few-shot tasks. This balance has proven valuable across a wide range of domains, including medical imaging, remote sensing, fine-grained visual recognition, and natural language understanding. In many of these settings, data scarcity is inevitable, and the ability to adapt powerful pretrained models with only a small amount of supervision can be transformative.

The rise of pretrained visual and language backbones has reshaped how few-shot learning is approached. What once required intricate meta-learning algorithms can now often be achieved through smart use of pretrained representations and lightweight adaptation methods. As foundation models continue to grow and improve, their role in few-shot learning will only become more central, enabling efficient, scalable solutions in data-limited environments.

Prompting and In-Context Learning

Large language models (LLMs) have dramatically reshaped the landscape of few-shot and zero-shot learning, especially within natural language processing. Unlike traditional machine learning systems that require explicit fine-tuning or retraining for each new task, LLMs exhibit the remarkable ability to perform tasks directly through examples and instructions embedded in their input. This phenomenon, often referred to as in-context learning, was popularized by the work of Brown et al. (2020) with GPT-3. Their results showed that a sufficiently large model could perform few-shot or even zero-shot inference simply by conditioning on natural-language prompts-*without any gradient updates or parameter changes*. In other words, LLMs learn to "interpret" demonstrations provided at inference time, behaving as though they had been trained on the new task all along.

This capability is striking because it suggests that LLMs internalize broad patterns of language,

reasoning, and task structure during pretraining. When presented with a prompt containing instructions or examples, the model can generalize to new settings by leveraging its massive store of learned contextual associations. For example, a user might provide two or three examples of a sentiment classification task and then ask the model to classify a new sentence. Even though the model has never been explicitly trained on that specific task, it can often infer the underlying pattern from context alone. Similarly, in zero-shot settings, the model can perform tasks like translation, summarization, or question answering based solely on a clear instruction, relying on the rich semantic knowledge encoded in its pretrained representations.

As the capabilities of LLMs have grown, so too has the importance of prompt engineering, the practice of crafting effective instructions or example formats to steer the model's behavior. Prompt engineering allows users to adapt LLMs to new tasks simply by phrasing the request in a particular way, often achieving surprisingly strong performance without any retraining. Techniques such as chain-of-thought prompting, instruction templates, and task-specific prompt patterns have become crucial tools for extracting consistent results from models in both few-shot and zero-shot scenarios.

Building on these ideas, prompt tuning offers a more systematic and parameter-efficient way to adapt LLMs to specific applications. Instead of modifying the model's core parameters, prompt tuning introduces small, learnable vectors-often called *soft prompts*-that are prepended to the model's input. These prompts represent a tiny fraction of the model's overall size but can be optimized for a particular task with minimal data. This makes prompt tuning particularly attractive in low-resource settings, where full fine-tuning would be computationally expensive or prone to overfitting. By training only a lightweight prompt module while keeping the main model frozen, researchers achieve strong performance with significantly reduced computational cost.

Together, in-context learning, prompt engineering, and prompt tuning illustrate a shift toward more flexible and user-driven ways of adapting language models. Instead of relying on heavy retraining, modern NLP workflows increasingly harness the inherent generalization abilities of LLMs, enabling parameter-efficient and data-efficient adaptation. As these models continue to grow in scale and capability, their role in few-shot and zero-shot learning will only deepen, offering powerful tools for tackling tasks in domains where labeled data are limited or unavailable.

Multimodal Contrastive Pretraining and Zero-Shot Vision

Contrastive pretraining on image–text pairs has emerged as a powerful technique for enabling zero-shot classification, fundamentally changing how models learn to connect visual and linguistic information. Unlike traditional supervised approaches, which rely on labeled examples for every class of interest, contrastive pretraining allows models to learn rich, aligned embeddings from large collections of paired images and textual descriptions. The core idea is intuitive: the model is trained to bring representations of matching images and text closer together in a shared embedding space, while pushing representations of mismatched pairs farther apart. This process encourages the model to capture semantic relationships between visual content and natural language, effectively teaching it to "understand" images in terms of textual concepts.

A landmark example of this approach is CLIP (Contrastive Language–Image Pretraining) introduced by Radford et al. (2021). CLIP leverages hundreds of millions of image–text pairs to train dual encoders-one for images and one for text-so that their embeddings can be directly compared. Once trained, the model can perform zero-shot image classification by simply mapping textual class descriptions into the visual embedding space. For instance, given an image and a set of class labels such as "cat," "dog," or "horse," CLIP computes the similarity between the image embedding and each textual label embedding. The label with the highest similarity is selected, enabling accurate classification without requiring a single labeled image from the target classes. This capability represents a dramatic departure from conventional approaches: the model can generalize to completely new categories simply by understanding their semantic descriptions.

The implications of this are profound. Zero-shot transfer with CLIP and similar models eliminates the need for task-specific datasets in many cases, making it particularly valuable in domains where labeled data are scarce, costly, or time-consuming to collect. It also enables more flexible AI systems, where new categories can be introduced at inference time simply by providing textual descriptions. Beyond classification, contrastive pretraining has proven useful in other vision-language tasks such as retrieval, captioning, and multimodal reasoning, highlighting the versatility of aligned embeddings.

Moreover, the success of contrastive pretraining demonstrates the broader principle that large-scale, multimodal pretraining can imbue models with generalizable knowledge. By learning to connect visual patterns to natural language, the model internalizes rich conceptual representations that

transfer well to downstream tasks without task-specific fine-tuning. This approach has inspired a wave of research exploring extensions such as promptable vision–language models, fine-grained attribute-based zero-shot learning, and applications in scientific imaging, medical diagnostics, and robotics.

In summary, contrastive image–text pretraining, exemplified by CLIP, represents a major advance in zero-shot learning. By aligning visual and textual representations, these models enable powerful generalization to unseen classes, providing a practical and scalable solution for tasks where labeled examples are limited or unavailable. This paradigm illustrates how leveraging naturally paired data at scale can bridge the gap between perception and language, bringing AI closer to human-like generalization.

Hybrid and Adapter-based Approaches

Adapters and hybrid techniques have emerged as a practical and efficient approach for adapting large pretrained models to new tasks, especially in low-data scenarios. Rather than fine-tuning the entire backbone-which can contain hundreds of millions or even billions of parameters-these methods introduce small, trainable modules into an otherwise frozen network. The frozen backbone retains the broad knowledge and general representations acquired during pretraining, while the lightweight adapters provide the flexibility needed to specialize the model for a particular task. This approach strikes a balance between performance and computational efficiency, allowing models to adapt effectively without the expense and risk of full fine-tuning.

The intuition behind adapters is straightforward: pretrained models encode a wealth of general-purpose features that are often transferable across many tasks. However, each new task may require subtle adjustments or task-specific knowledge that is not fully captured by the frozen backbone. By inserting small trainable layers-typically a few bottleneck layers or parameter-efficient transformations-between existing layers of the network, adapters provide the model with the ability to modify its representations locally. These adjustments are sufficient to capture task-specific nuances while keeping most of the pretrained parameters unchanged. As a result, adapters reduce memory usage, accelerate training, and lower the risk of overfitting, which is particularly important in few-shot or low-label settings.

Hybrid techniques extend this concept by combining multiple adaptation strategies. For example, some methods use adapters alongside lightweight fine-tuning of select layers, prompt tuning, or residual task-specific modules. This combination allows the model to benefit from both the stability of the frozen backbone and the flexibility of task-specific components. In practice, hybrid approaches have proven effective across a wide range of domains, from natural language processing to computer vision and multimodal applications. For instance, in NLP, adapters can enable a large language model to specialize in sentiment analysis, domain-specific classification, or question answering with only a small number of labeled examples. In vision tasks, similar strategies allow pretrained convolutional or transformer-based models to adapt to fine-grained recognition, medical imaging, or remote sensing applications without expensive retraining.

One of the key advantages of adapters and hybrid techniques is their parameter efficiency. Since only a small fraction of the network is updated during training, these methods require significantly less compute and storage compared to full fine-tuning. This makes them particularly attractive for deploying large models in resource-constrained environments, such as mobile devices, edge computing, or cloud-based services with cost limitations. At the same time, adapters maintain strong performance, often approaching or matching the accuracy of fully fine-tuned models in few-shot or low-data scenarios.

In summary, adapters and hybrid adaptation techniques offer a practical and elegant solution to the challenges of low-label learning. By combining frozen pretrained backbones with small, trainable modules, these methods enable effective task-specific adaptation while dramatically reducing computational costs. They highlight a broader trend in modern machine learning: leveraging large pretrained models as a stable foundation while using lightweight, efficient mechanisms to customize behavior for new tasks. As model sizes continue to grow, such parameter-efficient strategies are likely to play an increasingly important role in making state-of-the-art models accessible, adaptable, and scalable across a wide range of applications.

Cross-Domain Few-Shot Learning and Applications

Datasets play a crucial role in evaluating and comparing few-shot learning methods, providing standardized benchmarks that allow researchers to measure progress and identify strengths and limitations of different approaches. Among the most widely used datasets are Omniglot and MiniImageNet**.** Omniglot, often described as the "transcribed alphabet dataset," contains thousands of handwritten characters from dozens of alphabets. Its relatively small image size and diverse set of classes make it an ideal benchmark for initial few-shot experiments, particularly for evaluating a model's ability to recognize entirely new character classes from just a few examples. MiniImageNet, on the other hand, is a scaled-down version of the more familiar ImageNet dataset, containing 100 classes

with 600 images per class. It is commonly used to assess few-shot classification performance in natural images and provides a more challenging and realistic test bed than Omniglot due to greater intra-class variability and visual complexity.

While these standard benchmarks are invaluable, there is increasing recognition that few-shot learning methods must be evaluated beyond single-domain tasks to assess their generalization and transfer capabilities. This has led to the development of cross-domain few-shot learning benchmarks, which test a model's ability to adapt when the training and testing distributions differ significantly. For example, a model trained on natural images such as MiniImageNet may be evaluated on entirely different domains like medical imagery, satellite images, or sketches. These cross-domain evaluations expose the limitations of models that overfit to a particular dataset or domain and provide insight into the robustness and versatility of different few-shot learning approaches.

Cross-domain few-shot learning is particularly important in practical applications. In medical AI, for instance, large amounts of labeled data are typically available in one hospital or imaging modality, but models must generalize to other hospitals, devices, or patient populations where data is scarce. Similarly, in environmental monitoring, a model trained on wildlife photographs from one region may need to classify species in another region with very different visual characteristics. Standard within-domain benchmarks cannot capture these challenges, making cross-domain evaluation essential for understanding how well few-shot learning methods will perform in real-world, dynamic settings.

Moreover, cross-domain evaluation highlights the importance of robust representations and pretraining strategies. Models that rely heavily on domain-specific cues may perform well on standard benchmarks but fail to generalize to unseen domains. In contrast, methods that learn transferable features-through metric learning, meta-learning, or large-scale pretraining-tend to exhibit stronger cross-domain performance. By systematically testing models across diverse domains, researchers can identify which techniques truly capture generalizable patterns rather than memorizing idiosyncrasies of a particular dataset.

In summary, while Omniglot and MiniImageNet serve as foundational benchmarks for developing and comparing few-shot learning models, cross-domain few-shot learning provides a more realistic and demanding evaluation framework. It challenges models to transfer knowledge across domains, simulating the kinds of low-data scenarios encountered in real-world applications. By incorporating both standard and cross-domain benchmarks, researchers gain a more complete understanding of a model's strengths, limitations, and potential for practical deployment.

Evaluation and Reproducibility

Episodic evaluation with multiple seeds and confidence intervals is standard. Reproducibility requires consistent dataset splits and augmentation policies.

Open Challenges

Despite the significant progress in few-shot and zero-shot learning, several key challenges remain that continue to drive research in this area. One of the most pressing issues is domain-shift robustness. Many models perform well when the training and testing data come from similar distributions, but in real-world scenarios, data often differ in unexpected ways. For example, a model trained on high-quality, curated images might struggle when applied to low-resolution medical scans, satellite imagery, or photos captured under different lighting conditions. Developing models that can generalize reliably across such shifts is critical, especially in safety-sensitive applications like healthcare, autonomous driving, and security, where errors can have serious consequences.

Another major challenge is uncertainty estimation. In few-shot and zero-shot settings, models often make predictions with limited evidence, which can lead to overconfident mistakes. Being able to quantify uncertainty allows practitioners to know when to trust model predictions and when to seek additional human input or data. Techniques such as Bayesian methods, ensemble models, and probabilistic embeddings are being explored to address this, but achieving accurate, calibrated uncertainty estimates in low-data regimes remains an open problem.

Integrating multimodal prior knowledge is also an important frontier. Humans learn from multiple sources-text, images, audio, and prior experiences-and can combine these signals to understand new concepts quickly. Similarly, AI systems that can fuse information from diverse modalities have the potential to generalize more effectively in data-scarce environments. For instance, combining textual descriptions with visual features or leveraging structured knowledge graphs alongside sensory data could improve zero-shot and few-shot performance, but doing so efficiently and reliably poses significant modeling and computational challenges.

Finally, a central concern across all applications is reducing annotation and computational costs while maintaining reliable performance. Collecting labeled data is often expensive, time-consuming, or

even impractical in specialized domains such as medical imaging, rare-language NLP, or anomaly detection. Large pretrained models and complex meta-learning frameworks can achieve impressive results, but they come with high computational demands, limiting accessibility for many organizations. Developing lightweight, parameter-efficient methods-such as adapters, prompt tuning, or hybrid strategies-offers a path forward, but balancing efficiency, accuracy, and generalization is still an ongoing challenge.

Addressing these open problems is not only a technical endeavor but also a prerequisite for the responsible deployment of AI in high-stakes environments. Models must be robust, trustworthy, and adaptable to truly meet real-world needs. By tackling domain shifts, improving uncertainty estimation, integrating multimodal knowledge, and reducing resource requirements, future research can bring AI systems closer to human-like flexibility and reliability, enabling them to perform effectively even when labeled data are scarce or tasks are entirely novel.

Methodology

Few-shot and zero-shot learning methods can be broadly categorized into three main groups: metric-based approaches**,** optimization-based meta-learners, and prompt-based models. Each of these paradigms offers a distinct strategy for tackling the challenges of learning with little or no labeled data, reflecting different assumptions about how knowledge can be transferred and generalized across tasks. Understanding these categories provides a useful framework for evaluating, comparing, and designing models for low-data scenarios.

Metric-based approaches form one of the most intuitive and widely studied families of few-shot learning methods. The central idea is that models can perform classification by learning a feature embedding space in which examples from the same class are close to one another, while examples from different classes are farther apart. At test time, new examples are classified by comparing their embeddings to those of a small support set of labeled examples. Models such as Prototypical Networks exemplify this approach: they compute a prototype for each class by averaging the embeddings of the support examples and assign labels to query instances based on distance to these prototypes, typically using Euclidean or cosine distance. Other metric-based models, such as Matching Networks or Relation Networks, extend this concept by incorporating attention mechanisms or learned similarity functions to improve performance on more complex tasks. The appeal of metric-based approaches lies in their simplicity, interpretability, and computational efficiency. They require little adaptation at test time, making them particularly suitable for domains where computational resources are limited or rapid inference is necessary.

Optimization-based meta-learners take a different perspective. Instead of relying primarily on embedding comparisons, these models focus on learning parameters that can be quickly adapted to new tasks using only a few gradient updates. A canonical example is Model-Agnostic Meta-Learning (MAML), which seeks an initialization of model parameters such that fine-tuning on a small support set yields strong performance on a new task. The intuition is analogous to how humans transfer skills across domains: prior experience prepares us to learn new tasks efficiently, even with minimal data. Variants of MAML and related methods, such as Reptile or Meta-SGD, explore improvements in stability, efficiency, and adaptability, while maintaining the core principle of meta-learning the learning process itself. Optimization-based methods are particularly powerful when tasks vary significantly or when complex adaptation is required, as they provide a structured mechanism for adjusting model parameters quickly without overfitting to the small available dataset.

Prompt-based approaches represent the third major category and have emerged primarily from advances in large pretrained language and multimodal models. Unlike metric- or optimization-based methods, prompt-based models can perform inference on new tasks without any gradient updates, relying instead on carefully crafted task descriptions or example prompts. These models, often based on GPT-style architectures or multimodal models like CLIP, interpret natural language prompts to condition their predictions. In few-shot settings, a small number of examples can be included in the prompt (in-context learning), while in zero-shot settings, task instructions alone are sufficient. Prompt-based methods are highly flexible and have demonstrated impressive generalization across NLP tasks, image captioning, and even vision-language reasoning, highlighting the potential of pretrained models to encode broad prior knowledge that can be leveraged in novel contexts.

To explore the practical applications of these approaches, we implemented Prototypical Networks for few-shot classification tasks and conducted prompt-based evaluations using GPT-style models for zero-shot inference. Prototypical Networks were applied to MiniImageNet, a standard benchmark for few-shot image classification, which contains 100 classes of natural images with 600 examples per class. This dataset allows us to evaluate how well metric-based embeddings support generalization when only a few labeled examples are available. For zero-shot tasks, we leveraged both CLIP and GPT-style models. CLIP provides a framework for multimodal zero-shot evaluation by mapping textual

class descriptions to visual embeddings, enabling classification of unseen categories without additional training. GPT-based prompt evaluation was conducted across GLUE, a widely used suite of NLP tasks that tests capabilities in natural language understanding, including sentiment analysis, textual entailment, and question answering. By using prompts and task instructions alone, we assessed the ability of pretrained language models to generalize to tasks for which they had not been explicitly trained.

This experimental design highlights the complementary strengths of each approach. Metric-based methods excel when a small labeled support set is available and embeddings can be learned to capture task-specific relationships. Optimization-based meta-learners are particularly effective in settings where the model must rapidly adapt to tasks with subtle differences in structure or distribution. Prompt-based models, in contrast, offer a parameter-efficient alternative that leverages the general knowledge embedded in large pretrained networks, performing well even in zero-shot scenarios where labeled data are entirely unavailable. Together, these categories provide a comprehensive toolkit for tackling low-data problems across diverse domains, from computer vision to natural language processing.

In summary, categorizing few-shot and zero-shot methods into metric-based, optimization-based, and prompt-based approaches provides both conceptual clarity and practical guidance for research and application. Our implementations on Prototypical Networks for MiniImageNet, CLIP for multimodal zero-shot classification, and GPT-style prompt evaluations on GLUE illustrate the real-world utility of these methods and demonstrate how different strategies can be applied depending on task characteristics and data availability. By understanding the strengths and limitations of each paradigm, researchers and practitioners can design more effective, flexible, and efficient models for low-data learning scenarios.

Experiments

To empirically explore the effectiveness of few-shot and zero-shot learning approaches, we conducted a series of pilot experiments using representative models and datasets. For few-shot learning, we focused on a subset of MiniImageNet, implementing a 5-way 1-shot classification setup with Prototypical Networks. In this configuration, each episode consists of five classes, with one labeled support example per class and five query examples used to evaluate the model's predictions. This episodic structure mirrors the low-data learning scenario that few-shot models are designed to handle, forcing the network to generalize from extremely limited labeled information. By repeating this process across multiple episodes, the model learns to construct meaningful class prototypes and classify novel instances based on their proximity to these reference points in the embedding space.

During training, we performed 100 episodes per epoch over a total of 10 epochs, providing sufficient exposure to diverse task compositions while maintaining computational efficiency. Each episode presents a unique combination of classes, encouraging the model to learn a robust embedding space that can accommodate previously unseen classes at test time. This experimental setup is widely used in the few-shot learning literature, as it balances the need for repeated task-specific adaptation with practical considerations of training time and computational resources. The network's performance was evaluated using accuracy, averaged across episodes, as well as the standard deviation to assess consistency and stability of predictions under the high-variance conditions typical of few-shot learning.

For zero-shot evaluation, we leveraged CLIP, a multimodal model trained on large-scale image–text pairs. Unlike few-shot experiments, zero-shot classification does not rely on labeled support examples for the new classes. Instead, CLIP maps textual descriptions of unseen categories into the same embedding space as the images, allowing the model to determine class membership by measuring similarity between the visual and textual representations. This setup allows us to directly assess the model's ability to generalize to entirely novel categories without task-specific training, providing insight into the effectiveness of contrastive pretraining and semantic alignment in enabling zero-shot inference.

Together, these pilot experiments provide a comparative perspective on low-data learning. The few-shot experiments highlight the capability of metric-based approaches, like Prototypical Networks, to construct task-specific representations from minimal examples. In contrast, the zero-shot evaluation with CLIP demonstrates the power of pretrained multimodal embeddings to classify unseen categories purely from semantic descriptions. By reporting both accuracy and variability across episodes, we are able to quantify not only the central tendency of model performance but also its reliability under challenging low-data conditions.

This experimental framework illustrates how few-shot and zero-shot methods can be systematically evaluated, providing a foundation for understanding their strengths, limitations, and potential for practical deployment. These pilot studies serve as a stepping stone for more extensive evaluations across diverse datasets, tasks, and adaptation strategies,

helping to inform future research directions in data-efficient machine learning.

Results and Discussion

Our pilot experiments illustrate the strengths and limitations of current few-shot and zero-shot learning methods, providing insights into their practical performance and potential applications. In the few-shot setting, Prototypical Networks achieved consistently high accuracy on the 5-way 1-shot tasks, confirming their ability to generalize effectively from a very limited number of examples. By computing class prototypes and comparing query examples to these reference points in the embedding space, the model was able to correctly classify unseen instances despite having access to only a single labeled example per class. This result underscores the power of metric-based approaches in low-data regimes: when the embedding space is well-structured, even minimal supervision can yield strong predictive performance. Moreover, the relatively low variance across episodes indicates that Prototypical Networks are not only accurate but also stable, producing reliable predictions even under the high-variance conditions inherent to few-shot tasks.

In the zero-shot setting, CLIP demonstrated strong transfer capabilities, effectively classifying previously unseen categories based solely on textual descriptions. This highlights the potential of large-scale multimodal pretraining, where models learn rich, aligned representations of images and text from massive datasets. By leveraging this semantic alignment, CLIP can generalize to a wide variety of classes without requiring any task-specific labeled examples, offering a practical solution for scenarios where data collection is expensive, time-consuming, or otherwise infeasible. The zero-shot performance also illustrates the flexibility of pretrained foundation models, which can adapt to diverse tasks simply through appropriate textual prompts, bypassing the need for traditional gradient-based fine-tuning.

Despite these successes, several challenges remain. First, cross-domain adaptation continues to be a significant hurdle. While Prototypical Networks perform well within the same domain, their accuracy can degrade when applied to tasks that differ substantially from the training data, such as transferring from natural images to medical or satellite imagery. Similarly, even large-scale models like CLIP can struggle when textual descriptions do not perfectly capture the target domain or when domain-specific visual characteristics are underrepresented in the pretraining data. Second, uncertainty estimation remains an open problem in both few-shot and zero-shot settings. Models often produce confident predictions even when they encounter ambiguous or out-of-distribution examples, which can be particularly problematic in high-stakes applications like healthcare or autonomous systems. Finally, achieving efficient adaptation in low-resource domains is still a challenge. While pretrained models and metric-based approaches reduce the need for large labeled datasets, they can still be computationally expensive to deploy, and fine-tuning even a small number of parameters may be infeasible in some practical scenarios.

Our results illustrate both the promise and the limitations of current low-data learning approaches. Prototypical Networks and CLIP demonstrate that it is possible to achieve strong generalization from limited supervision or even zero examples, highlighting the value of structured embeddings and large-scale pretraining. At the same time, the challenges in domain transfer, uncertainty quantification, and computational efficiency emphasize the need for continued research in building more robust, adaptable, and resource-efficient models. These findings provide guidance for future work, suggesting that combining robust embeddings, multimodal priors, and lightweight adaptation strategies may be a fruitful path toward scalable and reliable few-shot and zero-shot learning.

Conclusion

Few-shot and zero-shot learning represent a transformative shift in the design and application of machine learning models, offering the ability to generalize from limited or even no labeled data. Traditional supervised learning approaches have long relied on large, carefully labeled datasets to achieve high performance, but in many real-world settings, such data are scarce, expensive, or practically impossible to collect. Few-shot learning addresses this challenge by enabling models to learn new concepts from just a handful of labeled examples, while zero-shot learning pushes the boundary further, allowing models to perform tasks they have never seen before by leveraging auxiliary knowledge, such as textual descriptions, attributes, or prior pretraining. Together, these paradigms expand the scope of machine learning systems, making them more flexible, efficient, and capable of handling diverse tasks in low-data regimes.

A variety of methodological approaches have been developed to tackle these problems, each with its own strengths and limitations. Metric-based methods, such as Prototypical Networks, Matching Networks, and Relation Networks, rely on the principle that learning a meaningful embedding space allows models to classify new examples by comparing them to a small support set. These approaches are intuitive and computationally efficient, providing reliable performance even with extremely limited supervision. They excel when the

embedding space captures the essential similarities and differences between classes, but their performance may degrade under significant domain shifts or when task complexity increases. Optimization-based meta-learners, on the other hand, focus on learning parameters or update rules that can be rapidly adapted to new tasks using only a few gradient steps. Techniques like MAML and its variants are particularly effective in scenarios where tasks vary substantially or require nuanced adaptation, offering flexibility and principled mechanisms for transfer learning, though they can be computationally intensive. Prompt-based methods, which have emerged in the era of large pretrained models, leverage the knowledge embedded in foundation models such as GPT and CLIP to perform inference directly from task descriptions or in-context examples. These models are highly flexible and enable zero-shot or few-shot performance without modifying the model parameters, making them both parameter-efficient and scalable.

Our pilot experiments illustrate the complementary strengths of these approaches. Using a subset of MiniImageNet, Prototypical Networks demonstrated strong performance on 5-way 1-shot classification tasks, effectively generalizing from extremely limited labeled data. The episodic training strategy allowed the model to construct robust class prototypes and accurately classify unseen examples, highlighting the power of metric-based embeddings in few-shot scenarios. For zero-shot evaluation, we leveraged CLIP, a multimodal model pretrained on vast image–text datasets. CLIP successfully classified unseen categories based solely on textual descriptions, showcasing the remarkable generalization capabilities afforded by large-scale multimodal pretraining. Together, these results confirm that both structured embedding spaces and pretrained knowledge can provide effective solutions for low-data tasks, while also revealing the challenges that remain in achieving consistent performance across domains.

Despite these successes, several open challenges and opportunities for future research remain. One critical area is cross-domain robustness. While few-shot and zero-shot models perform well within the domains on which they were trained or pre-exposed, their performance often declines when faced with tasks from substantially different distributions, such as transferring from natural images to medical or satellite imagery. Improving domain generalization is essential for deploying these models in real-world, heterogeneous environments. Another promising direction involves hybrid approaches that combine the strengths of multiple paradigms-for instance, integrating metric-based embeddings with optimization-based adaptation or combining pretrained multimodal models with lightweight adapters and prompt tuning. Such hybrid strategies may offer the flexibility to adapt to diverse tasks while maintaining computational efficiency. Finally, achieving scalable deployment in practical applications remains a major concern. Models must not only perform well under low-data conditions but also operate efficiently in terms of memory, compute, and energy, particularly when applied in resource-constrained or time-sensitive contexts. Parameter-efficient methods, lightweight adapters, and prompt-based adaptation provide promising avenues to address these concerns.

In summary, few-shot and zero-shot learning have opened new possibilities for data-efficient machine learning, enabling models to generalize from minimal supervision and even perform unseen tasks. Metric-based, optimization-based, and prompt-based approaches each offer unique advantages, as evidenced by the strong performance of Prototypical Networks and multimodal models like CLIP in pilot experiments. However, challenges remain in ensuring robust cross-domain transfer, efficient adaptation, and practical scalability. By addressing these challenges, future research can push these paradigms closer to human-like flexibility, making machine learning systems more versatile, accessible, and capable in real-world, low-data settings.

Conflict of Interest

The authors declare no conflict of interest in conducting this research.

References

1. Vinyals, T., Blundell, C., Lillicrap, T. & Wierstra, D. (2016). Matching networks for one shot learning. NeurIPS.
2. Snell, J., Swersky, K. & Zemel, R. (2017). Prototypical networks for few-shot learning. NeurIPS.
3. Finn, C., Abbeel, P. & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. ICML.
4. Brown, T.B. & et. al. (2020). Language models are few-shot learners. arXiv:2005, 14165.
5. Radford, A. & et. al. (2021). Learning transferable visual models from natural language supervision. arXiv:2103.00020.
6. Sung, F. & et. al. (2018). Learning to compare: Relation network for few-shot learning. CVPR.