Context-Adaptive Inference: Bridging Statistical and Foundation Models

This manuscript (permalink) was automatically generated from AdaptInfer/context-review@f99ee82 on September 4, 2025.

Authors

- **Ben Lengerich**
  0000-0001-8690-9554 · blengerich · ben_lengerich
  Department of Statistics, University of Wisconsin-Madison · Funded by None

- **Caleb N. Ellington**
  0000-0001-7029-8023 · cnellington · probablybots
  Computational Biology Department, Carnegie Mellon University · Funded by None

- **Yue Yao**
  0009-0000-8195-3943 · YueYao-stat
  Department of Statistics, University of Wisconsin-Madison · Funded by None

- **Dong Liu**
  0009-0009-6815-8297 · NoakLiu
  Department of Computer Science, Yale University · Funded by None

✉ — Correspondence possible via GitHub Issues

Abstract

Context-adaptive inference enables models to adjust their behavior across individuals, environments, or tasks. This adaptivity may be *explicit*, through parameterized functions of context, or *implicit*, as in foundation models that respond to prompts and support in-context learning. In this review, we connect recent developments in varying-coefficient models, contextualized learning, and in-context learning. We highlight how foundation models can serve as flexible encoders of context, and how statistical methods offer structure and interpretability. We propose a unified view of context-adaptive inference and outline open challenges in developing scalable, principled, and personalized models that adapt to the complexities of real-world data.

Introduction

A convenient simplifying assumption in statistical modeling is that observations are independent and identically distributed (i.i.d.). This assumption allows us to use a single model to make predictions across all data points. But in practice, this assumption rarely holds. Data are collected across different individuals, environments, and tasks – each with their own characteristics, constraints, and dynamics.

To model this heterogeneity, a growing class of methods aim to make inference *adaptive to context*. These include varying-coefficient models in statistics, transfer and meta-learning in machine learning, and in-context learning in large foundation models. Though these approaches arise from different traditions, they share a common goal: to use contextual information – whether covariates, environments, or support sets – to inform sample-specific inference.

We formalize this by assuming each observation x_i is drawn from a distribution governed by parameters θ_i:

$$x_i \sim P(x; \theta_i).$$

In population models, the assumption is that $\theta_i = \theta$ for all i. In context-adaptive models, we instead posit that the parameters vary with context:

$$\theta_i = f(c_i) \quad \text{or} \quad \theta_i \sim P(\theta \mid c_i),$$

where c_i captures the relevant covariates or environment for observation i. The goal is to estimate either a deterministic function f or a conditional distribution over parameters.

This shift raises new modeling challenges. Estimating a unique θ_i from a single observation is ill-posed unless we impose structure—smoothness, sparsity, shared representations, or latent grouping. And as adaptivity becomes more implicit (e.g., via neural networks or black-box inference), we must develop tools to recover, interpret, or constrain the underlying parameter variation.

In this review, we examine methods that use context to guide inference, either by specifying how parameters change with covariates or by learning to adapt behavior implicitly. We begin with classical models that impose explicit structure—such as varying-coefficient models and multi-task learning—and then turn to more flexible approaches like meta-learning and in-context learning with foundation models. Though these methods arise from different traditions, they share a common goal: to tailor inference to the local characteristics of each observation or task. Along the way, we highlight recurring themes: complex models often decompose into simpler, context-specific components; foundation models can both adapt to and generate context; and context-awareness challenges classical assumptions of homogeneity. These perspectives offer a unifying lens on recent advances and open new directions for building adaptive, interpretable, and personalized models.

From Population Assumptions to Context-Adaptive Inference

Most statistical and machine learning models begin with a foundational assumption: that all samples are drawn independently and identically from a shared population distribution. This assumption simplifies estimation and enables generalization from limited data, but it collapses in the presence of meaningful heterogeneity.

In practice, data often reflect differences across individuals, environments, or conditions. These differences may stem from biological variation, temporal drift, site effects, or shifts in measurement context. Treating heterogeneous data as if it were homogeneous can obscure real effects, inflate variance, and lead to brittle predictions.

Failure Modes of Population Models

Even when traditional models appear to fit aggregate data well, they may hide systematic failure modes.

Mode Collapse
When one subpopulation is much larger than another, standard models are biased toward the dominant group, underrepresenting the minority group in both fit and predictions.

Outlier Sensitivity
In the parameter-averaging regime, small but extreme groups can disproportionately distort the global model, especially in methods like ordinary least squares.

Phantom Populations
When multiple subpopulations are equally represented, the global model may fit none of them well, instead converging to a solution that represents a non-existent average case.

These behaviors reflect a deeper problem: the assumption of identically distributed samples is not just incorrect, but actively harmful in heterogeneous settings.

Toward Context-Aware Models

To account for heterogeneity, we must relax the assumption of shared parameters and allow the data-generating process to vary across samples. A general formulation assumes each observation is governed by its own latent parameters:

$$x_i \sim P(x; \theta_i),$$

However, estimating N free parameters from N samples is underdetermined. Context-aware approaches resolve this by introducing structure on how parameters vary, often by assuming that θ_i depends on an observed context c_i:

$$\theta_i = f(c_i) \quad \text{or} \quad \theta_i \sim P(\theta \mid c_i).$$

This formulation makes the model estimable, but it raises new challenges. How should f be chosen? How smooth, flexible, or structured should it be? The remainder of this review explores different answers to this question, and shows how implicit and explicit representations of context can lead to powerful, personalized models.

Early Remedies: Grouped and Distance-Based Models

Before diving into flexible estimators of $f(c)$, we review early modeling strategies that attempt to break away from homogeneity.

Conditional and Clustered Models

One approach is to group observations into C contexts, either by manually defining conditions (e.g. male vs. female) or using unsupervised clustering. Each group is then assigned a distinct parameter vector:

$$\{\hat{\theta}_0, \ldots, \hat{\theta}_C\} = \arg\max_{\theta_0, \ldots, \theta_C} \sum_{c \in \mathcal{C}} \ell(X_c; \theta_c),$$

where $\ell(X; \theta)$ is the log-likelihood of θ on X and c specifies the covariate group that samples are assigned to. This reduces variance but limits granularity. It assumes that all members of a group share the same distribution and fails to capture variation within a group.

Distance-Regularized Estimation

A more flexible alternative assumes that observations with similar contexts should have similar parameters. This is encoded as a regularization penalty that discourages large differences in θ_i for nearby c_i:

$$\{\hat{\theta}_0, \ldots, \hat{\theta}_N\} = \arg\max_{\theta_0, \ldots, \theta_N} \left(\sum_i \ell(x_i; \theta_i) - \sum_{i,j} \frac{\|\theta_i - \theta_j\|}{D(c_i, c_j)} \right),$$

where $D(c_i, c_j)$ is a distance metric between contexts. This approach allows for smoother parameter variation but requires careful choice of D and regularization strength λ to balance bias and variance. The choice of distance metric D and regularization strength λ controls the bias–variance tradeoff.

Parametric Varying-coefficient models

Original paper (based on a smoothing spline function): [1] Markov networks: [2] Linear varying-coefficient models assume that parameters vary linearly with covariates, a much stronger assumption than the classic varying-coefficient model but making a conceptual leap that allows us to define a form for the relationship between the parameters and covariates.

$$\hat{\theta}_0, \ldots, \hat{\theta}_N = \widehat{A} C^T$$
$$\widehat{A} = \arg\max_A \sum_i \ell(x_i; A c_i)$$

TODO: Note that they achieve distance-matching by using a distance metric under Euclidean distance, which is a special case of the distance-regularized estimation above.

Semi-parametric varying-coefficient Models

Original paper: [3] 2-step estimation with RBF kernels: [4]

Classic varying-coefficient models assume that models with similar covariates have similar parameters, or – more formally – that changes in parameters are smooth over the covariate space. This assumption is encoded as a sample weighting, often using a kernel, where the relevance of a sample to a model is equivalent to its kernel similarity over the covariate space.

$$\hat{\theta}_0, \dots, \hat{\theta}_N = \arg \max_{\theta_0, \dots, \theta_N} \sum_{i,j} \frac{K(c_i, c_j)}{\sum_k K(c_i, c_k)} \ell(x_j; \theta_i)$$

This estimator is the simplest to recover N unique parameter estimates. However, the assumption here is contradictory to the partition model estimator. When the relationship between covariates and parameters is discontinuous or abrupt, this estimator will fail.

Contextualized Models

Seminal work [5] Contextualized ML generalization and applications: [6], [7], [8], [9], [10], [11], [12], [13]

Contextualized models make the assumption that parameters are some function of context, but make no assumption on the form of that function. In this regime, we seek to estimate the function often using a deep learner (if we have some differentiable proxy for probability):

$$\widehat{f} = \arg \max_{f \in \mathcal{F}} \sum_i \ell(x_i; f(c_i))$$

Latent-structure Models

Partition Models

Markov networks: [14] Partition models also assume that parameters can be partitioned into homogeneous groups over the covariate space, but make no assumption about where these partitions occur. This allows the use of information from different groups in estimating a model for a each covariate. Partition model estimators are most often utilized to infer abrupt model changes over time and take the form

$$\hat{\theta}_0, \dots, \hat{\theta}_N = \arg \max_{\theta_0, \dots, \theta_N} \sum_i \ell(x_i; \theta_i) + \sum_{i=2}^N \mathrm{TV}(\theta_i, \theta_{i-1})$$

Where the regularizaiton term might take the form

$$\mathrm{TV}(\theta_i, \theta_{i-1}) = |\theta_i - \theta_{i-1}|$$

This still fails to recover a unique parameter estimate for each sample, but gets closer to the spirit of personalized modeling by putting the model likelihood and partition regularizer in competition to find the optimal partitions.

Fine-tuned Models and Transfer Learning

Review: [15] Noted in foundational literature for linear varying coefficient models [3]

Estimate a population model, freeze these parameters, and then include a smaller set of personalized parameters to estimate on a smaller subpopulation.

$$\widehat{\gamma} = \arg \max_{\gamma} = \ell(\gamma; X)$$
$$\widehat{\theta_c} = \arg \max_{\theta_c} = \ell(\theta_c; \widehat{\gamma}, X_c)$$

Context-informed and Latent-structure models

Seminal paper: [16]

Key idea: negative information sharing. Different models should be pushed apart.

$$\hat{\theta}_0, \ldots, \hat{\theta}_N = \arg\max_{\theta_0, \ldots, \theta_N, D} \sum_{i=0}^{N} \prod_{j=0 s.t. D(c_i, c_j) < d}^{N} P(x_j; \theta_i) P(\theta_i; \theta_j)$$

A Spectrum of Context-Awareness

Context-aware models can be viewed along a spectrum of assumptions about the relationship between context and parameters.

Global models: $\theta_i = \theta$ for all i
Grouped models: $\theta_i = \theta_c$ for some finite set of groups
Smooth models: $\theta_i = f(c_i)$, with f assumed to be continuous or low-complexity
Latent models: $\theta_i \sim P(\theta | c_i)$, with f learned implicitly

Each of these choices encodes different beliefs about how parameters vary. The next section formalizes this variation and examines general principles for adaptivity in statistical modeling.

Relevant references:

- Can Subpopulation Shifts Explain Disagreement in Model Generalization? [17]

Principles of Context-Adaptive Inference

What makes a model adaptive? When is it good for a model to be adaptive? While the appeal of adaptivity lies in flexibility and personalized inference, not all adaptivity is good adaptivity. In this section, we formalize the core principles that underlie adaptive modeling.

1. Adaptivity requires flexibility

A model cannot adapt unless it has the capacity to represent multiple behaviors. Flexibility may take the form of nonlinearity, hierarchical structure, or modular components that allow different responses in different settings.

- Interaction effects in regression models [18]
- Hierarchical models that allow for varying effects across groups
- Meta-learning and mixtures-of-experts models that learn to adapt based on context
- Varying-coefficient models that allow coefficients to change with context [1]

2. Adaptivity requires a signal of heterogeneity

- Varying-coefficient models adapt parameters based on observed context [1]
- Contextual bandits adapt actions to context features [19]
- Multi-domain models adapt across known environments or inferred partitions [20]

3. Modularity improves adaptivity

Adaptive systems are easier to design, debug, and interpret when built from modular parts. Modularity supports targeted adaptation, transferability, and disentanglement.

- []

4. Adaptivity implies selectivity

Adaptation must be earned. Overreacting to limited data leads to overfitting. The best adaptive methods include mechanisms for deciding when not to adapt. - Lepski's method [21] - Aggregation of classifiers [22]

5. Adaptivity is bounded by data efficiency

[23]

6. Adaptivity is not a free lunch

Adaptivity improves performance when heterogeneity is real and informative, but it can degrade performance when variation is spurious. Key tradeoffs include:

- **Bias vs. variance**: More flexible adaptation can reduce bias but increase variance
- **Stability vs. personalization**: Highly adaptive models may overfit to noise or adversarial context
- **Inference cost**: Adaptive inference may be more computationally intensive than global prediction

Understanding these tradeoffs is essential when designing systems for real-world deployment.

When Adaptivity Fails: Common Failure Modes

Even when all the ingredients are present, adaptivity can backfire. Common failure modes include:

- Spurious Adaptation: Adapting to unstable or confounded features [24]
- Overfitting in Low-Data Contexts: Attempting fine-grained adaptation with insufficient signal
- Modularity Mis-specification: Adapting in the wrong units or groupings [25]
- Feedback Loops: Models that change the data distribution they rely on [26]

Related references:

Explicit Adaptivity: Structured Estimation of $f(c)$

In classical statistical modeling, all observations are typically assumed to share a common set of parameters. However, modern datasets often display significant heterogeneity across individuals, locations, or experimental conditions, making this assumption unrealistic in many real-world applications. To better capture such heterogeneity, recent approaches model parameters as explicit functions of observed context, formalized as $\theta_i = f(c_i)$, where f maps each context to a sample-specific parameter [27].

This section systematically reviews explicit adaptivity methods, with a focus on structured estimation of $f(c)$. We begin by revisiting classical varying-coefficient models, which provide a conceptual and methodological foundation for modeling context-dependent effects. We then categorize recent advances in explicit adaptivity according to three principal strategies for estimating $f(c)$: (1) smooth nonparametric models that generalize classical techniques, (2) structurally constrained models that incorporate domain-specific knowledge such as spatial or network structure, and (3) learned function approximators that leverage machine learning methods for high-dimensional or complex contexts.

Finally, we summarize key theoretical developments and highlight promising directions for future research in this rapidly evolving field.

Classical Varying-Coefficient Models: A Foundation

Varying-coefficient models (VCMs) are a foundational tool for modeling heterogeneity, as they allow model parameters to vary smoothly with observed context variables [27,28,29]. In their original formulation, the regression coefficients are treated as nonparametric functions of low-dimensional covariates, such as time or age. The standard VCM takes the form

$$y_i = \sum_{j=1}^{p} \beta_j(c_i) x_{ij} + \varepsilon_i,$$

where each $\beta_j(c)$ is an unknown smooth function, typically estimated using kernel smoothing, local polynomials, or penalized splines [28].

This approach provides greater flexibility than fixed-coefficient models and is widely used for longitudinal and functional data analysis. The assumption of smoothness makes estimation and theoretical analysis more tractable, but also imposes limitations. Classical VCMs work best when the context is low-dimensional and continuous. They may struggle with abrupt changes, discontinuities, or high-dimensional and structured covariates. In such cases, interpretability and accuracy can be compromised, motivating the development of a variety of modern extensions, which will be discussed in the following sections.

Advances in Modeling $f(c)$

Recent years have seen substantial progress in the modeling of $f(c)$, the function mapping context to model parameters. These advances can be grouped into three major strategies: (1) **smooth non-parametric models** that extend classical flexibility; (2) **structurally constrained approaches** that encode domain knowledge such as spatial or network topology; and (3) high-capacity **learned function approximators** from machine learning designed for high-dimensional, unstructured contexts. Each strategy addresses specific challenges in modeling heterogeneity, and together they provide a comprehensive toolkit for explicit adaptivity.

Smooth Non-parametric Models

This family of models generalizes the classical VCM by expressing $f(c)$ as a flexible, smooth function estimated with basis expansions and regularization. Common approaches include spline-based methods, local polynomial regression, and RKHS-based frameworks. For instance, [28] developed a semi-nonparametric VCM using RKHS techniques for imaging genetics, enabling the model to capture complex nonlinear effects. Such methods are central to generalized additive models, supporting both flexibility and interpretability. Theoretical work has shown that penalized splines and kernel methods offer strong statistical guarantees in moderate dimensions, although computational cost and overfitting can become issues as the dimension of c increases.

Structurally Constrained Models

Another direction focuses on incorporating structural information into $f(c)$, especially when the context is discrete, clustered, or topologically organized.

Piecewise-Constant and Partition-Based Models. Here, model parameters are allowed to remain constant within specific regions or clusters of the context space, rather than vary smoothly.

Approaches include classical grouped estimators and modern partition models, which may learn changepoints using regularization tools like total variation penalties or the fused lasso. This framework is particularly effective for data with abrupt transitions or heterogeneous subgroups.

Structured Regularization for Spatial, Graph, and Network Data. When context exhibits known structure, regularization terms can be designed to promote similarity among neighboring coefficients [30]. For example, spatially varying-coefficient models have been applied to problems in geographical analysis and econometrics, where local effects are expected to vary across adjacent regions [31,32,33,34]. On networked data, the network VCM of [35] generalizes these ideas by learning both the latent positions and the parameter functions on graphs, allowing the model to accommodate complex relational heterogeneity. Such structural constraints allow models to leverage domain knowledge, improving efficiency and interpretability where smooth models may struggle.

Learned Function Approximators

A third class of methods is rooted in modern machine learning, leveraging high-capacity models to approximate $f(c)$ directly from data. These approaches are especially valuable when the context is high-dimensional or unstructured, where classical assumptions may no longer be sufficient.

Tree-Based Ensembles. Gradient boosting decision trees (GBDTs) and related ensemble methods are well suited to tabular and mixed-type contexts. The framework developed by [36] extends varying-coefficient models by integrating gradient boosting, achieving strong predictive performance with a level of interpretability. These models are typically easier to train and tune than deep neural networks, and their structure lends itself to interpretation with tools such as SHAP.

Deep Neural Networks. For contexts defined by complex, high-dimensional features such as images, text, or sequential data, deep neural networks offer unique advantages for modeling $f(c)$. These architectures can learn adaptive, data-driven representations that capture intricate relationships beyond the scope of classical models. Applications include personalized medicine, natural language processing, and behavioral science, where outcomes may depend on subtle or latent features of the context.

The decision between these machine learning approaches depends on the specific characteristics of the data, the priority placed on interpretability, and computational considerations. Collectively, these advances have significantly broadened the scope of explicit adaptivity, making it feasible to model heterogeneity in ever more complex settings.

Key Theoretical Advances

The expanding landscape of varying-coefficient models (VCMs) has been supported by substantial theoretical progress, which secures the validity of flexible modeling strategies and guides their practical use. The nature of these theoretical results often reflects the core structural assumptions of each model class.

Theory for Smooth Non-parametric Models. For classical VCMs based on kernel smoothing, local polynomial estimation, or penalized splines, extensive theoretical work has characterized their convergence rates and statistical efficiency. Under standard regularity conditions, these estimators are known to achieve minimax optimality for function estimation in moderate dimensions [27]. Recent developments, such as the work of [28], have established asymptotic normality in semi-nonparametric settings, which enables valid confidence interval construction and hypothesis testing even in complex applications.

Theory for Structurally Constrained Models. When discrete or network structure is incorporated into VCMs, theoretical analysis focuses on identifiability, regularization properties, and conditions for consistent estimation. For example, [35] provide non-asymptotic error bounds for estimators in network VCMs, demonstrating that consistency can be attained when the underlying graph topology satisfies certain connectivity properties. In piecewise-constant and partition-based models, results from change-point analysis and total variation regularization guarantee that abrupt parameter changes can be recovered accurately under suitable sparsity and signal strength conditions.

Theory for High-Capacity and Learned Models. The incorporation of machine learning models into VCMs introduces new theoretical challenges. For high-dimensional and sparse settings, oracle inequalities and penalized likelihood theory establish conditions for consistent variable selection and accurate estimation, as seen in methods based on boosting and other regularization techniques [36,37]. In the context of neural network-based VCMs, the theory is still developing, with current research focused on understanding generalization properties and identifiability in non-convex optimization. This remains an active and important frontier for both statistical and machine learning communities.

These theoretical advances provide a rigorous foundation for explicit adaptivity, ensuring that VCMs can be deployed confidently across a wide range of complex and structured modeling scenarios.

Context-Aware Efficiency Principles and Design

The efficiency of context-adaptive methods hinges on several key design principles that balance computational tractability with statistical accuracy. These principles guide the development of methods that can scale to large datasets while maintaining interpretability and robustness.

Context-aware efficiency often relies on sparsity assumptions that limit the number of context-dependent parameters. This can be achieved through group sparsity, which encourages entire groups of context-dependent parameters to be zero simultaneously [38], hierarchical regularization that applies different regularization strengths to different levels of context specificity [39], and adaptive thresholding that dynamically adjusts sparsity levels based on context complexity.

Efficient context-adaptive inference can be achieved through computational strategies that allocate resources based on context. Early stopping terminates optimization early for contexts where convergence is rapid [40], while context-dependent sampling uses different sampling strategies for different contexts [41]. Caching and warm-starting leverage solutions from similar contexts to accelerate optimization, particularly effective when contexts exhibit smooth variation [42].

The design of context-aware methods often involves balancing computational efficiency with interpretability. Linear context functions are more interpretable but may require more parameters, while explicit context encoding improves interpretability but may increase computational cost. Local context modeling provides better interpretability but may be less efficient for large-scale applications. These trade-offs must be carefully considered based on the specific requirements of the application domain, as demonstrated in recent work on adaptive optimization methods [43].

Synthesis and Future Directions

Selecting an appropriate modeling strategy for $f(c)$ involves weighing flexibility, interpretability, computational cost, and the extent of available domain knowledge. Learned function approximators, such as deep neural networks, offer unmatched capacity for modeling complex, high-dimensional relationships. However, classical smooth models and structurally constrained approaches often provide greater interpretability, transparency, and statistical efficiency. The choice of prior assumptions and the scalability of the estimation procedure are also central considerations in applied contexts.

Looking forward, several trends are shaping the field. One important direction is the integration of varying-coefficient models with foundation models from natural language processing and computer vision. By using pre-trained embeddings as context variables c_i, it becomes possible to incorporate large amounts of prior knowledge and extend VCMs to multi-modal and unstructured data sources. Another active area concerns the principled combination of cross-modal contexts, bringing together information from text, images, and structured covariates within a unified VCM framework.

Advances in interpretability and visualization for high-dimensional or black-box coefficient functions are equally important. Developing tools that allow users to understand and trust model outputs is critical for the adoption of VCMs in sensitive areas such as healthcare and policy analysis.

Finally, closing the gap between methodological innovation and practical deployment remains a priority. Although the literature has produced many powerful variants of VCMs, practical adoption is often limited by the availability of software and the clarity of methodological guidance [29]. Continued investment in user-friendly implementations, open-source libraries, and empirical benchmarks will facilitate broader adoption and greater impact.

In summary, explicit adaptivity through structured estimation of $f(c)$ now forms a core paradigm at the interface of statistical modeling and machine learning. Future progress will focus not only on expanding the expressive power of these models, but also on making them more accessible, interpretable, and practically useful in real-world applications.

Implicit Adaptivity: Emergent Contextualization in Complex Models

Introduction: From Explicit to Implicit Adaptivity.

Traditional models often describe how parameters change by directly specifying a function of context, for example through expressions like $\theta_i = f(c_i)$, where the link between context c_i and parameters θ_i is fully explicit. In contrast, many modern machine learning systems adapt in fundamentally different ways. Large neural network architectures—particularly foundation models that are now central to state-of-the-art AI research [44]—show a capacity for adaptation that does not arise from any predefined mapping. Instead, their flexibility emerges naturally from the structure of the model and the breadth of the data seen during training. This phenomenon is known as implicit adaptivity.

Unlike explicit approaches, implicit adaptivity does not depend on directly mapping context to model parameters, nor does it always require context to be formally defined. Such models, by training on large and diverse datasets, internalize broad statistical regularities. As a result, they often display context-sensitive behavior at inference time, even when the notion of context is only implicit or distributed across the input. This capacity for emergent adaptation is especially prominent in foundation models, which can generalize to new tasks and domains without parameter updates, relying solely on the information provided within the input or prompt.

In this section, we offer a systematic review of the mechanisms underlying implicit adaptation. We first discuss the core architectural principles that support context-aware computation in neural networks. Next, we examine how meta-learning frameworks deliberately promote adaptation across diverse tasks. Finally, we focus on the advanced phenomenon of in-context learning in foundation models, which highlights the frontiers of implicit adaptivity in modern machine learning. Through this progression, we aim to clarify the foundations and significance of implicit adaptivity for current and future AI systems.

Foundations of Implicit Adaptation

The capacity for implicit adaptation does not originate from a single mechanism, but reflects a range of capabilities grounded in fundamental principles of neural network design. Unlike approaches that adjust parameters by directly mapping context to coefficients, implicit adaptation emerges from the way information is processed within a model, even when the global parameters remain fixed. To provide a basis for understanding more advanced forms of adaptation, such as in-context learning, this section reviews the architectural components that enable context-aware computation. We begin with simple context-as-input models and then discuss the more dynamic forms of conditioning enabled by attention mechanisms.

Architectural Conditioning via Context Inputs

The simplest form of implicit adaptation appears in neural network models that directly incorporate context as part of their input. In models written as $y_i = g([x_i, c_i]; \Phi)$, context features c_i are concatenated with the primary features x_i, and the mapping g is determined by a single set of fixed global weights Φ. Even though these parameters do not change during inference, the network's nonlinear structure allows it to capture complex interactions. As a result, the relationship between x_i and y_i can vary depending on the specific value of c_i.

This basic yet powerful principle is central to many conditional prediction tasks. For example, personalized recommendation systems often combine a user embedding (as context) with item features to predict ratings. Similarly, in multi-task learning frameworks, shared networks learn representations conditioned on task or environment identifiers, which allows a single model to solve multiple related problems [45].

Interaction Effects and Attention Mechanisms

Modern architectures go beyond simple input concatenation by introducing interaction layers that support richer context dependence. These can include feature-wise multiplications, gating modules, or context-dependent normalization. Among these innovations, the attention mechanism stands out as the foundation of the Transformer architecture [46].

Attention allows a model to assign varying degrees of importance to different parts of an input sequence, depending on the overall context. In the self-attention mechanism, each element in a sequence computes a set of query, key, and value vectors. The model then evaluates the relevance of each element to every other element, and these relevance scores determine a weighted sum of the value vectors. This process enables the model to focus on the most relevant contextual information for each step in computation. The ability to adapt processing dynamically in this way is not dictated by explicit parameter functions, but emerges from the network's internal organization. Such mechanisms make possible the complex forms of adaptation observed in large language models and set the stage for advanced phenomena like in-context learning.

Amortized Inference and Meta-Learning

Moving beyond fixed architectures that implicitly adapt, another family of methods deliberately trains models to become efficient learners. These approaches, broadly termed meta-learning or "learning to learn," distribute the cost of adaptation across a diverse training phase. As a result, models can make rapid, task-specific adjustments during inference. Rather than focusing on solving a single problem, these methods train models to learn the process of problem-solving itself. This perspective provides an important conceptual foundation for understanding the in-context learning capabilities of foundation models.

Amortized Inference

Amortized inference represents a more systematic form of implicit adaptation. In this setting, a model learns a reusable function that enables rapid inference for new data points, effectively distributing the computational cost over the training phase. In traditional Bayesian inference, calculating the posterior distribution for each new data point is computationally demanding. Amortized inference addresses this challenge by training an "inference network" to approximate these calculations. A classic example is the encoder in a Variational Autoencoder (VAE), which is optimized to map high-dimensional observations directly to the parameters, such as mean and variance, of an approximate posterior distribution over a latent space [47]. The inference network thus learns a complex, black-box mapping from the data context to distributional parameters. Once learned, this mapping can be efficiently applied to any new input at test time, providing a fast feed-forward approximation to a traditionally costly inference process.

Meta-Learning: Learning to Learn

Meta-learning builds upon these ideas by training models on a broad distribution of related tasks. The explicit goal is to enable efficient adaptation to new tasks. Instead of optimizing performance for any single task, meta-learning focuses on developing a transferable adaptation strategy or a parameter initialization that supports rapid learning in novel settings [48].

Gradient-based meta-learning frameworks such as Model-Agnostic Meta-Learning (MAML) illustrate this principle. In these frameworks, the model discovers a set of initial parameters that can be quickly adapted to a new task with only a small number of gradient updates [49]. Training proceeds in a nested loop: the inner loop simulates adaptation to individual tasks, while the outer loop updates the initial parameters to improve adaptability across tasks. As a result, the capacity for adaptation becomes encoded in the meta-learned parameters themselves. When confronted with a new task at inference, the model can rapidly achieve strong performance using just a few examples, without the need for a hand-crafted mapping from context to parameters. This stands in clear contrast to explicit approaches, which rely on constructing and estimating a direct mapping from context to model coefficients.

In-Context Learning in Foundation Models

The most powerful and, arguably, most enigmatic form of implicit adaptivity is **in-context learning (ICL)**, an emergent capability of large-scale foundation models. This phenomenon has become a central focus of modern AI research, as it represents a significant shift in how models learn and adapt to new tasks. This section provides an expanded review of ICL, beginning with a description of the core phenomenon, then deconstructing the key factors that influence its performance, reviewing the leading hypotheses for its underlying mechanisms, and concluding with its current limitations and open questions.

The Phenomenon of Few-Shot In-Context Learning

First systematically demonstrated in large language models such as GPT-3 [50], ICL is the ability of a model to perform a new task after being conditioned on just a few examples provided in its input prompt. Critically, this adaptation occurs entirely within a single forward pass, without any updates to the model's weights. For instance, a model can be prompted with a few English-to-French translation pairs and then successfully translate a new word, effectively learning the task on the fly. This capability supports a broad range of applications, including few-shot classification, following complex instructions, and even inducing and applying simple algorithms from examples.

Deconstructing ICL: Key Influencing Factors

The Role of Scale. A critical finding is that ICL is an *emergent ability* that appears only after a model surpasses a certain threshold in scale (in terms of parameters, data, and computation). Recent work has shown that larger models do not just improve quantitatively at ICL; they may also learn in qualitatively different ways, suggesting that scale enables a fundamental shift in capability rather than a simple performance boost [51].

Prompt Engineering and Example Selection. The performance of ICL is highly sensitive to the composition of the prompt. The format, order, and selection of the in-context examples can dramatically affect the model's output. Counterintuitively, research has shown that the distribution of the input examples, rather than the correctness of their labels, often matters more for effective ICL. This suggests that the model is primarily learning a task format or an input-output mapping from the provided examples, rather than learning the underlying concepts from the labels themselves [52].

Hypothesized Mechanisms: How Does ICL Work?

The underlying mechanisms that enable ICL are not fully understood and remain an active area of research. Several leading hypotheses have emerged, viewing ICL through the lenses of meta-learning, Bayesian inference, and specific architectural components.

ICL as Implicit Meta-Learning. The most prominent theory posits that transformers learn to implement general-purpose learning algorithms within their forward pass. During pre-training on vast and diverse datasets, the model is exposed to a multitude of tasks and patterns. This process is thought to implicitly train the model as a meta-learner, allowing it to recognize abstract task structures within a prompt and then execute a learned optimization process on the provided examples to solve the task for a new query [53,54].

ICL as Implicit Bayesian Inference. A complementary and powerful perspective understands ICL as a form of implicit Bayesian inference. In this view, the model learns a broad prior over a large class of functions during its pre-training phase. The in-context examples provided in the prompt act as evidence, which the model uses to perform a Bayesian update, resulting in a posterior predictive distribution for the final query. This framework provides a compelling explanation for how models can generalize from very few examples [55].

The Role of Induction Heads. From a more mechanistic, architectural perspective, researchers have identified specific attention head patterns, dubbed "induction heads," that appear to be crucial for ICL. These specialized heads are hypothesized to form circuits that can scan the context for repeated patterns and then copy or complete them, providing a basic mechanism for pattern completion and generalization from in-context examples [56].

Limitations and Open Questions

Despite its remarkable capabilities, ICL faces significant limitations with respect to transparency, explicit control, and robustness. The adaptation process is opaque, making it difficult to debug or predict failure modes. Furthermore, performance can be brittle and highly sensitive to small changes in the prompt. As summarized in recent surveys, key open questions include developing a more complete theoretical understanding of ICL, improving its reliability, and establishing methods for controlling its behavior in high-stakes applications [57].

Comparative Synthesis: Implicit versus Explicit Adaptivity

Implicit and explicit adaptation strategies represent two fundamentally different philosophies for modeling heterogeneity, each with distinct strengths and limitations. The optimal choice between these approaches depends on the goals of analysis, the structure and scale of available data, and the need for interpretability or regulatory compliance in the application domain.

- **Implicit Adaptivity**: This strategy offers exceptional flexibility and scalability, making it well suited for high-dimensional or unstructured data and efficient at inference. However, the adaptation mechanisms are typically opaque, making it challenging to interpret or control the model's decision process. In applications like healthcare or autonomous systems, this lack of transparency can hinder trust, validation, and responsible deployment.

- **Explicit Adaptivity**: In contrast, explicit models provide direct, interpretable mappings from context to parameters through functions such as $f(c)$. This structure supports clear visualization, statistical analysis, and the formulation of scientific hypotheses. It also enables more direct scrutiny and control of the model's reasoning. Nevertheless, explicit methods rely heavily on domain expertise to specify an appropriate functional form, and may struggle to accommodate unstructured or highly complex context spaces. If the assumed structure is misspecified, the model's performance and generalizability can be severely limited.

In summary, these two paradigms illustrate a fundamental trade-off between expressive capacity and transparent reasoning. Practitioners should carefully weigh these considerations, often choosing or blending approaches based on the unique demands of the task. For clarity, a comparative table or figure can further highlight the strengths and limitations of each strategy across various real-world applications.

Open Challenges and the Motivation for Interpretability

The rise of powerful implicit adaptation methods, particularly in-context learning, raises critical open research questions regarding their diagnosis, control, and reliability. As these models are deployed in increasingly high-stakes applications, understanding their failure modes is not just an academic exercise but a practical necessity [44]. It is important to develop systematic methods for assessing when and why in-context learning is likely to fail, and to create techniques for interpreting and, where possible, steering the adaptation process. While direct control remains elusive, recent prompting techniques like Chain-of-Thought suggest that structuring the context can guide the model's internal reasoning process, offering a limited but important form of behavioral control [58]. A thorough understanding of the theoretical limits and practical capabilities of implicit adaptivity remains a central topic for ongoing research.

These considerations motivate a growing search for techniques that can make the adaptation process more transparent by "making the implicit explicit." Such methods aim to bridge the gap between the powerful but opaque capabilities of implicit models and the need for trustworthy, reliable AI. This research can be broadly categorized into several areas, including post-hoc interpretability approaches that seek to explain individual predictions [59], surrogate modeling where a simpler, interpretable model is trained to mimic the complex model's behavior, and strategies for extracting modular structure from trained models. A prime example of the latter is the line of work probing language models to determine if they have learned factual knowledge in a structured, accessible way [60]. By surfacing the latent structure inside these systems, researchers can enhance trust, promote modularity, and improve the readiness of adaptive models for deployment in real-world settings. This line of work provides a conceptual transition to subsequent sections, which explore the integration of interpretability with adaptive modeling.

Making Implicit Adaptivity Explicit: Local Models, Surrogates and Post Hoc Approximations

Motivation

Building on the prior discussion of implicit adaptivity, we now turn to methods that expose, approximate, or control those adaptive mechanisms.
Implicit adaptivity allows powerful models, including foundation models, to adjust behavior without explicitly representing a mapping from context to parameters [44]. This flexibility hides why and how adaptation occurs, limits modular reuse, and complicates auditing, personalization, and failure diagnosis. Making adaptivity explicit improves alignment with downstream goals, enables modular composition, and supports debugging and error attribution. It also fits the call for a more rigorous science of interpretability with defined objectives and evaluation criteria [61,62].
This chapter reviews practical approaches for surfacing structure, the assumptions they rely on, and how to evaluate their faithfulness and utility.

Approaches

Efforts to make implicit adaptation explicit span complementary strategies that differ in assumptions, granularity, and computational cost. We group them into six families:

1. surrogate modeling for local approximation,
2. prototype- and neighbor-based reasoning,
3. diagnostics for amortized inference,
4. disentanglement and bottleneck methods,
5. parameter extraction and probing, and
6. emerging approaches that leverage large language models as post-hoc explainers.

Surrogate Modeling

This line of work approximates a black-box $h(x, c)$ with an interpretable model in a small neighborhood, so that local behavior and a local view of $f(c)$ can be inspected. A formal template is

$$\hat{g}\_x\_0, c\_0 = \arg\min_{g \in \mathcal{G}} \mathbb{E}\_(x, c) \sim \mathcal{N}_{x_0, c_0} \left[\ell\big(h(x, c), g(x, c)\big) \right] + \Omega(g),$$

where $\mathcal{N}_{x_0, c_0}$ defines a locality (e.g., kernel weights), ℓ measures fidelity, and Ω controls complexity. A convenient local goodness-of-fit is

$$R^2_{\text{local}} = 1 - \frac{\sum_i w_i \left(h_i - g_i\right)^2}{\sum_i w_i \left(h_i - \bar{h}\right)^2}, \qquad w_i \propto \kappa\big((x_i, c_i), (x_0, c_0)\big).$$

LIME perturbs inputs and fits a locality-weighted linear surrogate [63]; SHAP / DeepSHAP provide additive attributions based on Shapley values [64]. Integrated Gradients and DeepLIFT link attribution to path-integrated sensitivity or reference-based contributions [65,66]. These methods are most reliable when the model is near-linear in the chosen neighborhood and perturbations remain near the data manifold; consequently, a rigorous analysis involves stating the neighborhood definition, reporting the surrogate's goodness-of-fit, and assessing stability across seeds and baselines.

Prototype and Nearest-Neighbor Methods

Here, a decision is grounded by reference to similar cases in representation space, which supports case-based explanations and modular updates. ProtoPNet learns a library of visual prototypes to implement "this looks like that" reasoning [67]. Deep k-nearest neighbors audits predictions by querying neighbors in activation space and can flag distribution shift [**PapernotMcDanielGoodfellow2018?**]. Influence functions link a prediction to influential training points for data-centric debugging [68]. This line of work connects naturally to exemplar models and contextual bandits, where decisions are justified via comparisons to context-matched exemplars. Reports should include prototype coverage and diversity, neighbor quality checks, and the effect of editing prototypes or influential examples.

Amortization Diagnostics

For amortized inference systems (e.g., VAEs), the encoder $q_\phi(\theta \mid x)$ can be treated as an implicit $f(c)$. Diagnostics measure amortization gaps and identify suboptimal inference or collapse [69]. Useful outputs include calibration under shift and posterior predictive checks, together with ablations that vary encoder capacity or add limited iterative refinement. This clarifies when the learned mapping is faithful versus when it underfits the target posterior.

Disentangled and Bottlenecked Representations

The aim is to expose factors that align with distinct contextual causes, making changes traceable and controllable. β-VAE encourages more factorized latents [**HigginsMattheyPalAubry2017?**], while the Deep Variational Information Bottleneck promotes predictive minimality that can suppress spurious context [70]. Concept-based methods such as TCAV and ACE map latent directions to human concepts and test sensitivity at the concept level [**KimGhorbaniZaharia2018?**,**GhorbaniWexlerZouKim2019?**]. Fully unsupervised disentanglement is often ill-posed without inductive bias or weak supervision [71]. Reports should include concept validity tests, factor stability across runs, and simple interventions that demonstrate controllability.

Parameter Extraction and Probing

This family locates where adaptation is encoded and exposes handles for inspection or edits. Linear probes test what is linearly decodable from intermediate layers [72]; edge probing examines specific linguistic structure in contextualized representations [73]. Model editing methods such as ROME can modify stored factual associations directly in weights [74], while "knowledge neurons" seek units linked to particular facts [75]. Evaluations should include pre/post-edit behavior, the locality and persistence of edits, and any side effects on unrelated capabilities.

LLMs as Post-hoc Explainers

Recent work uses in-context prompting to elicit rationales, counterfactuals, or error hypotheses from large language models for a target system [76]. These explanations can be useful but must be validated for faithfulness, for example by checking agreement with surrogate attributions, reproducing input–output behavior, and testing stability to prompt variations. Explanations should be treated as statistical estimators with stated objectives and evaluation criteria [62].

Trade-offs

Fidelity vs. Interpretability

High-fidelity surrogates capture the target model's behavior more accurately, yet they often grow in complexity and lose readability. A crisp statement of the design goal is

$$\min \_g \in \mathcal{G} \ \underbrace{\phi\_\text{fid}(g;U)} \_\text{faithfulness on use set } U\backslash+\lambda$$

$$underbrace\psi\_\text{simplicity}(g)\_\text{sparsity / size / semantic load},$$

where ϕ_{fid} can be local R^2, AUC, or rank correlation with h, and $\psi_{\text{simplicity}}$ can be sparsity, tree depth, rule count, or active concept count. If a simple surrogate underfits, consider structured regularization (e.g., monotonic constraints, grouped sparsity, concept bottlenecks). If a complex surrogate is needed, accompany it with readable summaries (partial dependence snapshots, distilled rule sets, compact concept reports).

Local vs. Global Scope

Local surrogates aim for $g_{x_0,c_0} \approx h$ only on $\mathcal{N}_{x_0,c_0}$, whereas a global surrogate seeks $g_{\text{global}} \approx h$ across the domain, potentially smoothing away distinct regimes. Hybrid schemes combine both:

$$g(x,c) = \sum_{k=1}^{K} w_k(x,c)\, g_k(x,c), \qquad \sum_k w_k(x,c) = 1, \quad w_k \geq 0,$$

with local experts g_k and soft assignment w_k. Report the neighborhood definition, coverage (fraction of test cases with acceptable local fit), and disagreements between local and global views; flag regions where the global surrogate is unreliable.

Approximation vs. Control

Coarse modularization makes control and auditing simpler because edits act on a small number of levers, yet residual error can be large. Fine-grained extraction, such as neuron- or weight-level edits, can achieve precise behavioral changes but may introduce unintended side effects. Define the intended edit surface in advance (concepts, features, prototypes, submodules, parameters). For coarse modules, measure the residual gap to the base model and verify that edits improve target behavior without harming unaffected cases. For fine-grained edits, quantify locality and collateral effects using a held-out audit suite with counterfactuals, canary tasks, and out-of-distribution probes. Maintain versioned edits, enable rollback, and document the scope of validity.

Open Questions

Reusable Modules

A central question is whether we can isolate portable skills or routines from large models and reuse them across tasks without degrading overall capability [44]. Concretely, a reusable module should satisfy portability, isolation, composability, and stability. Promising directions include concept bottlenecks that expose human-aligned interfaces, prototype libraries as swappable reference sets, sparse adapters that confine changes to limited parameter subsets, and routing mechanisms that select modules based on context. Evaluation should track transfer performance, sample efficiency, interference on held-out capabilities, and robustness under domain shift.

Performance Gains

When does making structure explicit improve robustness or efficiency compared to purely implicit adaptation? Benefits are most likely when domain priors are reliable, data are scarce, or safety constraints limit free-form behavior. Explicit structure is promising when context topology is known (spatial or graph), when spurious correlations should be suppressed, and when explanations must be auditable. To assess this, fix capacity and training budget and vary only the explicit structure (prototypes, disentanglement, bottlenecks). Stress tests should cover diverse distributional challenges, including covariate shift, concept shift, long-tail classes, and adversarially correlated features. Account for costs such as concept annotation, extra hyperparameters, and potential in-domain accuracy loss.

Abstraction Level

Another open issue is the appropriate level at which to represent structure: parameters (weights, neurons), functions (local surrogates, concept scorers, routing policies), or latent causes (disentangled or causal factors). Choose based on the use case. For safety patches, lower-level handles allow precise edits but require guardrails and monitoring. For scientific or policy communication, function- or concept-level interfaces are often more stable and auditable. Optimize three objectives in tension: faithfulness to the underlying model, usability for the target audience, and stability under shift. Tooling should support movement between levels (e.g., distilling weight-level edits into concept summaries or lifting local surrogates into compact global reports).

Notes on Classical Post-hoc Methods

LIME, SHAP, and gradient-based methods such as Integrated Gradients and DeepLIFT remain common tools for context-adaptive interpretation. Their usefulness depends on careful design and transparent reporting. Explanations should be treated as statistical estimators with stated objectives and evaluation criteria [61,62].

Scope and locality

Local surrogate methods require a clear definition of the neighborhood in which the explanation is valid. The sampling scheme, kernel width, and surrogate capacity determine which aspects of the black box can be recovered. When context variables are present, the explanation should be conditioned on the relevant context and the valid region should be described.

Attribution methods in practice

Attribution based on gradients is sensitive to baseline selection, path construction, input scaling, and preprocessing. Baselines should have clear domain meaning, and sensitivity analyses should show how conclusions change under alternative baselines. For perturbation-based surrogates, report the perturbation distribution and any constraints that keep samples on the data manifold.

Faithfulness and robustness

Faithfulness and robustness should be checked rather than assumed. Useful checks include deletion and insertion curves, counterfactual tests, randomization tests, stability under small input and seed perturbations, and for local surrogates a local goodness-of-fit such as a neighborhood R^2. The evaluation metric should match the stated objective of the explanation [61,62].

Minimal reporting checklist

- Data slice and context used for the explanation, with a description of the locality or neighborhood.
- Surrogate specification, including model family, regularization, and kernel or sampling parameters.
- Faithfulness metrics, for example local R^2, deletion and insertion area, counterfactual validity.
- Sensitivity analyses over baselines, random seeds, and small perturbations, with uncertainty estimates.
- Computational budget and constraints that may affect explanation quality.
- Known limitations and failure modes observed during validation.

From post hoc analysis to design

When the goal is control, auditing, or policy communication, insights from post-hoc analysis can inform the design of explicit context-to-parameter structure. In such cases, use post-hoc findings to specify prototypes, bottlenecks, or concept interfaces that are trained and validated directly, rather than relying only on after-the-fact rationales [61,62]. Taken together, these tools bridge black-box adaptation and structured inference and prepare the ground for designs where context-to-parameter structure is specified and trained end to end.

Implications for classical models

These tools can also clarify how traditional models, for example, logistic regression with interaction terms or generalized additive models to admit a local adaptation view: a simple global form paired with context-sensitive weights or features. Reading such models through the lens of local surrogates and concept interfaces helps align classical estimation with modern, context-adaptive practice.

Context-Invariant Training: A View from the Converse

Most of this review discusses the importance of context in tailoring predictions. The converse view is to ask about the robustness of a model: can we learn features so that one simple predictor works across sites, cohorts, or time—despite shifting environments and nuisance cues? Training for context invariance targets out-of-distribution (OOD) generalization by prioritizing signals whose relationship to the target is stable across environments while down-weighting shortcuts that fluctuate. Standard Empirical Risk Minimization (ERM) [77] often latches onto spurious, environment-specific correlations. In practice, this means using multiple environments during training and favoring representations that make a single readout perform well everywhere.

The first method for invariant prediction with modern deep learning problems and techniques is Invariant Risk Minimization (IRM), which ties robustness to learning invariant (causally stable) predictors across multiple training environments [78]. IRM learns a representation Φ so that the predictor w is simultaneously optimal for every training environment e with respect to the risk $R^e(\cdot)$. The original optimization problem is bi-leveled and hard to solve. To overcome computation difficulty, the author proposes a surrogate model IRMv1, which adds a penalty forcing the per-environment risk to be stationary for a shared “dummy” classifier (gradient at $w = 1$ near zero). This construction connects invariance to out-of-distribution (OOD) generalization by encouraging predictors aligned with causal mechanisms that persist across environments.

However, there are several risks of IRM: in linear models IRM often fails to recover the invariant predictor, and in nonlinear settings IRM can fail catastrophically unless the test data are sufficiently similar to training—undercutting its goal of handling distribution shift (changes in $P(X)$ with $P(Y|X)$ fixed) [24]. Thus, IRM offers no mechanism to reduce sensitivity when those shifts are amplified at test time. To address the covariate shift situation, Risk Extrapolation (REx) allows extrapolation beyond the convex hull and optimize directly over the vector of per-environment risks, with the two instantiations, MM-REx and V-REx. MM-REx performs robust optimization over affine combinations of the environment risks (weights sum to 1, can be negative), while V-REx is a simpler surrogate that minimizes the mean risk plus the variance of risks across environments [79].

Unlike IRM that assumes multiple observed environments and seeks a representation for which the same classifier is optimal in every environment, one can assume that some samples share an identifier. The paper [80] decomposes features into core (whose class-conditional distribution is stable across domains) and style (e.g., brightness, pose) that vary with domain. Under this assumption, the CoRe estimator promotes robustness by penalizing the conditional variance of the prediction or loss within groups with the same class label and identifer (Y, ID).

Adversarial Robustness as Context-Invariant Training

Related references:

- Towards Deep Learning Models Resistant to Adversarial Attacks [81]
- Robustness May Be at Odds with Accuracy [82]

Training methods for Context-Invariant Models

- Just Train Twice: Improving Group Robustness without Training Group Information [83]
- Environment Inference for Invariant Learning [84]
- Distributionally Robust Neural Networks for Group Shifts [25]

Applications, Case Studies, Evaluation Metrics, and Tools

Implementation Across Sectors

TODO: Detailed examination of context-adaptive models in sectors like healthcare and finance.

Relevant references:

- [85]
- [86]

Context-Aware Efficiency in Practice

The principles of context-aware efficiency find practical applications across diverse domains, demonstrating how computational and statistical efficiency can be achieved through intelligent context utilization.

In healthcare applications, context-aware efficiency enables adaptive imaging protocols that adjust scan parameters based on patient context such as age, symptoms, and medical history, reducing unnecessary radiation exposure. Personalized screening schedules optimize screening frequency based on individual risk factors and previous results, while resource allocation systems efficiently distribute limited healthcare resources based on patient acuity and context.

Financial services leverage context-aware efficiency principles in risk assessment by adapting risk models based on market conditions, economic indicators, and individual borrower characteristics. Fraud detection systems use context-dependent thresholds and sampling strategies to balance detection accuracy with computational cost, while portfolio optimization dynamically adjusts rebalancing frequency based on market volatility and transaction costs [87].

Industrial applications benefit from context-aware efficiency through predictive maintenance systems that adapt maintenance schedules based on equipment context including age, usage patterns, and environmental conditions [88]. Quality control implements context-dependent sampling strategies that focus computational resources on high-risk production batches, and inventory management uses context-aware forecasting to optimize stock levels across different product categories and market conditions.

A notable example of context-aware efficiency is adaptive clinical trial design, where trial parameters are dynamically adjusted based on accumulating evidence while maintaining statistical validity. Population enrichment refines patient selection criteria based on early trial results, and dose finding optimizes treatment dosages based on individual patient responses and safety profiles. These applications demonstrate how context-aware efficiency principles can lead to substantial improvements in both computational performance and real-world outcomes.

Performance Evaluation

TODO: Successes, failures, and comparative analyses of context-adaptive models across applications.

Survey of Tools

TODO: Reviewing current technological supports for context-adaptive models.

Selection and Usage Guidance

TODO: Offering practical advice on tool selection and use for optimal outcomes.

Future Trends and Opportunities with Foundation Models

Emerging Technologies

TODO: Identifying upcoming technologies and predicting their impact on context-adaptive learning.

Advances in Methodologies

TODO: Speculating on potential future methodological enhancements.

Expanding Frameworks with Foundation Models

Foundation models refer to large-scale, general-purpose neural networks, predominantly transformer-based architectures, trained on vast datasets using self-supervised learning [89]. These models have significantly transformed modern statistical modeling and machine learning due to their flexibility, adaptability, and strong performance across diverse domains. Notably, large language models (LLMs) such as GPT-4 [90] and LLaMA-3.1 [91] have achieved substantial advancements in natural language processing (NLP), demonstrating proficiency in tasks ranging from text generation and summarization to question-answering and dialogue systems. Beyond NLP, foundation models also excel in multimodal (text-vision) tasks [92], text embedding generation [93], and structured tabular data analysis [94], highlighting their broad applicability.

A key strength of foundation models lies in their capacity to dynamically adapt to different contexts provided by inputs. This adaptability is primarily achieved through techniques such as prompting, which involves designing queries to guide the model's behavior implicitly, allowing task-specific responses without additional fine-tuning [95]. Furthermore, mixture-of-experts (MoE) architectures amplify this contextual adaptability by employing routing mechanisms that select specialized sub-models or "experts" tailored to specific input data, thus optimizing computational efficiency and performance [96].

Foundation Models as Context

Foundation models offer significant opportunities by supplying context-aware information that enhances various stages of statistical modeling and inference:

Feature Extraction and Interpretation: Foundation models transform raw, unstructured data into structured and interpretable representations. For example, targeted prompts enable LLMs to extract insightful features from text, providing meaningful insights and facilitating interpretability [99]. This allows statistical models to operate directly on semantically meaningful features rather than on raw, less interpretable data.

Contextualized Representations for Downstream Modeling: Foundation models produce adaptable embeddings and intermediate representations useful as inputs for downstream models, such as decision trees or linear models [100]. These embeddings significantly enhance the training of both complex, black-box models [101] and simpler statistical methods like n-gram-based analyses [102], thereby broadening the application scope and effectiveness of statistical approaches.

Post-hoc Interpretability: Foundation models support interpretability by generating natural-language explanations for decisions made by complex models. This capability enhances transparency and trust in statistical inference, providing clear insights into how and why certain predictions or decisions are made [103].

Recent innovations underscore the role of foundation models in context-sensitive inference and enhanced interpretability:

FLAN-MoE (Fine-tuned Language Model with Mixture of Experts) [104] combines instruction tuning with expert selection, dynamically activating relevant sub-models based on the context. This method significantly improves performance across diverse NLP tasks, offering superior few-shot and zero-shot capabilities. It also facilitates interpretability through explicit expert activations. Future directions may explore advanced expert-selection techniques and multilingual capabilities.

LMPriors (Pre-Trained Language Models as Task-Specific Priors) [105] leverages semantic insights from pre-trained models like GPT-3 to guide tasks such as causal inference, feature selection, and reinforcement learning. This method markedly enhances decision accuracy and efficiency without requiring extensive supervised datasets. However, it necessitates careful prompt engineering to mitigate biases and ethical concerns.

Mixture of In-Context Experts (MoICE) [105] introduces a dynamic routing mechanism within attention heads, utilizing multiple Rotary Position Embeddings (RoPE) angles to effectively capture token positions in sequences. MoICE significantly enhances performance on long-context sequences and retrieval-augmented generation tasks by ensuring complete contextual coverage. Efficiency is achieved through selective router training, and interpretability is improved by explicitly visualizing attention distributions, providing detailed insights into the model's reasoning process.

Open Problems

Theoretical Challenges

TODO: Critically examining unresolved theoretical issues like identifiability, etc.

Ethical and Regulatory Considerations

TODO: Discussing the ethical landscape and regulatory challenges, with focus on benefits of interpretability and regulatability.

Complexity in Implementation

TODO: Addressing obstacles in practical applications and gathering insights from real-world data.

TODO: Other open problems?

Conclusion

Overview of Insights

TODO: Summarizing the main findings and contributions of this review.

Context-Aware Efficiency: A Unifying Framework

The principles of context-aware efficiency emerge as a unifying theme across the diverse methods surveyed in this review. This framework provides a systematic approach to designing methods that are both computationally tractable and statistically principled.

Several fundamental insights emerge from our analysis. Rather than being a nuisance parameter, context provides information that can be leveraged to improve both statistical and computational efficiency. Methods that adapt their computational strategy based on context often achieve better performance than those that use fixed approaches. The design of context-aware methods requires careful consideration of how to balance computational efficiency with interpretability and regulatory compliance.

Future research in context-aware efficiency should focus on developing methods that can efficiently handle high-dimensional, multimodal context information, creating systems that can adaptively allocate computational resources based on context complexity and urgency, investigating how efficiency principles learned in one domain can be transferred to others, and ensuring that context-aware efficiency methods can be deployed in regulated environments while maintaining interpretability [106].

The development of context-aware efficiency principles has implications beyond statistical modeling. More efficient methods reduce computational costs and environmental impact, enabling sustainable computing practices. Efficient methods also democratize AI by enabling deployment of sophisticated models on resource-constrained devices. Furthermore, context-aware efficiency enables deployment of personalized models in time-critical applications, supporting real-time decision making.

As we move toward an era of increasingly personalized and context-aware statistical inference, the principles outlined in this review provide a foundation for developing methods that are both theoretically sound and practically useful.

Future Directions

TODO: Discussing potential developments and innovations in context-adaptive statistical inference.

References

1. **Varying-Coefficient Models**
Trevor Hastie, Robert Tibshirani
Journal of the Royal Statistical Society Series B: Statistical Methodology (1993-09-01)
https://doi.org/gmfvmb
DOI: 10.1111/j.2517-6161.1993.tb01939.x

2. **Bayesian Edge Regression in Undirected Graphical Models to Characterize Interpatient Heterogeneity in Cancer**
Zeya Wang, Veerabhadran Baladandayuthapani, Ahmed O Kaseb, Hesham M Amin, Manal M Hassan, Wenyi Wang, Jeffrey S Morris
Journal of the American Statistical Association (2022-01-05) https://doi.org/gt68hr
DOI: 10.1080/01621459.2021.2000866 · PMID: 36090952 · PMCID: PMC9454401

3. **Statistical estimation in varying coefficient models**
Jianqing Fan, Wenyang Zhang
The Annals of Statistics (1999-10-01) https://doi.org/dsxd4s
DOI: 10.1214/aos/1017939139

4. **Time-Varying Coefficient Model Estimation Through Radial Basis Functions**
Juan Sosa, Lina Buitrago
arXiv (2021-03-02) https://arxiv.org/abs/2103.00315

5. **Contextual Explanation Networks**
Maruan Al-Shedivat, Avinava Dubey, Eric P Xing
arXiv (2017) https://doi.org/gt68h9
DOI: 10.48550/arxiv.1705.10301

6. **Contextualized Machine Learning**
Benjamin Lengerich, Caleb N Ellington, Andrea Rubbi, Manolis Kellis, Eric P Xing
arXiv (2023) https://doi.org/gt68jg
DOI: 10.48550/arxiv.2310.11340

7. **NOTMAD: Estimating Bayesian Networks with Sample-Specific Structures and Parameters**
Ben Lengerich, Caleb Ellington, Bryon Aragam, Eric P Xing, Manolis Kellis
arXiv (2021) https://doi.org/gt68jc
DOI: 10.48550/arxiv.2111.01104

8. **Contextualized: Heterogeneous Modeling Toolbox**
Caleb N Ellington, Benjamin J Lengerich, Wesley Lo, Aaron Alvarez, Andrea Rubbi, Manolis Kellis, Eric P Xing
Journal of Open Source Software (2024-05-08) https://doi.org/gt68h8
DOI: 10.21105/joss.06469

9. **Contextualized Policy Recovery: Modeling and Interpreting Medical Decisions with Adaptive Imitation Learning**
Jannik Deuschel, Caleb N Ellington, Yingtao Luo, Benjamin J Lengerich, Pascal Friederich, Eric P Xing
arXiv (2023) https://doi.org/gt68jf
DOI: 10.48550/arxiv.2310.07918

10. **Automated interpretable discovery of heterogeneous treatment effectiveness: A COVID-19 case study**
Benjamin J Lengerich, Mark E Nunnally, Yin Aphinyanaphongs, Caleb Ellington, Rich Caruana
Journal of Biomedical Informatics (2022-06) https://doi.org/gt68h5
DOI: 10.1016/j.jbi.2022.104086 · PMID: 35504543 · PMCID: PMC9055753

11. **Discriminative Subtyping of Lung Cancers from Histopathology Images via Contextual Deep Learning**
Benjamin J Lengerich, Maruan Al-Shedivat, Amir Alavi, Jennifer Williams, Sami Labbaki, Eric P Xing
Cold Spring Harbor Laboratory (2020-06-26) https://doi.org/gt68h6
DOI: 10.1101/2020.06.25.20140053

12. **Learning to Estimate Sample-specific Transcriptional Networks for 7000 Tumors**
Caleb N Ellington, Benjamin J Lengerich, Thomas BK Watkins, Jiekun Yang, Abhinav Adduri, Sazan Mahbub, Hanxi Xiao, Manolis Kellis, Eric P Xing
Cold Spring Harbor Laboratory (2023-12-04) https://doi.org/gt68h7
DOI: 10.1101/2023.12.01.569658

13. **Contextual Feature Selection with Conditional Stochastic Gates**
Ram Dyuthi Sristi, Ofir Lindenbaum, Shira Lifshitz, Maria Lavzin, Jackie Schiller, Gal Mishne, Hadas Benisty
arXiv (2023) https://doi.org/gt68jh
DOI: 10.48550/arxiv.2312.14254

14. **Estimating time-varying networks**
Mladen Kolar, Le Song, Amr Ahmed, Eric P Xing
The Annals of Applied Statistics (2010-03-01) https://doi.org/b3rn6q
DOI: 10.1214/09-aoas308

15. **When Personalization Harms: Reconsidering the Use of Group Attributes in Prediction**
Vinith M Suriyakumar, Marzyeh Ghassemi, Berk Ustun
arXiv (2022) https://doi.org/gt68jd
DOI: 10.48550/arxiv.2206.02058

16. **Learning Sample-Specific Models with Low-Rank Personalized Regression**
Benjamin Lengerich, Bryon Aragam, Eric P Xing
arXiv (2019) https://doi.org/gt68jb
DOI: 10.48550/arxiv.1910.06939

17. **Sketch-Based Anomaly Detection in Streaming Graphs**
Siddharth Bhatia, Mohit Wadhwa, Kenji Kawaguchi, Neil Shah, Philip S Yu, Bryan Hooi
arXiv (2023-07-18) https://arxiv.org/abs/2106.04486

18. **Intelligible Models for HealthCare**
Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, Noemie Elhadad
Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2015-08-10) https://doi.org/gftgxk
DOI: 10.1145/2783258.2788613

19. **Adapting multi-armed bandits policies to contextual bandits scenarios**
David Cortes
arXiv (2019-11-26) https://arxiv.org/abs/1811.04383

20. **Environment Inference for Invariant Learning**
Elliot Creager, JÃ¶rn-Henrik Jacobsen, Richard Zemel

arXiv (2021-07-16) https://arxiv.org/abs/2010.07249

21. **Lepski's Method and Adaptive Estimation of Nonlinear Integral Functionals of Density**
Rajarshi Mukherjee, Eric Tchetgen Tchetgen, James Robins
arXiv (2016-01-12) https://arxiv.org/abs/1508.00249

22. **Optimal Rates of Aggregation**
Alexandre B Tsybakov
Lecture Notes in Computer Science (2003) https://doi.org/czntw5
DOI: 10.1007/978-3-540-45167-9_23

23. **Optimal Estimation of Change in a Population of Parameters**
Ramya Korlakai Vinayak, Weihao Kong, Sham M Kakade
arXiv (2019-12-02) https://arxiv.org/abs/1911.12568

24. **The Risks of Invariant Risk Minimization**
Elan Rosenfeld, Pradeep Ravikumar, Andrej Risteski
arXiv (2021-03-30) https://arxiv.org/abs/2010.05761

25. **Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization**
Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, Percy Liang
arXiv (2020-04-03) https://arxiv.org/abs/1911.08731

26. **The Selective Labels Problem**
Himabindu Lakkaraju, Jon Kleinberg, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan
Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2017-08-04) https://doi.org/ggd7hz
DOI: 10.1145/3097983.3098066 · PMID: 29780658 · PMCID: PMC5958915

27. **Varying-coefficient models**
Trevor Hastie, Robert Tibshirani
Journal of the Royal Statistical Society: Series B (Methodological)

28. **Semi-nonparametric Varying Coefficients Models for Imaging Genetics**
Ting Li, Yang Yu, Xiao Wang, JS Marron, Hongtu Zhu
Statistica Sinica
DOI: 10.5705/ss.202024.0118

29. **Publication Trends on the Varying Coefficients Model: Estimating the Actual (Under)Utilization of a Highly Acclaimed Method for Studying Statistical Interactions**
Assaf Botzer
Publications
DOI: 10.3390/publications13020019

30. **Graph-based regularization for regression problems with alignment and highly-correlated designs**
Yuan Li, Benjamin Mark, Garvesh Raskutti, Rebecca Willett, Hyebin Song, David Neiman
arXiv (2019-10-15) https://arxiv.org/abs/1803.07658

31. **Fast Spatio-Temporally Varying Coefficient Modeling With Reluctant Interaction Selection**
Daisuke Murakami, Shinichiro Shirota, Seiji Kajita, Mami Kajita
Geographical Analysis
DOI: 10.1111/gean.70005

32. **Spatially Varying Coefficient Models for Estimating Heterogeneous Mixture Effects**

Jacob Englert, Howard Chang

33. **Varying-Coefficient Panel Models with Spatial Dependence**
    Yiqing Hu, Qingyuan Zhao
    Journal of Econometrics
    DOI: 10.1016/j.jeconom.2024.105883

34. **Urban Economic Modeling with Spatially Varying Coefficients**
    Chongliang Luo, Yihong Du, Peng Zhao
    Regional Science and Urban Economics
    DOI: 10.1016/j.regsciurbeco.2024.104009

35. **Network Varying Coefficient Model**
    Xinyan Fan, Kuangnan Fang, Wei Lan, Chih-Ling Tsai
    Journal of the American Statistical Association
    DOI: 10.1080/01621459.2025.2470481

36. **Boosted Trees for Varying-Coefficient Models**
    Yunfei Wang, Qiang Sun
    Machine Learning
    DOI: 10.1007/s00180-025-01603-8

37. **XGBoost-Inspired Estimation for High-Dimensional Varying Coefficient Models**
    Yu Cheng, Dongdong Yang, Denny Zhou

38. **Model selection and estimation in regression with grouped variables**
    Ming Yuan, Yi Lin
    Journal of the Royal Statistical Society: Series B (Statistical Methodology)

39. **Regression shrinkage and selection via the lasso**
    Robert Tibshirani
    Journal of the Royal Statistical Society: Series B (Methodological)

40. **Optimization methods for large-scale machine learning**
    Léon Bottou, Frank E Curtis, Jorge Nocedal
    SIAM Review

41. **Contextual Bandits with Cross-Learning**
    Santiago Balseiro, Negin Golrezaei, Mohammad Mahdian, Vahab Mirrokni, Jon Schneider
    Advances in Neural Information Processing Systems

42. **Distributed optimization and statistical learning via the alternating direction method of multipliers**
    Stephen Boyd, Neal Parikh, Eric Chu, Brendan Peleato, Jonathan Eckstein
    Foundations and Trends in Machine Learning

43. **Adam: A method for stochastic optimization**
    Diederik P Kingma, Jimmy Ba

44. **On the Opportunities and Risks of Foundation Models**
    Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora
    arXiv preprint arXiv:2108.07258

45. **An Overview of Multi-Task Learning in Deep Neural Networks**
    Sebastian Ruder

46. **Attention Is All You Need**
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, Illia Polosukhin
Advances in Neural Information Processing Systems 30

47. **Auto-Encoding Variational Bayes**
Diederik P Kingma, Max Welling

48. **Meta-Learning in Neural Networks: A Survey**
Timothy Hospedales, Antreas Antoniou, Paul Micaelli, Amos Storkey
IEEE Transactions on Pattern Analysis and Machine Intelligence
DOI: 10.1109/tpami.2021.3079209

49. **Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks**
Chelsea Finn, Pieter Abbeel, Sergey Levine
Proceedings of the 34th International Conference on Machine Learning

50. **Language Models are Few-Shot Learners**
Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, … Dario Amodei
Advances in Neural Information Processing Systems 33

51. **Emergent Abilities of Large Language Models**
Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, … William Fedus
Transactions on Machine Learning Research

52. **Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?**
Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, Luke Zettlemoyer
Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing

53. **Why Can GPT Learn In-Context? Language Models as Meta-Learners**
Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, Furu Wei

54. **Transformers as Support Vector Machines**
Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, Samet Oymak
Proceedings of the 40th International Conference on Machine Learning

55. **An Explanation of In-context Learning as Implicit Bayesian Inference**
Sang Michael Xie, Aditi Raghunathan, Percy Liang, Tengyu Ma

56. **In-context Learning and Induction Heads**
Catherine Olsson

57. **A Survey for In-context Learning**
Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Fei Huang, Xin Li

58. **Chain-of-Thought Prompting Elicits Reasoning in Large Language Models**
Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou
Advances in Neural Information Processing Systems 35

59. **Explainable AI: A Review of Machine Learning Interpretability Methods**
Pantelis Linardatos, Vasilis Papastefanopoulos, Sotiris Kotsiantis

Entropy
DOI: 10.3390/e23010018

60. **Language Models as Knowledge Bases?**
Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick SH Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller
Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing

61. **Towards a Rigorous Science of Interpretable Machine Learning**
Finale Doshi-Velez, Been Kim

62. **Rethinking Explainable Machine Learning as Applied Statistics**
Sebastian Bordt, Eric Raidl, Ulrike von Luxburg

63. **"Why Should I Trust You?": Explaining the Predictions of Any Classifier**
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin
Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining
DOI: 10.1145/2939672.2939778

64. **A Unified Approach to Interpreting Model Predictions**
Scott M Lundberg, Su-In Lee
Advances in Neural Information Processing Systems

65. **Axiomatic Attribution for Deep Networks**
Mukund Sundararajan, Ankur Taly, Qiqi Yan
Proceedings of the 34th International Conference on Machine Learning

66. **Learning Important Features Through Propagating Activation Differences**
Avanti Shrikumar, Peyton Greenside, Anshul Kundaje

67. **This Looks Like That: Deep Learning for Interpretable Image Recognition**
Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Jonathan Su, Cynthia Rudin
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
DOI: 10.1109/cvpr.2019.00914

68. **Understanding Black-box Predictions via Influence Functions**
Pang Wei Koh, Percy Liang
Proceedings of the 34th International Conference on Machine Learning

69. **Inference Suboptimality in Variational Autoencoders**
Chris Cremer, Quaid Morris, David Duvenaud
Proceedings of the 35th International Conference on Machine Learning

70. **Deep Variational Information Bottleneck**
Alexander A Alemi, Ian Fischer, Joshua V Dillon, Kevin Murphy
International Conference on Learning Representations

71. **Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations**
Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, Olivier Bachem
International Conference on Machine Learning

72. **Understanding Intermediate Layers Using Linear Classifier Probes**
Guillaume Alain, Yoshua Bengio

73. **What Do You Learn from Context? Probing for Sentence Structure in Contextualized Word Representations**
Ian Tenney, Patrick Xia, Berlin Chen
Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)

74. **Locating and Editing Factual Associations in GPT**
Kevin Meng, David Bau, Alex Andonian, Yonatan Belinkov
Advances in Neural Information Processing Systems

75. **Knowledge Neurons in Pretrained Transformers**
Damai Dai, Li Dong, Yutao Sun, Shuming Ma, Furu Wei
Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)

76. **In-Context Explainers: Harnessing LLMs for Explaining Black Box Models**
Nicholas Kroeger, Dan Ley, Satyapriya Krishna, Chirag Agarwal, Himabindu Lakkaraju

77. **Principles of Risk Minimization for Learning Theory**
Vladimir N Vapnik
(1991) https://proceedings.neurips.cc/paper/1991/file/ff4d5fbbafdf976cfdc032e3bde78de5-Abstract.html

78. **Invariant Risk Minimization**
Martin Arjovsky, LÃ©on Bottou, Ishaan Gulrajani, David Lopez-Paz
arXiv (2020-03-31) https://arxiv.org/abs/1907.02893

79. **Out-of-Distribution Generalization via Risk Extrapolation (REx)**
David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, Aaron Courville
arXiv (2021-02-26) https://arxiv.org/abs/2003.00688

80. **Conditional Variance Penalties and Domain Shift Robustness**
Christina Heinze-Deml, Nicolai Meinshausen
arXiv (2019-04-16) https://arxiv.org/abs/1710.11469

81. **Towards Deep Learning Models Resistant to Adversarial Attacks**
Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu
arXiv (2019-09-06) https://arxiv.org/abs/1706.06083

82. **Robustness May Be at Odds with Accuracy**
Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, Aleksander Madry
arXiv (2019-09-10) https://arxiv.org/abs/1805.12152

83. **On the Sample Complexity of Adversarial Multi-Source PAC Learning**
Nikola Konstantinov, Elias Frantar, Dan Alistarh, Christoph H Lampert
arXiv (2020-07-01) https://arxiv.org/abs/2002.10384

84. **Conflict-Averse Gradient Descent for Multi-task Learning**
Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, Qiang Liu
arXiv (2024-02-22) https://arxiv.org/abs/2110.14048

85. **Exact Inference for Transformed Large-Scale Varying Coefficient Models with Applications**
Tianyu Chen, Robert Habans, Thomas Douthat, Jenna Losh, Lida Chalangar Jalili Dehkharghani, Li-Hsiang Lin

Journal of Data Science (2025-01-01) https://doi.org/g9t2rs
DOI: 10.6339/25-jds1181

86. **Variable Selection for Generalized Single-Index Varying-Coefficient Models with Applications to Synergistic G × E Interactions**
Shunjie Guan, Xu Liu, Yuehua Cui
Mathematics (2025-01-31) https://doi.org/g9t2rp
DOI: 10.3390/math13030469

87. **Asset Allocation with Regime Shifts and Long-Horizon Risks**
Andrew Ang, Geert Bekaert
Review of Financial Studies

88. **A Model-based method for remaining useful life prediction of machinery**
Yaguo Lei, Naipeng Li, Stanislaw Gontarz, Jing Lin, Slawomir Radkowski, Jacek Dybala
IEEE Transactions on Reliability

89. **On the Opportunities and Risks of Foundation Models**
Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, … Percy Liang
arXiv (2021) https://doi.org/hw3v
DOI: 10.48550/arxiv.2108.07258

90. **GPT-4 Technical Report**
OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, … Barret Zoph
arXiv (2023) https://doi.org/grx4cb
DOI: 10.48550/arxiv.2303.08774

91. **The Llama 3 Herd of Models**
Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, … Zhiyu Ma
arXiv (2024) https://doi.org/ndw6
DOI: 10.48550/arxiv.2407.21783

92. **Learning Transferable Visual Models From Natural Language Supervision**
Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, … Ilya Sutskever
arXiv (2021) https://doi.org/hs7z
DOI: 10.48550/arxiv.2103.00020

93. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**
Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
arXiv (2018) https://doi.org/hm65
DOI: 10.48550/arxiv.1810.04805

94. **TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second**
Noah Hollmann, Samuel Müller, Katharina Eggensperger, Frank Hutter
arXiv (2022) https://doi.org/g9t22b
DOI: 10.48550/arxiv.2207.01848

95. **Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing**
Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, Graham Neubig
ACM Computing Surveys (2023-01-16) https://doi.org/gq5fh2
DOI: 10.1145/3560815

96. **Mixture of experts: a literature survey**
Saeed Masoudnia, Reza Ebrahimpour
Artificial Intelligence Review (2012-05-12) https://doi.org/f59sxs
DOI: 10.1007/s10462-012-9338-y

97. **CHiLL: Zero-shot Custom Interpretable Feature Extraction from Clinical Notes with Large Language Models**
Denis Jered McInerney, Geoffrey Young, Jan-Willem van de Meent, Byron C Wallace
arXiv (2023) https://doi.org/g9t22g
DOI: 10.48550/arxiv.2302.12343

98. **Learning Interpretable Style Embeddings via Prompting LLMs**
Ajay Patel, Delip Rao, Ansh Kothary, Kathleen McKeown, Chris Callison-Burch
arXiv (2023) https://doi.org/g9t22h
DOI: 10.48550/arxiv.2305.12696

99. **Tree Prompting: Efficient Task Adaptation without Fine-Tuning**
Chandan Singh, John Morris, Alexander Rush, Jianfeng Gao, Yuntian Deng
Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (2023) https://doi.org/gtgrkq
DOI: 10.18653/v1/2023.emnlp-main.384

100. **What Can Transformers Learn In-Context? A Case Study of Simple Function Classes**
Shivam Garg, Dimitris Tsipras, Percy Liang, Gregory Valiant
arXiv (2022) https://doi.org/g9t22c
DOI: 10.48550/arxiv.2208.01066

101. **One Embedder, Any Task: Instruction-Finetuned Text Embeddings**
Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, Tao Yu
arXiv (2022) https://doi.org/g9t22f
DOI: 10.48550/arxiv.2212.09741

102. **Augmenting interpretable models with large language models during training**
Chandan Singh, Armin Askari, Rich Caruana, Jianfeng Gao
Nature Communications (2023-11-30) https://doi.org/g9t2z9
DOI: 10.1038/s41467-023-43713-1 · PMID: 38036543 · PMCID: PMC10689442

103. **Explaining Datasets in Words: Statistical Models with Natural Language Parameters**
Ruiqi Zhong, Heng Wang, Dan Klein, Jacob Steinhardt
arXiv (2024) https://doi.org/g9t22k
DOI: 10.48550/arxiv.2409.08466

104. **Mixture-of-Experts Meets Instruction Tuning:A Winning Combination for Large Language Models**
Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, … Denny Zhou
arXiv (2023) https://doi.org/g9t22j
DOI: 10.48550/arxiv.2305.14705

105. **LMPriors: Pre-Trained Language Models as Task-Specific Priors**
Kristy Choi, Chris Cundy, Sanjari Srivastava, Stefano Ermon
arXiv (2022) https://doi.org/g9t22d
DOI: 10.48550/arxiv.2210.12530

106. **Efficient and Effective Query Context-Aware Learning-to-Rank Model for Sequential Recommendation**
Andrii Dzhoha, Alisa Mironenko, Evgeny Labzin, Vladimir Vlasov, Maarten Versteegh, Marjan Celikik