Context-Adaptive Inference: Bridging Statistical and Foundation Models

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

- **Rikhil Rao**
  ⓘ [0009-0003-5221-1125](#) · ◯ [rikhilr](#)
  Department of Computer Science, University of Wisconsin - Madison · Funded by None

- **Jiaqi Wang**
  ⓘ [0009-0003-8531-9490](#) · ◯ [w-jiaqi](#)
  Paul G. Allen School of Computer Science & Engineering, University of Washington · Funded by None

- **Jingyun Jia**
  ⓘ [0009-0006-3241-3485](#) · ◯ [Clouddelta](#)
  Department of Statistics, University of Wisconsin-Madison · 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.

Emerging Theoretical Bridges and Scope of This Review

Recent theoretical work has revealed that the boundary between explicit and implicit approaches to context adaptation may be narrower than it first appears. A growing line of research has shown that transformers trained on regression tasks can implement classical estimators such as ordinary least squares, ridge regression, and even nonlinear kernel regression in context [1,2,3]. In other words, while varying-coefficient models specify a mapping from context to parameters explicitly, in-context learning can implicitly perform the same estimators through internal computation. This emerging

bridge motivates a unified treatment of explicit statistical models and implicit foundation models within a single conceptual framework.

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.

Related Surveys and Reviews

Several surveys have examined specific aspects of context-adaptive inference, but they have largely remained confined to individual methodological traditions. Classical statistical surveys focus on varying-coefficient models and related structured regression methods. In machine learning, surveys on transfer and meta-learning emphasize task adaptation and shared representations, while recent work on foundation models explores the implicit adaptation capabilities of large pretrained models. Table 1 summarizes the scope and coverage of representative surveys.

Survey	Topic Focus	Scope	Coverage of Adaptivity	Gap Relative to This Work
TODO: Add references and details for each survey.	TODO: Add references and details for each survey.	TODO: Add references and details for each survey.	TODO: Add references and details for each survey.	TODO: Add references and details for each survey.

Table 1: Representative surveys and their coverage of context-adaptive inference. Most focus on a single methodological tradition and do not connect explicit and implicit approaches.

While existing surveys have reviewed individual components of this landscape—such as varying-coefficient models, meta-learning, or foundation models—they have remained largely siloed. This article provides the first comprehensive review that unifies explicit and implicit context-adaptive methods under a common framework. By situating classical statistical models, modern machine learning methods, and foundation models along a shared spectrum of context-adaptive inference, we highlight common principles and distinctive challenges.

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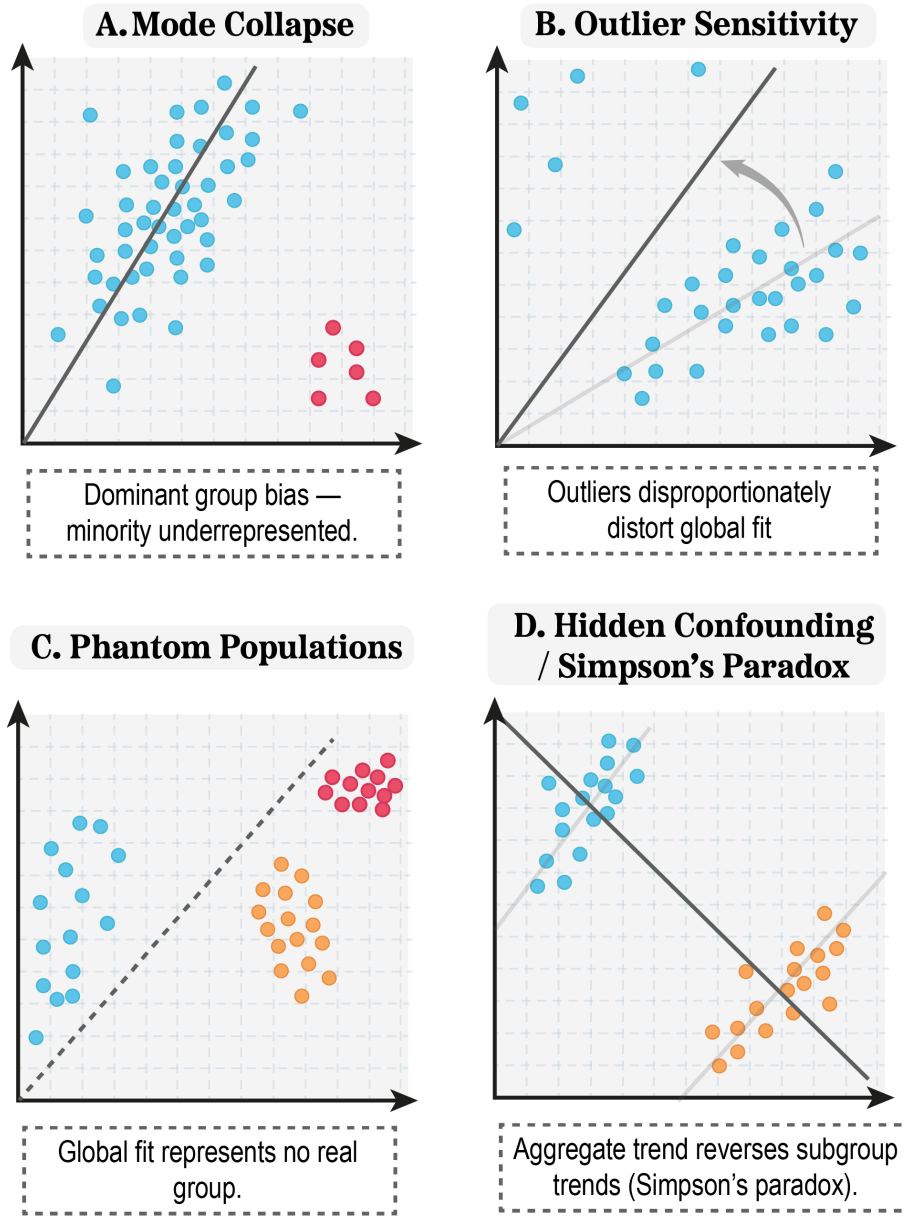
Mode Collapse
When one subpopulation is much larger than another, standard models are biased toward the dominant group, underrepresenting the minority group in both fit and predictions.

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.

Figure 1: Failure Modes of Population Models. Illustrative schematics of common failure types when fitting a single global model to heterogeneous data. (A) **Mode Collapse**: the dominant group drives the fit, underrepresenting the minority. (B) **Outlier Sensitivity**: extreme points distort the global line, shifting predictions away from the majority. (C) **Phantom Populations**: the global fit represents no actual subgroup, but an artificial average. (D) **Hidden Confounding / Simpson's Paradox**: aggregate trends reverse subgroup trends, obscuring true relationships.

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.

A classical example of this challenge arises in causal inference. Following the Neyman–Rubin potential outcomes framework, we let $Y(1)$ and $Y(0)$ denote the outcomes that would be observed under treatment and control, respectively. The average treatment effect (ATE) is then $E[Y(1) - Y(0)]$, or more generally the conditional average treatment effect (CATE) given covariates. Standard approaches often condition only on X, while heterogeneous treatment effect (HTE) models incorporate additional context C to capture systematic variation across subpopulations (Figure 2).
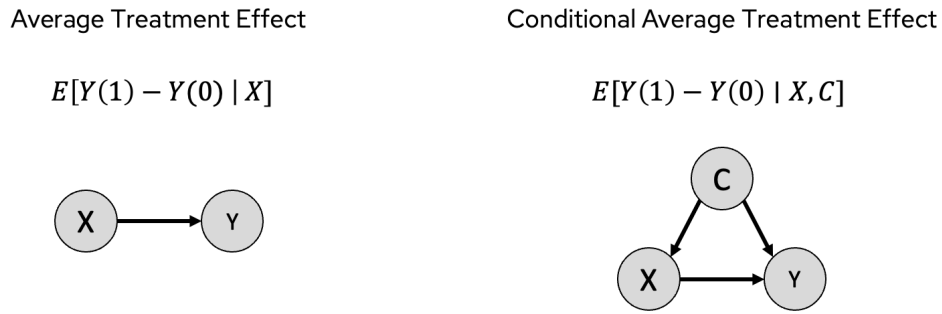
Figure 2: Heterogeneous treatment effects. Left: average treatment effect (ATE) conditional on X, implicitly assuming homogeneity across contexts. Right: conditional average treatment effect (CATE) that allows treatment effects to vary systematically with additional context C.

These models highlight both the promise and the challenges of choosing and estimating $f(c)$.

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 and Semi-parametric Varying-Coefficient Models

Varying-coefficient models (VCMs) provide one of the earliest formal frameworks for explicit adaptivity. Parametric VCMs assume that parameters vary linearly with covariates, a restrictive but interpretable assumption [4]. The estimation can be written as

$$\hat{A} = \arg \max_A \sum_i \ell(x_i; Ac_i).$$

This formulation can be interpreted as a special case of distance-regularized estimation where the distance metric is Euclidean. Related developments in graphical models extend this idea to structured dependencies [5].

Semi-parametric VCMs relax the linearity assumption by requiring only that parameter variation be smooth. This is commonly encoded through kernel weighting, where the relevance of each sample is determined by its similarity in the covariate space [6,7]. These models are more flexible but may fail when the true relationship between covariates and parameters is discontinuous.

Contextualized Models

Contextualized models take a fully non-parametric approach, introduced in [8]. They assume that parameters are functions of context, $f(c)$, but do not restrict the form of f. Instead, f is estimated directly, often with deep neural networks as function approximators:

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

This framework has been widely applied, from machine learning toolboxes [9,10] to personalized genomics [11,12], biomedical informatics [13,14,15], and contextual feature selection [16]. These examples highlight how contextual signals can drive adaptation without assuming a fixed functional form.

Partition and Latent-Structure Models

Partition models extend the contextualized framework by assuming that parameters can be divided into homogeneous groups, while leaving group boundaries to be inferred. This design is useful for capturing abrupt changes over covariates such as time. Estimation typically balances the likelihood with a penalty on parameter differences between adjacent samples, often expressed through a Total Variation (TV) penalty [17]:

$$\{\hat{\theta}_0, \ldots, \hat{\theta}_N\} = \arg \max_{\theta_0, \ldots, \theta_N} \left(\sum_i \ell(x_i; \theta_i) + \lambda \sum_{i=2}^N \|\theta_i - \theta_{i-1}\| \right).$$

By encouraging piecewise-constant structures, partition models get closer to personalized modeling, balancing fit and parsimony.

Fine-tuned Models and Transfer Learning

Another practical strategy for handling heterogeneity is fine-tuning. A global population model is first estimated, and then a smaller set of parameters is updated for particular subpopulations. This idea underlies transfer learning, where large pre-trained models are adapted to new tasks with limited additional training [18]. Fine-tuning balances the bias–variance tradeoff by borrowing statistical strength from large datasets while preserving flexibility for local adaptation. This notion was already recognized in early VCM literature as a form of semi-parametric estimation [6].

Models for Explicit Subgroup Separation

Most adaptive methods encourage parameters for similar contexts to converge, but recent work explores the opposite: ensuring that models for distinct subgroups remain separated. This prevents minority subgroups from collapsing into majority patterns. Such "negative information sharing" is often implemented by learning representations that disentangle subgroup structure, bridging statistical partitioning with adversarial or contrastive learning objectives [19].

A Spectrum of Context-Awareness

Context-aware models can be organized 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**: $\theta_i \sim P(\theta \mid c_i)$, with f learned implicitly.

Each formulation encodes different beliefs about parameter variation. The next section formalizes these principles and examines general strategies for adaptivity in statistical modeling. For a discussion of how subpopulation shifts influence generalization, see [20].

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 beneficial. This section formalizes the core principles that underlie adaptive modeling and situates them within both classical statistics and recent advances in machine learning.

Adaptivity is best understood as a structured set of design choices rather than a single mechanism. Each principle described below highlights a different axis along which models can incorporate or restrict adaptation. Flexibility captures the representational capacity needed for adaptation, while signals of heterogeneity determine when adaptation is justified. Modularity helps organize adaptation into interpretable and transferable units, and selectivity guards against overfitting by controlling when adaptation is triggered. Data efficiency limits how finely we can adapt in practice, and tradeoffs remind us that adaptation is never free of cost. Together, these principles define both the promise and the risks of adaptive systems.

We organize this section into six subsections, each addressing one principle. Afterward, we discuss failure modes and conclude with a synthesis that connects these ideas to practical implications.

1. Adaptivity requires flexibility

The first principle concerns model capacity. A model must be able to represent multiple behaviors if it is to adapt. Without sufficient representational richness, adaptation becomes superficial, amounting only to noise-fitting rather than meaningful personalization. Flexibility provides the foundation that allows a model to express diverse responses across individuals, groups, or environments, rather than enforcing a single global rule.

Flexibility may arise from different modeling strategies. In classical statistics, regression models with interaction effects explicitly capture how predictors influence outcomes differently across contexts, while hierarchical and multilevel models let effects vary systematically across groups. Varying-coefficient models extend this further by allowing regression coefficients to evolve smoothly with contextual covariates [21]. In machine learning, meta-learning and mixture-of-experts architectures [22] offer dynamic allocation of capacity, training models to specialize on tasks or inputs as needed. Together, these approaches illustrate the common principle that without flexibility, adaptation has no meaningful space in which to operate.

2. Adaptivity requires a signal of heterogeneity

Flexibility alone is not enough. A model also requires observable signals that indicate how and why adaptation should occur. Without such signals, adaptive systems risk reacting to random fluctuations rather than capturing meaningful structure. In statistics, varying-coefficient regressions illustrate this idea by allowing parameters to change smoothly with observed covariates [21], while hierarchical models assume systematic group differences that provide a natural signal for adaptive pooling.

In machine learning, contextual bandits adapt decisions to side information that characterizes the current environment, while benchmarks like WILDS highlight that real-world datasets often contain distributional shifts and subgroup heterogeneity [23]. Recent work extends this further, modeling time-varying changes in continuous temporal domain generalization [24] or leveraging diversity across experts to separate stable from unstable patterns [25]. Across applications, from medicine to online platforms, heterogeneity signals provide the essential cues that justify adaptation.

3. Modularity improves adaptivity

Organizing adaptation into modular units improves interpretability and robustness. Instead of spreading changes across an entire system, modularity restricts variation to well-defined subcomponents that can be recombined, reused, or replaced. This structure provides three advantages: targeted adaptation, since adjustments are localized to the relevant parts of a model; transferability, because modules can be carried across tasks or domains; and disentanglement, since modular designs isolate distinct sources of variation.

A canonical example is the mixture-of-experts framework, where a gating network routes inputs to specialized experts trained for different data regimes [22]. By decomposing capacity in this way, models not only gain efficiency but also clarify which components are responsible for specific adaptive behaviors. Recent advances extend this principle in modern architectures: modular domain experts [26], adapter libraries for large language models [27], and mixtures of LoRA experts [28]. In applications ranging from language processing to computer vision, modularity has become a cornerstone of scalable adaptivity.

4. Adaptivity implies selectivity

Adaptation must not occur indiscriminately. Overreacting to noise or small fluctuations often leads to overfitting, which undermines the very purpose of adaptation. Selectivity provides the discipline that ensures adaptive mechanisms respond only when supported by reliable evidence.

Classical statistics formalized this principle through methods such as Lepski's rule for bandwidth selection, which balances bias and variance in nonparametric estimation [29]. Aggregation methods such as the weighted majority algorithm show how selective weighting of multiple models can improve robustness [30]. In modern machine learning, Bayesian rules can activate test-time updates only when uncertainty is manageable [31], while confidence-based strategies prevent unstable adjustments by holding back adaptation under weak signals [32]. Sparse expert models apply the same principle architecturally, activating only a few experts for easy inputs but engaging more capacity for difficult cases [33]. These safeguards demonstrate that good adaptation is selective adaptation.

5. Adaptivity is bounded by data efficiency

Even with flexibility, heterogeneity, modularity, and selectivity in place, the scope of adaptation is fundamentally constrained by the availability of data. Fine-grained adaptation requires sufficient samples to estimate context-specific effects reliably. When data are scarce, adaptive systems risk inflating variance, capturing noise, or overfitting to idiosyncratic patterns. This limitation is not tied to a particular method but reflects a general statistical truth: the ability to adapt cannot exceed the information provided by the data.

Meta-learning research illustrates this tension, as few-shot frameworks show both the promise of cross-task generalization and the sharp degradation that occurs when task diversity or sample size is insufficient [34]. Bayesian analyses of scaling laws for in-context learning formalize how the reliability of adaptation grows with data [35]. To mitigate these limits, modular reuse strategies have been developed, including adapter libraries [27] and modular domain experts. Practical applications, from medicine to recommendation systems, highlight the same lesson: adaptation cannot outpace the data that supports it.

6. Adaptivity is not a free lunch

Adaptivity offers benefits but also introduces costs. It can reduce bias and improve personalization, but at the expense of variance, computational resources, and stability. A model that adapts too readily may become fragile, inconsistent across runs, or difficult to interpret.

In statistical terms, this tension is captured by the classic bias and variance tradeoff [36]: increasing flexibility reduces systematic error but simultaneously increases estimation variance, especially in small-sample settings. Adaptive methods expand flexibility, which means they must also contend with this cost unless constrained by strong regularization or selectivity. In machine learning practice, these tradeoffs surface in multiple ways. Sparse expert models illustrate them clearly: while they scale efficiently, routing instability can cause experts to collapse or remain underused, undermining reliability [37]. Test-time adaptation can boost performance under distribution shift but may destabilize previously well-calibrated predictions. These examples show that adaptation is powerful but never free.

When Adaptivity Fails: Common Failure Modes

The six principles describe when adaptation should succeed, but in practice, failures remain common. Understanding these failure modes is crucial for designing safeguards, as they reveal the vulnerabilities of adaptive methods when principles are ignored or misapplied. Failure does not

necessarily mean that models cannot adapt, but rather that adaptation occurs in ways that are unstable, unjustified, or harmful.

Spurious adaptation. Models sometimes adapt to unstable or confounded features that correlate with outcomes only transiently. This phenomenon is closely related to shortcut learning in deep networks, where spurious correlations masquerade as useful signals [23,38]. Such adaptation may appear effective during training but fails catastrophically under distribution shift. The lesson here is that models must rely on stable signals of heterogeneity, not superficial correlations.

Overfitting in low-data contexts. Fine-grained adaptation requires sufficient signal. When the available data are limited, adaptive models tend to inflate variance and personalize to noise rather than meaningful structure. Meta-learning research illustrates this tension: although few-shot methods aim to generalize with minimal samples, they often degrade sharply when task diversity is low or heterogeneity is weak [34]. This failure mode underscores the principle that data efficiency sets unavoidable limits on adaptivity.

Modularity mis-specification. Although modularity can improve interpretability and transfer, poorly designed modules or unstable routing mechanisms can create new sources of error. Group-shift robustness studies reveal that when partitions are misaligned with true structure, adaptive pooling can worsen disparities across groups [39]. Similarly, analyses of mixture-of-experts models show that mis-specified routing can cause experts to collapse or remain underutilized [37]. These cases highlight that modularity is beneficial only when aligned with meaningful heterogeneity.

Feedback loops. Adaptive models can also alter the very distributions they rely on, especially in high-stakes applications such as recommendation, hiring, or credit scoring. This creates feedback loops where bias is reinforced rather than corrected. For example, an adaptive recommender system that over-personalizes may restrict exposure to diverse content, reshaping user behavior in ways that amplify initial bias. The selective labels problem in algorithmic evaluation illustrates how unobserved counterfactuals complicate learning from adaptively collected data [40]. These examples show that adaptation must be evaluated with attention to long-term interactions, not only short-term accuracy.

Taken together, these failure modes illustrate that adaptivity is double-edged: the same mechanisms that enable personalization and robustness can also entrench bias, waste data efficiency, or destabilize models if not carefully designed and monitored.
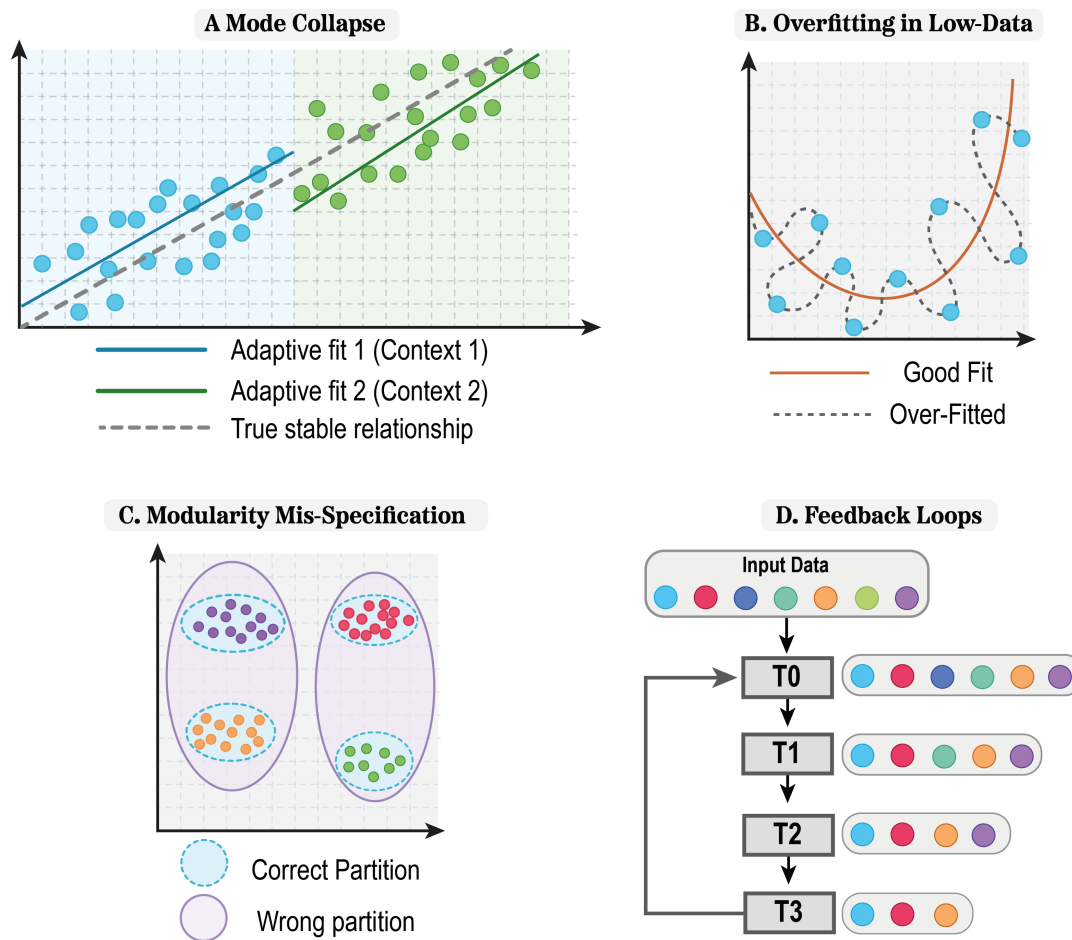
Figure 3: Failure Modes of Context-Adaptive Models. (A) **Mode Collapse**: adaptive fits diverge from the true stable relationship. (B) **Overfitting in Low-Data Contexts**: adaptation follows noise rather than signal. (C) **Modularity Mis-Specification**: incorrect partitions obscure the true structure. (D) **Feedback Loops**: adaptive decisions reshape the very data they rely on.

Synthesis and Implications

The principles and failure modes together provide a coherent framework for context-adaptive inference. Flexibility and heterogeneity define the capacity and justification for adaptation, ensuring that models have room to vary and meaningful signals to guide that variation. Modularity and selectivity organize adaptation into structured, interpretable, and disciplined forms, while data efficiency and tradeoffs impose the practical limits that prevent overreach. Failure modes remind us that these principles are not optional: neglecting them can lead to spurious adaptation, instability, or entrenched bias.

For practitioners, these insights translate into a design recipe. Begin by ensuring sufficient flexibility, but constrain it through modular structures that make adaptation interpretable and transferable. Seek out reliable signals of heterogeneity that justify adaptation, and incorporate explicit mechanisms of selectivity to guard against noise. Respect the limits imposed by data efficiency, recognizing that fine-grained personalization requires sufficient statistical support. Always weigh the tradeoffs explicitly, balancing personalization against stability, efficiency against interpretability, and short-term gains against long-term robustness. Evaluation criteria should extend beyond predictive accuracy to include calibration, fairness across subgroups, stability under distributional shift, and resilience to feedback loops.

By connecting classical statistical models with modern adaptive architectures, this framework provides both a conceptual map and practical guidance. It highlights that context-adaptive inference is

not a single technique but a set of principles that shape how adaptivity should be designed and deployed. When applied responsibly, these principles enable models that are flexible yet disciplined, personalized yet robust, and efficient yet interprepretable. This discussion prepares for the next section, where we turn to explicit adaptive models that operationalize these principles in practice.

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 [41], hierarchical regularization that applies different regularization strengths to different levels of context specificity [42,**tibshirani1996regression?**], 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 [43], while context-dependent sampling uses different sampling strategies for different contexts [44]. Caching and warm-starting leverage solutions from similar contexts to accelerate optimization, particularly effective when contexts exhibit smooth variation [45].

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 [46].

Adaptivity is bounded by data efficiency

Recent work underscores a practical limit: stronger adaptivity demands more informative data per context. When contexts are fine-grained or rapidly shifting, the effective sample size within each context shrinks, and models risk overfitting local noise rather than learning stable, transferable structure. Empirically, few-shot behaviors in foundation models improve with scale yet remain sensitive to prompt composition and example distribution, indicating that data efficiency constraints persist even when capacity is abundant [47,48,49]. Complementary scaling studies quantify how performance depends on data, model size, and compute, implying that adaptive behaviors are ultimately limited by sample budgets per context and compute allocation [35,50,51]. In classical and modern pipelines alike, improving data efficiency hinges on pooling information across related contexts (via smoothness, structural coupling, or amortized inference) while enforcing capacity control and early stopping to avoid brittle, context-specific artifacts [43]. These considerations motivate interpretation methods that report not only attributions but also context-conditional uncertainty and stability, clarifying when adaptive behavior is supported by evidence versus when it reflects data scarcity.

Formalization: data-efficiency constraints on adaptivity

Let contexts take values in a measurable space (), and suppose the per-context parameter is (c). For observation (x,y,c), consider a conditional model $(p_(yx,c))$ with loss $((; x,y,c))$. For a context

neighborhood (*(c) = {c': d(c,c') }*) under metric (*d*), define the effective sample size available to estimate (*(c)*) by [*N*(c,) ,=, {i=1}^n *w*(c_i,c),w_(c_i,c) K!(), *i w*(c_i,c)=1,] where (*K*) is a kernel. A kernel-regularized estimator with smoothness penalty (()=|*c (c)*|^2,c) solves [,=, {}; {i=1}^n (; *x_i,y_i,c_i*) , + , , ().] *Assuming local Lipschitzness in (c) and (L)-smooth, ()-strongly convex risk in (), a standard bias–variance decomposition yields for each component (j) [;; {}; +; {}; +; {},>0,]* which exhibits the adaptivity–data trade-off: finer locality (small ()) increases resolution but reduces (N_), inflating variance. Practical procedures pick () and () to balance these terms (e.g., via validation), and amortized approaches replace ((c)) by (f_(c)) with shared parameters () to increase (N_) through parameter sharing.

For computation, an early-stopped first-order method with step size () and (T(c)) context-dependent iterations satisfies (for smooth, strongly convex risk) the bound [(^{(T(c))}) - (^) ;; (1-)^{(T(c)},(({(0)})-(^)) , + , ,] linking compute allocation (T(c)) and data availability (N_(c,)) to the attainable excess risk at context (c).

Formal optimization view of context-aware efficiency

Let (f_!:!!!) be a context-conditioned predictor with shared parameters (). Given per-context compute budgets (T(c)) and a global regularizer (()), a resource-aware training objective is [{}; {(x,y,c)}, (f_(x,c),y) , + , ,() {c}, (f, T(c), c) B,] where (()) models compute/latency. The Lagrangian relaxation [{}; {(x,y,c)}, (f_(x,c),y) + ,() + , {c}, (f, T(c), c)] trades off accuracy and compute via (). For mixture-of-experts or sparsity-inducing designs, let (=(*1,,M*)) and a gating ((mc)). A compute-aware sparsity penalty is [() ,=, {m=1}^M *m,|m|2^2* , + , , {c}, {m=1}^M (mc),] encouraging few active modules per context. Under smoothness and strong convexity, the optimality conditions yield KKT stationarity [_(,+ ,+ ,_c,) ,=, 0, ,(_c, - B)=0, .] This perspective clarifies that context-aware efficiency arises from jointly selecting representation sharing, per-context compute allocation (T(c)), and sparsity in active submodules subject to resource budgets.

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 [21].

A familiar example of explicit adaptivity is multi-task learning, where context is defined by task identity. Traditional multi-task learning (left) assigns each task its own head on top of shared representations, while context-flagged models (right) pass task identity directly as an input, enabling richer parameter sharing. This illustrates how explicit conditioning on context variables can unify tasks within a single model and provides an intuitive entry point to more general forms of explicit adaptivity (Figure 4).
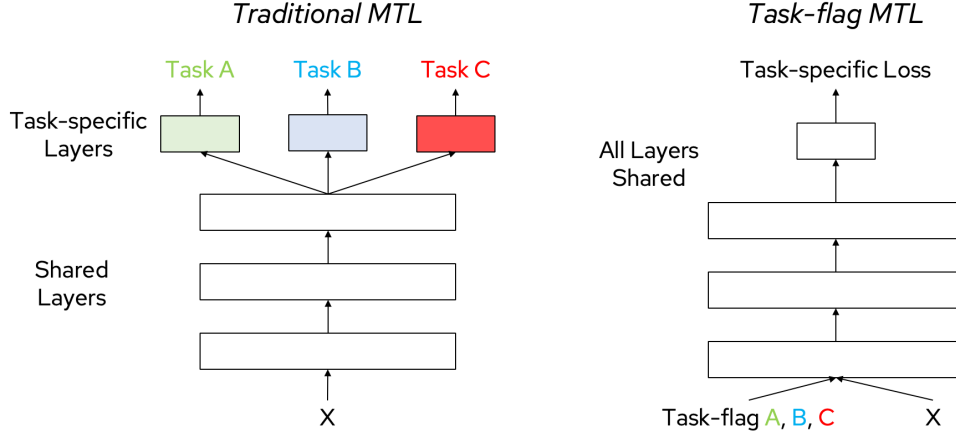
Figure 4: Multi-task learning as explicit adaptivity. In traditional MTL (left), each task has its own head on top of shared layers. In context-flagged models (right), the task identity is provided as an input, enabling a shared model to adapt across tasks.

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 [21,52]. 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.

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 machine learning methods 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, 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.

Structured Regularization for Graphical and Network Models

The origins of structurally constrained models can be traced to early work on covariance selection. Dempster (1972) demonstrated that zeros in the inverse covariance matrix correspond directly to conditional independencies, introducing the principle that sparsity reflects structure [53]. This idea was later consolidated by Lauritzen (1996), whose influential monograph systematized probabilistic graphical models and highlighted how independence assumptions can be embedded into statistical estimation [**dlauritzen1996graphical?**]. Together, these works established the conceptual foundation that explicit structure can guide inference in high-dimensional settings.

As high-dimensional data became common, scalable estimation procedures emerged to make these ideas practical. Meinshausen and Bühlmann (2006) proposed neighborhood selection, recasting graph recovery as a series of sparse regression problems that infer conditional dependencies node by node [54]. Shortly thereafter, Friedman, Hastie, and Tibshirani (2008) developed the graphical lasso, a convex penalized likelihood method that directly estimates sparse precision matrices [55]. These contributions showed that sparsity-inducing penalties could recover large network structures reliably, thereby providing concrete tools for estimating $f(c)$ when context corresponds to a structured dependency pattern such as a graph.

Building on these advances, later research recognized that networks themselves may vary across contexts. Guo, Levina, Michailidis, and Zhu (2011) introduced penalties that jointly estimate multiple graphical models, encouraging sparsity within each network while borrowing strength across related groups [56]. Danaher, Wang, and Witten (2014) extended this framework with the Joint Graphical Lasso, which balances shared structure and context-specific edges across multiple populations [57]. These developments illustrate how structured regularization transforms explicit adaptivity into a principled strategy: instead of estimating networks independently, one can pool information selectively across contexts (where context c is the group or task identity), making the estimation of the parameter function $f(c)$ both interpretable and statistically efficient.

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.

A key design principle is that splits of the context space can mimic what we might otherwise treat as distinct "tasks." By introducing hierarchical partitions, we can capture heterogeneity at multiple levels: sample-level variation within each context, and task-level switching across contexts. This perspective connects classical partition-based models with multi-task learning, highlighting how explicit splits of context define where parameters should be shared versus differentiated (Figure 5).

## Hierarchical Encoding of Context Enables Multi–Level Adaptivity
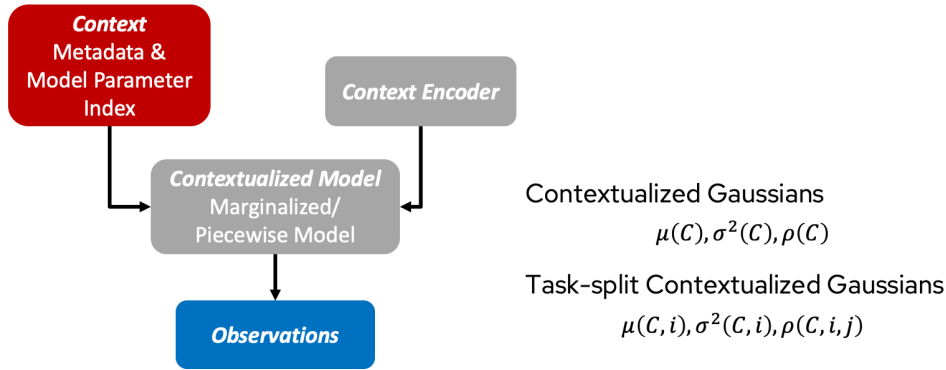
Figure 5: Hierarchical splits of context enable multi-level adaptivity. Explicit adaptivity can partition the context space into piecewise models, with parameters indexed both by context c and task identity (i, j). Such splits allow sample-level heterogeneity to be captured within contexts, while high-level partitions mimic task boundaries and enable task switching.

A subtle but important point is that the boundary between "parametric" and "nonparametric" adaptivity is porous. If we fit **simple parametric models within each context** – for observed contexts c or latent subcontexts Z – and then **aggregate across contexts**, the resulting conditional

$$P(Y \mid X, C) \ = \ \int P(Y \mid X, C, Z) \, dP(Z \mid C)$$

can display rich, multimodal behavior that looks nonparametric. In other words, **global flexibility can emerge from compositional, context-specific parametrics**. When component families are identifiable (or suitably regularized) and the context-to-mixture map is constrained (e.g., smoothness/TV/sparsity over c), the aggregate model remains estimable and interpretable while avoiding overflexible, ill-posed mixtures.

Nonparametric inference from context–adaptive parameters
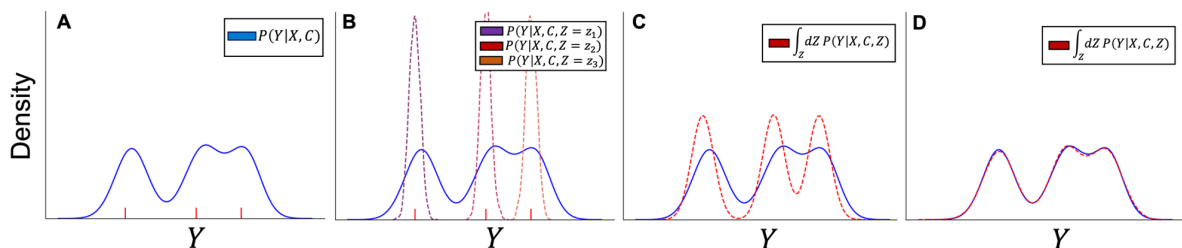
Figure 6: Compositional inference: nonparametric flexibility from parametric context-specific models. (A) Overall conditional $P(Y \mid X, C)$. (B) Context-specific components $P(Y \mid X, C, Z = z_i)$ for latent subgroups Z. (C) Recombination via marginalization $\int_Z P(Y \mid X, C, Z)$. (D) Aggregated distribution showing how structured parametric pieces yield multimodal, nonparametric-like behavior.

This perspective motivates flexible function approximators: trees and neural networks can be read as learning either the **context-to-mixture weights** or **local parametric maps**, providing similar global

flexibility with different inductive biases.

Structured Regularization for Spatial, Graph, and Network Data. When context exhibits known structure, regularization terms can be designed to promote similarity among neighboring coefficients. 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 [58,59]. On networked data, the network VCM of [60] 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.

Beyond spatial and single-network constraints, Bayesian approaches allow explicit modeling of multiple related graphical models across contexts. Rather than estimating each network independently or pooling across all data, these methods place structured priors that encourage information sharing when appropriate. For example, [61] introduced Bayesian inference for GGMs with lattice structure, demonstrating how spatial priors can capture context-dependence across neighboring sites. Building on this idea, [62] proposed a Bayesian framework with a Markov random field prior and spike-and-slab formulation to learn when edges should be shared across sample groups, improving estimation and quantifying inter-context similarity. More recently, [63] extended these principles to covariate-dependent graph learning, where network structure varies smoothly with observed covariates. Their dual group spike-and-slab prior enables multi-level selection at node, covariate, and local levels, providing a flexible and interpretable framework for heterogeneous biological networks. Together, these advances illustrate how Bayesian structural priors make adaptivity explicit in graphical models, supporting both efficient estimation and scientific interpretability.

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 data. A representative example is Tree Boosted Varying-Coefficient Models, introduced by Zhou and Hooker (2019), where GBDTs are applied to estimate context-dependent coefficient functions within a VCM framework [64]. This approach offers a useful balance among flexibility, predictive accuracy, and interpretability, while typically being easier to train and tune than deep neural networks. More recently, Zakrisson and Lindholm (2024) proposed a tree-based varying coefficient model that incorporates cyclic gradient boosting machines (CGBM). Their method enables dimension-wise early stopping and provides feature importance measures, thereby enhancing interpretability and offering additional regularization [65].

Overall, tree-based VCMs achieve strong predictive performance and retain a model structure that lends itself to interpretation, particularly when combined with tools such as SHAP for explaining model outputs.

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 [21]. More specifically, Lu, Zhang, and Zhu (2008) established both consistency and asymptotic normality for penalized spline estimators when using a sufficient number of knots and appropriate penalty terms [66], enabling valid inference through confidence intervals and hypothesis testing. These results provide a solid theoretical foundation even in relatively complex modeling contexts.

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, [60] 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. 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, a wide range of complex and structured modeling scenarios.

Sparsity and Incomplete Measurements as Context

A central practical challenge in combining real-world datasets is inconsistent measurement: different cohorts or institutions often collect different subsets of features. One dataset may contain detailed laboratory values, another may focus on imaging or physiological measurements, and a third may emphasize clinical outcomes. If such cohorts are naively pooled, the resulting feature matrix is sparse and unbalanced. If incomplete samples are discarded, data efficiency collapses.

Context-adaptive models provide a natural resolution by treating **measurement sparsity itself as context.** Rather than ignoring missingness, the model learns to adjust its parameterization according to which features are observed. In effect, each measurement policy (labs-only, vitals-only, multimodal) defines a context, and explicit adaptivity allows estimation that respects these differences while still

sharing information. This perspective reframes missingness from a nuisance into structured signal: it encodes which sources of evidence are available and how they should be combined.
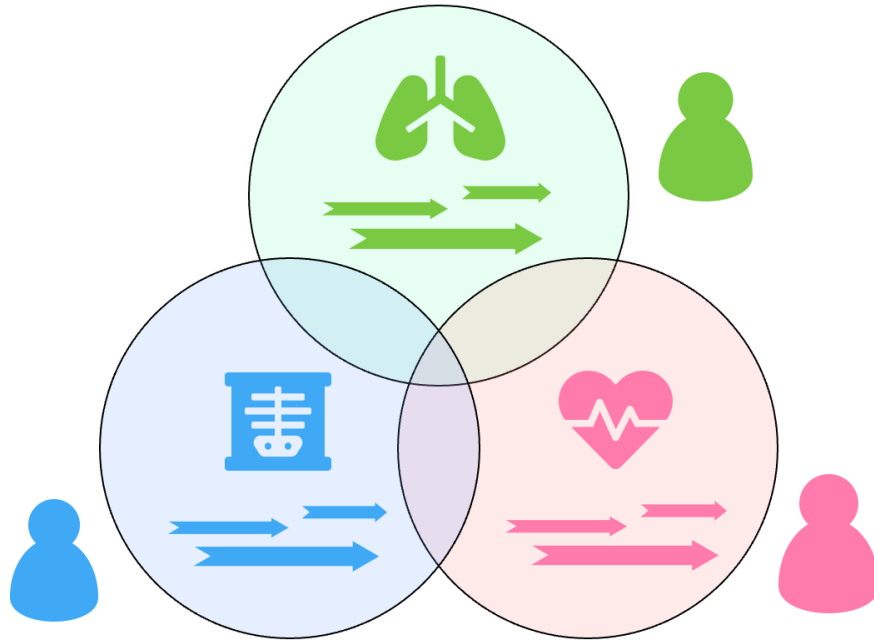
Figure 7: Patterns of missingness as context. Each dataset (e.g., cohort with labs, cohort with vitals, cohort with imaging) provides a different subset of measurements. Context-adaptive models allow integration by conditioning on measurement availability, enabling learning from fewer samples with more heterogeneous features.

Figure 7 illustrates this idea: each cohort contributes a different subset of measurements (lungs, labs, vitals), and explicit adaptivity enables integration across cohorts. By conditioning on measurement availability, we can achieve greater sample efficiency, learning from fewer individuals but with richer heterogeneous features.

Evaluation of missingness-as-context models should report *mask-stratified metrics*, including worst-group performance, following group-robust evaluation practice [67,68]. Robustness should be probed with *mask-shift stress tests*, training under one measurement policy and testing under another, to quantify degradation and the benefit of contextualization, as formalized in the Domain Adaptation under Missingness Shift (DAMS) setting [68,69]. When imputation is used, authors should assess *imputation realism* by holding out observed entries under realistic mask distributions and reporting MAE/RMSE and calibration for $p(x_{\mathrm{missing}} \mid x_{\mathrm{observed}})$ [70,71]. For causal or estimation applications, conduct *ignorability sensitivity analyses*, contrasting MAR-based results with pattern-mixture or selection-model analyses under plausible MNAR mechanisms [72,73]. Finally, include *ablations* that remove mask/indicator inputs—and, for trees, disable default-direction routing—to confirm that gains derive from modeling the mask signal rather than artifacts [74,75]. Practical implementations of these ideas are widely available: **GRU-D** [76] and **BRITS** [77] provide mask- and time-aware sequence models, while **GAIN** [71] and **VAEAC** [70] offer open-source code for imputation under arbitrary masks. For tree ensembles, **XGBoost** supports sparsity-aware default-direction splits, making it straightforward to treat "NA" values as context without preprocessing [74].

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.

One central principle is the use of sparsity assumptions to limit the number of context-dependent parameters. This can be achieved through group sparsity, which encourages entire groups of parameters to be zero simultaneously [41], hierarchical regularization that applies different strengths of shrinkage to varying levels of context specificity [42], and adaptive thresholding that dynamically adjusts sparsity levels in accordance with context complexity.

Efficiency can also be enhanced through computational strategies that allocate resources adaptively. Early stopping terminates optimization for contexts where convergence occurs rapidly [43], while context-dependent sampling employs different sampling schemes across contexts [44]. Caching and warm-starting further accelerate optimization by leveraging solutions from similar contexts, particularly effective when contexts exhibit smooth variation [45].

A further consideration is the balance between efficiency and interpretability. Linear context functions are highly interpretable but may require many parameters, while explicit context encodings improve transparency at the potential cost of higher computational overhead. Local context modeling provides fine-grained interpretability but may be less scalable to large applications. Such trade-offs must be evaluated in light of application needs. For example, advanced adaptive optimizers like Adam can efficiently train complex, nonlinear models, but the resulting systems may be less interpretable than simpler alternatives [46]. In practice, these tensions underscore the ongoing challenge of designing methods that are simultaneously efficient, interpretable, and robust.

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 [52]. 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 [78] They 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 [79].

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 [80].

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 [81]. 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 [82].

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 [83]. 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 [47], 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 effectiveness of ICL is not guaranteed and depends heavily on several interacting factors, which have been the subject of extensive empirical investigation.

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 [48].

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 [49].

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 [84,85].

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 [86].

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 [87].

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 [88].

Theoretical Bridges Between Varying-Coefficient Models and In-Context Learning

Recent theoretical work has uncovered deep connections between classical varying-coefficient models and the mechanisms underlying in-context learning in transformers. Although these approaches arise from different traditions — one grounded in semi-parametric statistics, the other in large-scale deep learning — they can implement strikingly similar estimators. This section formalizes these parallels and reviews key theoretical results establishing these bridges.

Varying-Coefficient Models as Kernel Regression

Consider a semi-parametric varying-coefficient model in which each observation is governed by a parameter vector θ_i that depends smoothly on context c_i. For a new query context c^*, the parameter estimate is obtained by solving a locally weighted likelihood problem:

$$\hat{\theta}(c^*) = \arg\max_\theta \sum_{i=1}^n K_\lambda(c_i, c^*)\, \ell(x_i; \theta),$$

where K_λ is a kernel function that measures similarity between contexts and ℓ is the log-likelihood.

For regression with squared loss, this reduces to kernel ridge regression in the context space. Let $y = (y_1, \ldots, y_n)^\top$ and $K \in \mathbb{R}^{n \times n}$ be the Gram matrix with $K_{ij} = k(c_i, c_j)$. The prediction at c^* is

$$\hat{y}(c^*) = k(c^*)^\top (K + \lambda I)^{-1} y,$$

where $k(c^*) = (k(c^*, c_1), \ldots, k(c^*, c_n))^\top$. This expression highlights that varying-coefficient models perform kernel smoothing in the context space: nearby observations in context have greater influence on the parameter estimates at c^*.

Equivalently, the fitted model can be written as

$$\widehat{f}(x^*, c^*) = \sum_{i=1}^n \alpha_i(c^*)\, y_i,$$

where $\alpha_i(c^*)$ are normalized kernel weights determined entirely by the context similarities and the regularization parameter λ.

Transformers as Ridge and Kernel Regressors In-Context

A parallel line of work has demonstrated that transformers trained on simple regression tasks can learn to perform ridge or kernel regression within their forward pass, without any explicit instruction to do so.

[TODO: What learning algorithm is in-context learning? Investigations with linear models, Akyürek et al., 2022] show that for linear regression tasks, transformers can learn to implement the ridge regression estimator

$$\widehat{w} = (X^\top X + \lambda I)^{-1} X^\top y$$

directly from a sequence of in-context examples. Each example (x_i, y_i) is represented as a token, and the query token attends to the support tokens to compute the prediction for x^*; the attention mechanism learns to encode the solution to the regression problem.

[TODO: Transformers learn in-context by gradient descent, von Oswald et al., 2023] establish convergence results showing that gradient-based training of transformers on a distribution of regression tasks leads them to implement kernel regression in context, with the transformer's learned attention kernel playing the role of $k(c_i, c_j)$. [TODO: What can transformers learn in-context? A case study of simple function classes, Garg et al., 2023] and [TODO: Why can transformers learn in-context? Language models implicitly implement functions, Dai et al., 2023] extend these results, showing that attention layers can approximate kernel smoothers under appropriate scaling limits, and that learned representations can substitute for hand-specified kernels.

Finally, [TODO: Transformers as amortized Bayesian inference engines, Goyal et al., 2025] provide a Bayesian interpretation: transformers pre-trained on large distributions of tasks implicitly learn a prior over functions, and in-context learning corresponds to performing amortized Bayesian inference

with this prior. The in-context examples act as data, and the transformer's forward pass computes an approximate posterior predictive distribution for the query.

In all these cases, the support set within the prompt plays the same role as the neighborhood in context space for varying-coefficient models. The query token's prediction is obtained by aggregating information from the support tokens with learned similarity weights, which are implemented by the attention mechanism rather than an explicit kernel function.

Synthesis: Two Paths to the Same Estimators

Taken together, these results reveal a common form:

$$\widehat{f}(x^*, c^*) = \sum_{i=1}^{n} \alpha_i(c^*) \, y_i,$$

where the weights $\alpha_i(c^*)$ depend on the relationship between the query context c^* and support contexts $\{c_i\}$.

- In varying-coefficient models, $\alpha_i(c^*)$ are determined explicitly by a user-chosen kernel K_λ.
- In transformers, $\alpha_i(c^*)$ emerge implicitly from the learned attention patterns and internal computations after pretraining.

Thus, in-context learning and varying-coefficient modeling are not just philosophically aligned — they can implement the same family of estimators, one through explicit semi-parametric specification, the other through emergent, data-driven computation. This bridge motivates a unified framework for studying context-adaptive inference: explicit methods provide interpretability and structure, while implicit methods provide flexibility and scalability. Understanding how these two meet offers a promising path toward adaptive, interpretable models at scale.

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. The principal advantage of implicit methods lies in their remarkable flexibility and scalability. Leveraging large-scale pre-training on diverse datasets, these models can effectively adapt to high-dimensional and unstructured contexts, such as raw text, images, or other complex sensory data, where explicitly specifying a context function $f(c)$ is infeasible. Because adaptation is performed internally during the model's forward pass, inference is both rapid and adaptable. However, the mechanisms underlying this adaptability 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 [78]. 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 [89]. 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 [90], 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 [91]. 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 [78]. 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 [92,93].
This chapter reviews practical approaches for surfacing structure, the assumptions they rely on, and how to evaluate their faithfulness and utility.

From Implicit to Explicit Adaptivity

Implicit adaptivity is hidden, flexible, and hard to audit, while explicit adaptivity surfaces modular structure that is structured, auditable, and controllable. The transition highlights three key trade-offs developed in this section: **Fidelity vs. Interpretability**, **Local vs. Global Scope**, and **Approximation vs. Control**.
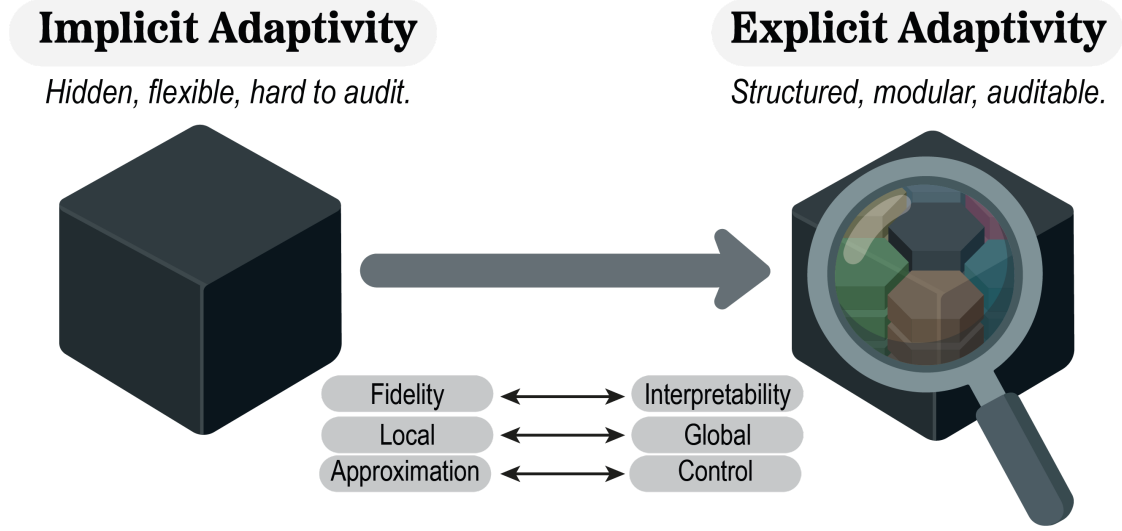
Figure 8: From Implicit to Explicit Adaptivity. A black-box model (left) represents implicit adaptation, which is hidden and opaque. Making adaptivity explicit (right) exposes structured components that can be inspected and controlled. The axes below highlight the trade-offs between fidelity and interpretability, local and global scope, and approximation and control.

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 [94]; SHAP / DeepSHAP provide additive attributions based on Shapley values [95]. Integrated Gradients and DeepLIFT link attribution to path-integrated sensitivity or reference-based contributions [96,97]. 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 [98]. Deep k-nearest neighbors audits predictions by querying neighbors in activation space and can flag distribution shift [99]. Influence functions link a prediction to influential training points for data-centric debugging [100]. 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 [101]. 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 [102], while the Deep Variational Information Bottleneck promotes predictive minimality that can suppress spurious context [103]. Concept-based methods such as TCAV and ACE map latent directions to human concepts and test sensitivity at the concept level [104,105]. Fully unsupervised disentanglement is often ill-posed without inductive bias or weak supervision [106]. 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 [107]; edge probing examines specific linguistic structure in contextualized representations [108]. Model editing methods such as ROME can modify stored factual associations directly in weights [109], while "knowledge neurons" seek units linked to particular facts [110]. 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 [111]. 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 [93].

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

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

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

where ϕ_fid can be local R^2, AUC, or rank correlation with h, and $\psi_\mathrm{simplicity}$ 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 $g_\mathrm{global} \approx h$ 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 [78]. 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 [92,93].

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 [92,93].

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 [92,93]. 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) [112] 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 [113]. 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) [114]. 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 [115].

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 [116] 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

While IRM focuses on learning models that are robust across different environments, adversarial robustness can be viewed as a subclass of the approach. Those different environments can be interpreted as fine-grained, synthetic perturbations around each data point rather than distinct real-world domains. Invariant learning generally seeks predictors whose performance remains stable when the data-generating context changes — for example, across hospitals, time periods, or demographic groups [117]. Adversarial robustness follows the same principle of invariance, but at a much finer scale: instead of using naturally occurring environments, it constructs synthetic "environments" through small, deliberate perturbations of the input data. These perturbations simulate local environmental shifts around each sample and expose the model to worst-case contexts. From this perspective, adversarial robustness is essentially context-invariant learning under infinitesimal, adversarially generated environments. Each adversarial example $x' = x + \delta$ (where $||||_p$) can be interpreted as belonging to a neighboring environment of the original input x. Training the model to perform consistently under such local shifts enforces a form of fine-grained invariance that complements the coarse-grained invariance targeted by IRM. The paper [118] addresses the vulnerability of deep learning models to adversarial attacks from the optimization view. Specifically, the authors interpret adversarial robustness as a min-max optimization problem, where the goal is to minimize the worst-case loss incurred by adversarial examples. The authors introduce the Projected Gradient Descent (PGD) as a universal first-order adversary and these adversarially perturbed samples are subsequently incorporated into the training loop to enhance robustness against such local context shifts. In this view, the environments in IRM correspond to multiple data domains, while those in adversarial training correspond to local neighborhoods of each sample—both formulations share the same objective of minimizing performance variation across shifts in context.

[119] provably demonstrates the trade-off between robustness and accuracy in machine learning models. The authors argue that adversarial training, while improving robustness to adversarial perturbations, can decrease the model's accuracy on clean data. This occurs because adversarial training forces the model to adjust its decision boundaries, which may lead to a loss in standard performance. The paper also shows that the representations learned by robust models align better with salient data characteristics and human perception, which suggests that robust models focus more on features that are meaningful and interpretable. At the same time, robust models tend to learn representations that align better with salient data characteristics and human perception, suggesting that robustness promotes the extraction of stable, semantically meaningful features, mirroring the goal of context-invariant learning at a smaller, instance-specific scale [120].

This perspective is directly applicable to the challenges faced by LLM-based Agents as surveyed in [121]. An autonomous agent does not operate in a sterile, curated dataset; it operates in the wild. The 'fine-grained, synthetic perturbations' explored in adversarial robustness research are a perfect model for the challenges an agent faces:

Perception Robustness: A small, imperceptible change to an image or a document an agent is analyzing (an adversarial perturbation) could cause it to completely misinterpret its environment and take a disastrous action.

Tool-Use Robustness: A slight rephrasing of a user's command could trick a non-robust agent into generating incorrect or malicious code for a tool to execute.

Training methods for Context-Invariant Models

While the principle of context-invariance is a powerful theoretical goal, several practical training methodologies have been developed to approximate it, primarily by enhancing robustness against group shifts. These methods vary in their assumptions, particularly regarding the availability of explicit group or environment labels for the training data.

A foundational approach, applicable when group labels are available, is Group Distributionally Robust Optimization (Group DRO). Unlike standard Empirical Risk Minimization (ERM) which minimizes the average loss over the entire dataset, formulated as:

$$\min_f \frac{1}{n} \sum_{i=1}^{n} L(f(x_i), y_i)$$

Group DRO's objective is to minimize the loss on the worst-performing data group. This is formally expressed as a min-max problem:

$$\min_f \max_{g \in \mathcal{G}} \mathbb{E}_{(x,y) \sim P_g}[L(f(x), y)]$$

where $\mathcal{G}$ represents the set of all predefined groups and P_g is the data distribution for a specific group g [122]. However, the authors identify a critical pitfall: in modern, overparameterized neural networks, this method can fail. Such models can easily memorize the entire training set, reducing the worst-case training loss to zero without actually learning a generalizable solution. The key insight from this work is that **strong regularization** (such as a high L2 penalty or aggressive early stopping) is essential. Regularization prevents the model from perfectly fitting the training data, forcing it to learn simpler, more robust features that generalize better to the worst-case groups on unseen data. The primary limitation of Group DRO is its reliance on fully annotated training data, a luxury seldom available in real-world scenarios. This challenge has spurred the development of methods that operate without explicit group information. These approaches cleverly leverage the inherent biases of standard models as a source of information. A simple and highly effective heuristic is Just Train Twice (JTT) [123]. This method operates in two stages: first, a standard ERM model is trained for several epochs. Second, the training examples that this initial model misclassified are identified and upweighted. A new model is then trained from scratch on this reweighted dataset. The underlying assumption is that a standard model's errors serve as an effective proxy for identifying examples from minority or difficult groups. By focusing the second stage of training on these hard examples, JTT improves worst-group performance without ever needing to know the group labels. Providing a more formalized framework, Environment Inference for Invariant Learning (EIIL) aims to bridge the gap between unlabeled data and invariant learning algorithms like IRM [124]. Similar to JTT, EIIL begins by training a standard ERM model. It then uses the biases of this reference model to automatically partition the dataset into several inferred "environments." For instance, examples the model confidently gets right might form one environment, while those it gets wrong form another. These algorithmically generated environment labels can then be fed into any off-the-shelf invariant learning method to train a final, robust model. EIIL essentially transforms the problem from one requiring

manual labels to one where environments can be discovered directly from the data itself. Together, these methods illustrate a clear progression from fully-supervised techniques to more practical approaches that cleverly infer hidden data structure, all aiming to build models that are more robust and invariant to challenging shifts in context.

Applications, Case Studies, Evaluation Metrics, and Tools

Implementation Across Sectors

Many real-world environments are dynamic and unpredictable, meaning that models built on static assumptions often fail when conditions shift. To remain reliable, models must be able to adapt to changing inputs, contexts, and behaviors. This adaptability is especially important in high-stakes domains where decisions directly affect human well-being or carry significant financial consequences. Two prominent examples are healthcare and finance. In healthcare, context-adaptive models enable more personalized treatment decisions and support early intervention by capturing the evolving state of patients and diseases. In finance, these models capture rapidly changing market conditions, allowing forecasts and risk assessments to remain accurate in volatile times.

Healthcare is one of the domains that benefits greatly from context-aware models because clinical and biomedical data are often hierarchical, exhibiting nested structures and evolving over time. For example, patients may have repeated measurements (e.g., vitals, labs) nested within visits, and these visits are themselves nested within broader care episodes. At the same time, disease trajectories and treatment responses are highly dynamic, requiring models that can adapt to changing contexts rather than assuming static relationships. Several reviews highlight the importance of methods that explicitly account for such complexity in longitudinal and multilevel health data [125,126]. One concrete example is a Bayesian multilevel time-varying joint model that captures complex structures while estimating diverse time-varying relationships, including both response–predictor and response–response dependencies [127]. In this framework, time-varying coefficients are flexibly estimated using Bayesian P-splines, and inference is performed through Markov Chain Monte Carlo (MCMC). The result is a computationally efficient algorithm that provides interpretable modeling of patient outcomes as they evolve over time.

In finance, context-aware models are particularly valuable for capturing the complex dynamics that unfold both over time and across countries, sectors, and assets, which together drive macroeconomic and market behavior. For instance, cross-sectional dependencies, which capture interconnectedness at the same point in time, emerge when shocks propagate differently across regions or industries, while temporal dependencies, which capture persistence across time, arise from persistent volatility clustering and regime changes. Several reviews and comparative studies emphasize the need for methods that can adapt to such heterogeneity in modern financial data [128,129]. A prominent line of work develops Bayesian matrix dynamic factor models (MDFMs), which provide a powerful framework for analyzing matrix-valued time series increasingly common in macro-finance applications [130]. These models incorporate multiple context-adaptive features. On the temporal side, an autoregressive factor process captures persistent comovement and improves recursive forecasting, while stochastic volatility, fat-tailed error distributions, and explicit COVID-19 outlier adjustments allow the model to remain robust under real-world market shocks. This approximate factor framework significantly improves computational efficiency while still capturing complex patterns of dependence, making MDFMs a flexible and scalable approach for modern financial data.

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 [131].

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 [132]. 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.

Formal metrics and evaluation

Let () denote the context space and (*) a test distribution over ((x,y,c)). For a predictor (), define the context-conditional risk* [(c) ,=, [, ((x,c), y) c ,],() ,=, {c_}, (c).] A context-stratified evaluation reports ((c)) across predefined bins or via a smoothed estimate ((c),(c)) for a measure ().

Adaptation efficiency for a procedure that adapts from (k) in-context examples (S_k(c)= {(x_j,y_j,c)}_{j=1}^k) is [_k(c) ,=, (*0c*) , - , ((S_k}c),*k* ,=, {c}, _k(c),] where (*0) is the non-adapted baseline and* ({S_k}) the adapted predictor. The data-efficiency curve (k_k) summarizes few-shot gains.

Transfer across contexts () with representation () can be measured by [() ,=, *{}({})* , - , *{}({})*,] quantifying performance retained by transferring () versus training from scratch. Robustness to context shift (Q) is [(;Q) ,=, *{Q}; ({}() - {}())*,] where (Q) encodes permissible shifts (e.g., f-divergence or Wasserstein balls over context marginals).

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 [131].

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 [132]. 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.

Contextualized Network Inference

One domain where context-adaptive models have shown particular promise is in network inference for genomics. Traditional approaches assume that all samples can be pooled into a single network, or that cohorts can be partitioned into homogeneous groups. These assumptions are often unrealistic: cancer, for example, exhibits both cross-patient heterogeneity and within-patient shifts in gene regulation.

Contextualized network models address this challenge by learning archetypal networks and then representing each sample as a mixture of these archetypes, weighted by its observed context. This formulation allows researchers to move beyond average-case networks and uncover mechanisms of disease, heterogeneity across patients, driver mutations, and structural hazards.
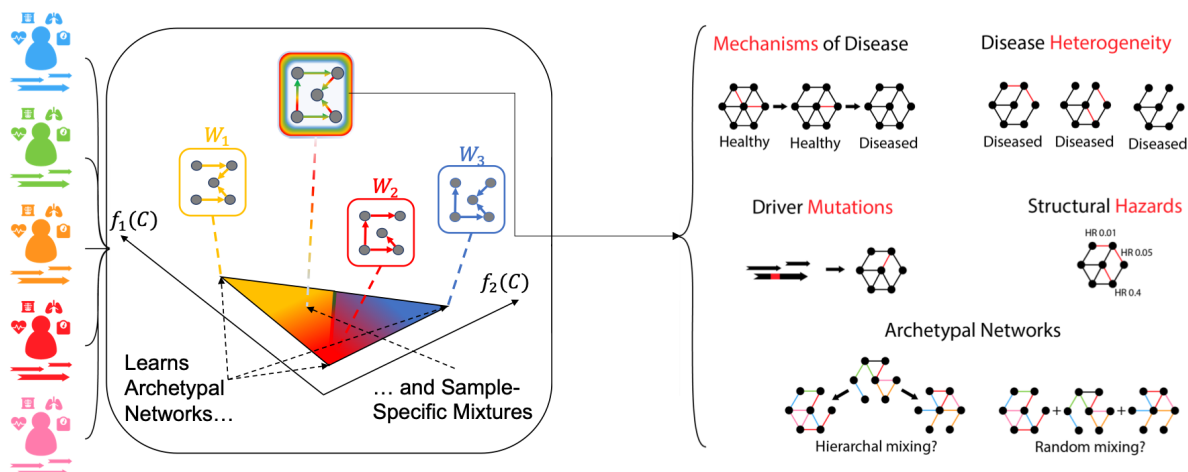
Figure 9: Contextualized networks enable inference of archetypal and sample-specific mixtures, unlocking new biological insights such as mechanisms of disease, disease heterogeneity, structural hazards, and driver mutations.

Performance Evaluation

Evaluating context-adaptive models requires careful consideration of predictive accuracy, robustness to variability, and scalability, with the emphasis varying by domain. Key aspects of performance evaluation include the choice of metrics, the handling of uncertainty, and assessment under stress or rare-event conditions.

In healthcare, evaluation prioritizes patient-specific predictive accuracy and calibrated uncertainty. Common metrics include mean squared error (MSE), concordance indices (C-index), and calibration curves, which measure how well models capture longitudinal patient trajectories and provide reliable uncertainty estimates. Multi-target Bayesian approaches and survival models demonstrate the importance of capturing correlations across outcomes and assessing credible interval coverage to quantify predictive confidence [133,134]. Evaluations in this domain also highlight trade-offs between model complexity, interpretability, and computational feasibility, since high-fidelity patient-level predictions can be costly to compute.

In finance and macro forecasting, performance evaluation emphasizes predictive accuracy under volatile conditions and resilience to structural breaks. Metrics such as root mean squared forecast error (RMSFE), log-likelihood, and stress-test performance are commonly used to assess how well models handle crises or abrupt shifts in data [130,135]. Probabilistic metrics, including posterior predictive checks and uncertainty bounds, provide additional insight into the reliability of forecasts, while chaos-informed diagnostics can highlight vulnerabilities to extreme events [136].

Across domains, consistent patterns emerge. Context-adaptive models outperform static baselines when variability is structured and partially predictable, but performance can degrade in data-sparse regimes or under unmodeled abrupt changes [137]. Evaluations therefore combine error-based measures, probabilistic calibration, and robustness tests to give a holistic view of model performance. The focus should be on these evaluation criteria, rather than the models themselves, to understand where and why context-adaptive approaches provide real advantages.

Survey of Tools

There are many technological supports that have emerged to support context-adaptive modeling. These tools provide the infrastructure, memory, and efficiency mechanisms that allow models to operate effectively in dynamic environments.

Retrieval-augmented generation (RAG) has become a core support for adaptivity, enabling models to incorporate new knowledge at inference time instead of relying only on static parameters. Recent surveys outline how RAG architectures combine dense retrievers, re-rankers, and generators into pipelines that continuously update with external information. This allows models to remain aligned with changing knowledge bases [138]. Beyond improving factuality, RAG also underpins adaptive behavior in AI-generated content, where external retrieval reduces hallucination and provides domain-specific grounding [139]. These systems depend on efficient vector search. Tools such as FAISS use approximate nearest neighbor algorithms to index billions of embeddings with low latency, while Milvus integrates distributed storage to scale such systems across production environments [140]. Together, retrieval pipelines and vector databases constitute the infrastructure through which context-adaptive models dynamically expand their accessible knowledge.

While retrieval addresses external knowledge, memory systems support continuity within ongoing interactions. Research on AI memory frameworks emphasizes how models require mechanisms to persist relevant context, get rid of redundancy, and resurface information at appropriate times [141]. Recent implementations such as MemoryOS illustrate how adaptive memory systems can summarize

past conversations, cluster related items, and strategically reinsert them into prompts, producing long-term coherence that can't be achieved with static context windows alone [142]. These memory architectures extend adaptivity from the level of just accessing facts to maintaining evolving histories, allowing models to not just adjust to new data, but also to be more consistent and contextually aware of their interactions.

Another critical support lies in scaling sequence length. Standard transformers suffer quadratic complexity and degraded performance as contexts grow, making it difficult to adapt to long or streaming data. New serving infrastructures such as StreamingLLM introduce rolling caches that let models handle long inputs without full recomputation, while frameworks like vLLM use paged attention to manage GPU memory efficiently during extended inference [143,144]. This long-context support shifts adaptability from handling snapshots of information to maintaining awareness across evolving information streams.

Selection and Usage Guidance

Deploying context-adaptive models effectively requires careful alignment between model capabilities, domain needs, and practical constraints.

In healthcare, where data is often hierarchical and time-varying, Bayesian multilevel models and generalized varying-coefficient frameworks are well suited because they can flexibly capture nonlinear interactions and evolving patient trajectories. In finance, high-dimensional time series demand scalability, making matrix dynamic factor models more appropriate than fully specified multivariate systems.

Domain priorities should drive tool choice. Clinical applications often require interpretable models that clinicians can trust, favoring spline-based or single-index approaches even if they sacrifice some predictive accuracy. In contrast, finance applications typically prioritize forecasting performance under volatility, where more complex factor models can offer a competitive edge despite reduced transparency.

Many context-adaptive models rely on resource-intensive inference methods such as MCMC, which may limit scalability. Approximate inference techniques like variational Bayes or stochastic optimization can mitigate this burden for large datasets. In real-time decision settings, long-context processing methods such as StreamingLLM or KV-cache compression provide efficiency gains but require specialized engineering and hardware support.

Finally, tool selection should reflect whether the primary objective is scientific insight or operational decision-making. Biomedical research benefits most from flexible, interpretable models that generate new hypotheses, whereas domains like trading demand models capable of rapid adaptation, scalable inference, and strong predictive accuracy under uncertainty.

There is no one-size-fits-all context-adaptive model. Successful deployment depends on tailoring tool choice to data structure, interpretability needs, computational constraints, and domain-specific goals.

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 [145]. 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 [146] and LLaMA-3.1 [147] 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 [148], text embedding generation [149], and structured tabular data analysis [150], 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 [151]. 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 [152].

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 [155]. 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 [156]. These embeddings significantly enhance the training of both complex, black-box models [157] and simpler statistical methods like n-gram-based analyses [158], 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 [159].

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) [160] 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) [161] 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) [161] 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

The rapid development of context-adaptive modeling has created both opportunities and challenges. This chapter outlines the most pressing open research questions and broader challenges that will shape the future of the field. We first detail five key technical research questions, covering modularity, the benefits of explicit structure, levels of abstraction, and remaining barriers. We then turn to the broader outlook, focusing on the ethical and societal implications of deploying these powerful adaptive systems.

Open Research Questions

Recent advances have broadened the scope of adaptive inference, but many questions remain unresolved. These questions cover several areas: modularity, the theoretical and practical benefits of explicit structure, the appropriate level of abstraction for intervention, barriers that limit adoption, and the balance between interpretable-by-design and post-hoc interpretability. Together, they define a research agenda that is both methodologically rich and practically significant.

First, researchers need to examine whether skills and routines can be modularized in a way that allows portability across tasks without interference. Second, the field must clarify under what conditions explicit structure provides measurable benefits. Third, it remains unclear at which level of abstraction such structure should be imposed, whether at the level of parameters, functions, or latent factors. Fourth, adoption is limited by both theoretical and practical barriers, including identifiability, generalization, and computational feasibility. Finally, the community must address the tension between building models that are interpretable from the start and those that rely on post-hoc explanations. The following subsections provide a more detailed discussion of these five questions.

Can Reusable Modules Enable Portability Across Tasks?

A central question is whether the skills or routines acquired by large models can be isolated and reused as portable modules across tasks without reducing overall performance [145]. The vision of modularity is to build an ecosystem of specialized components that can be composed when needed, instead of training a new large model for each task. Potential approaches include the study of concept bottlenecks that force models to represent information in human-understandable concepts, prototype libraries that act as case-based memory, and sparse adapters and routing mechanisms that selectively activate components according to context [162,163].

Applications illustrate the promise of this research. In healthcare, diagnostic modules could be reused across diseases. In natural language processing, syntax-aware modules might be applied across languages. However, modularity also introduces risks. Interactions between modules may cause interference, and poorly aligned modules may amplify biases. Future work should therefore design evaluation protocols that test not only portability and composability, but also isolation of unintended side effects and robustness to distribution shift [147].

What Are the Theoretical and Practical Benefits of Explicit Structure?

Clarifying the theoretical and practical benefits of explicit structure is an important open question. Implicit adaptation is highly flexible, but explicit structure may provide stronger guarantees of robustness and generalization under distribution shift. Practical benefits include greater interpretability, improved debugging, and the ability to incorporate domain knowledge directly.

To advance this agenda, systematic comparisons with implicit approaches are needed. Stress testing under covariate shift, concept drift, long-tail distributions, and adversarial correlations is particularly important, and benchmarks such as WILDS provide a useful starting point [164]. At the same time, researchers must weigh the costs of explicit structure. These costs include additional annotation, increased hyperparameter complexity, and potential reductions in in-domain accuracy [166]. A comprehensive evaluation framework that quantifies both theoretical guarantees and practical trade-offs remains to be established.

At What Level of Abstraction Should Explicit Structure Be Imposed?

Determining the appropriate level of abstraction for intervention remains a challenge. Parameter-level edits provide precise control but are brittle and can have unpredictable side effects [167]. Concept-level interventions provide stability and interpretability but may fail to capture the model's internal computations in full detail [168].

Intermediate levels may offer a balance. For example, function-level interventions or local surrogate models can capture mid-level abstractions that combine precision with stability. More importantly, future work should aim to develop methods that allow translation across levels. For instance, low-level parameter edits could be distilled into high-level conceptual summaries, while abstract concepts might be operationalized through concrete parameter changes. Such tools would make adaptive models more interpretable and more controllable in practice.

What Theoretical and Practical Barriers Remain?

Several barriers continue to limit the adoption of adaptive models. On the theoretical side, researchers have yet to establish strong guarantees for identifiability and generalization under distribution shift [166]. Extending these guarantees to high-dimensional and multimodal data remains an unsolved challenge.

Practical barriers are equally important. Training and deploying adaptive models requires significant computational and memory resources. Data limitations, such as biased sampling and noisy feedback, reduce reliability. Evaluation frameworks remain centered on accuracy, with insufficient attention to fairness, stability, and long-term robustness. Finally, the absence of standardized tools and implementation guidelines prevents many practitioners from applying state-of-the-art methods beyond research settings [169].

Interpretable-by-Design vs Post-hoc Interpretability: What Is the Right Path Forward?

A final open question concerns the balance between interpretable-by-design approaches and post-hoc interpretability. Interpretable-by-design models, such as varying coefficient models, provide transparency and faithfulness from the outset but may restrict predictive performance [4]. Post-hoc methods allow powerful foundation models to be explained after training, but explanations may be incomplete or unfaithful to the model's internal reasoning [145].

Progress in both directions suggests that the future lies in integration rather than a binary choice. Hybrid models may embed interpretable structures at their core while using post-hoc tools for flexibility. Promising directions include benchmarks that jointly evaluate adaptivity and interpretability, as well as human-in-the-loop workflows that allow domain experts to constrain and validate model adaptation in practice.

Broader Challenges and Future Outlook

While the previous section focused on research questions that can be addressed by new methods, theory, and experiments, broader challenges remain that extend beyond purely technical considerations. These challenges concern the responsible deployment of adaptive models in real-world environments, where issues such as ethics, fairness, and regulatory compliance play a critical role. Adaptive systems used in sensitive domains like healthcare and finance must satisfy principles of interpretability, auditability, and accountability in order to prevent harm and build public trust [145]. Regulators and practitioners will need to collaborate to create transparent standards so that adaptive decisions can be evaluated and explained.

Another set of challenges arises from the dynamic interaction between adaptive models and their environments. Feedback loops may amplify small initial biases, leading to systematic disadvantages for certain groups over time. Examples can be seen in credit scoring, hiring, and online recommendation systems, where early decisions influence future data collection and can entrench inequalities [170]. Addressing these risks requires methods that anticipate long-term effects, including simulation studies, formal analyses of dynamic systems, and model designs that incorporate fairness constraints directly during learning.

Looking ahead, the long-term vision for adaptive modeling is to develop systems that are not only powerful but also trustworthy. Progress requires moving beyond accuracy as the dominant evaluation criterion to include fairness, stability, and transparency. Human oversight should be an integral part of adaptive pipelines, enabling experts to guide and validate model behavior in practice. Sustainability is another important dimension, as the computational and environmental costs of adaptive models continue to grow. By combining technical innovation with responsible deployment, the field can ensure that adaptive inference contributes to both scientific progress and societal benefit.

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 [171].

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. **Transformers learn in-context by gradient descent**
   Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, Max Vladymyrov
   arXiv (2023-06-01) https://arxiv.org/abs/2212.07677

2. **What Can Transformers Learn In-Context? A Case Study of Simple Function Classes**
   Shivam Garg, Dimitris Tsipras, Percy Liang, Gregory Valiant
   arXiv (2023-08-15) https://arxiv.org/abs/2208.01066

3. **Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers**
   Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, Furu Wei
   arXiv (2023-05-16) https://arxiv.org/abs/2212.10559

4. **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

5. **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

6. **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

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

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

9. **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

10. **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

11. **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

12. **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

13. **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

14. **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

15. **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

16. **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

17. **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

18. **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

19. **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

20. **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

21. **Varying-Coefficient Models**
    Trevor J Hastie, Robert Tibshirani
    Journal of the Royal Statistical Society. Series B (Methodological) (1993)
    https://api.semanticscholar.org/CorpusID:125170071

22. **Adaptive Mixtures of Local Experts**
    Robert A Jacobs, Michael I Jordan, Steven J Nowlan, Geoffrey E Hinton
    Neural Computation (1991) https://api.semanticscholar.org/CorpusID:572361

23. **WILDS: A Benchmark of in-the-Wild Distribution Shifts**
    Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay
    Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, ... Percy Liang
    International Conference on Machine Learning (2020)
    https://api.semanticscholar.org/CorpusID:229156320

24. **Continuous Temporal Domain Generalization**
    Zekun Cai, Guangji Bai, Renhe Jiang, Xuan Song, Liang Zhao
    ArXiv (2024) https://api.semanticscholar.org/CorpusID:270063125

25. **LFME: A Simple Framework for Learning from Multiple Experts in Domain Generalization**
    Liang Chen, Yong Zhang, Yibing Song, Zhiqiang Shen, Lingqiao Liu
    ArXiv (2024) https://api.semanticscholar.org/CorpusID:273507138

26. **Scalable Multi-Domain Adaptation of Language Models using Modular Experts**
    Peter Schafhalter, Shun Liao, Yanqi Zhou, Chih-Kuan Yeh, Arun Kandoor, James Laudon
    ArXiv (2024) https://api.semanticscholar.org/CorpusID:273346210

27. **Towards Modular LLMs by Building and Reusing a Library of LoRAs**
    Oleksiy Ostapenko, Zhan Su, E Ponti, Laurent Charlin, Nicolas Le Roux, Matheus Pereira, Lucas
    Caccia, Alessandro Sordoni
    ArXiv (2024) https://api.semanticscholar.org/CorpusID:269922018

28. **Mixture of LoRA Experts**
    Xun Wu, Shaohan Huang, Furu Wei
    ArXiv (2024) https://api.semanticscholar.org/CorpusID:269293160

29. **Optimal pointwise adaptive methods in nonparametric estimation**
    Oleg V Lepski, Vladimir G Spokoiny
    Annals of Statistics (1997) https://api.semanticscholar.org/CorpusID:2635430

30. **The weighted majority algorithm**
    Nick Littlestone, Manfred K Warmuth
    30th Annual Symposium on Foundations of Computer Science (1989)
    https://api.semanticscholar.org/CorpusID:12843330

31. **Selective Test-Time Adaptation for Unsupervised Anomaly Detection using Neural
    Implicit Representations**
    Sameer Ambekar, Julia A Schnabel, Cosmin I Bercea
    ArXiv (2024) https://api.semanticscholar.org/CorpusID:273163004

32. **Test-Time Adaptation Induces Stronger Accuracy and Agreement-on-the-Line**
    Eungyeup Kim, Mingjie Sun, Christina Baek, Aditi Raghunathan, JZico Kolter
    Neural Information Processing Systems (2023)
    https://api.semanticscholar.org/CorpusID:263830662

33. **Harder Task Needs More Experts: Dynamic Routing in MoE Models**
Quzhe Huang, Zhenwei An, Zhuang Nan, Mingxu Tao, Chen Zhang, Yang Jin, Kun Xu, Liwei Chen, Songfang Huang, Yansong Feng
Annual Meeting of the Association for Computational Linguistics (2024)
https://api.semanticscholar.org/CorpusID:271895204

34. **Unsupervised Learning via Meta-Learning**
Kyle Hsu, Sergey Levine, Chelsea Finn
ArXiv (2018) https://api.semanticscholar.org/CorpusID:52922125

35. **Bayesian scaling laws for in-context learning**
Aryaman Arora, Daniel Jurafsky, Christopher Potts, Noah D Goodman
ArXiv (2024) https://api.semanticscholar.org/CorpusID:273507537

36. **An overview of statistical learning theory**
Vladimir Naumovich Vapnik
IEEE Transactions on Neural Networks (1999)
https://api.semanticscholar.org/CorpusID:6294728

37. **A Closer Look into Mixture-of-Experts in Large Language Models**
Ka Man Lo, Zeyu Huang, Zihan Qiu, Zili Wang, Jie Fu
ArXiv (2024) https://api.semanticscholar.org/CorpusID:270737825

38. **Shortcut learning in deep neural networks**
Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S Zemel, Wieland Brendel, Matthias Bethge, Felix Wichmann
Nature Machine Intelligence (2020) https://api.semanticscholar.org/CorpusID:215786368

39. **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 (2019) https://api.semanticscholar.org/CorpusID:208176471

40. **The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables**
Himabindu Lakkaraju, Jon M Kleinberg, Jure Leskovec, Jens Ludwig, Sendhil Mullainathan
Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2017) https://api.semanticscholar.org/CorpusID:2178920

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

42. **Data Analysis Using Regression and Multilevel/Hierarchical Models**
Andrew Gelman, Yu-Sung Su

43. **Optimization Methods for Large-Scale Machine Learning**
Léon Bottou, Frank E Curtis, Jorge Nocedal

44. **Contextual Bandits with Cross-learning**
Santiago R Balseiro, Negin Golrezaei, Mohammad Mahdian, Vahab S Mirrokni, Jon Schneider

45. **Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers**
Stephen P Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein
Found. Trends Mach. Learn.

46. **Adam: A Method for Stochastic Optimization**
Diederik P Kingma, Jimmy Ba

47. **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
ArXiv (2020) https://api.semanticscholar.org/CorpusID:218971783

48. **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
ArXiv (2022) https://api.semanticscholar.org/CorpusID:249674500

49. **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
ArXiv (2022) https://api.semanticscholar.org/CorpusID:247155069

50. **Scaling Laws for Neural Language Models**
Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, Dario Amodei
(2020) https://arxiv.org/abs/2001.08361

51. **Training Compute-Optimal Large Language Models**
Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, ... Laurent Sifre
(2022) https://arxiv.org/abs/2203.15556

52. **Publication Trends on the Varying Coefficients Model: Estimating the Actual (Under)Utilization of a Highly Acclaimed Method for Studying Statistical Interactions**
Assaf Botzer
Publ. (2025) https://api.semanticscholar.org/CorpusID:277710491

53. **Covariance Selection**
AP Dempster
Biometrics (1972-03) https://doi.org/d5t49s
DOI: 10.2307/2528966

54. **High-dimensional graphs and variable selection with the Lasso**
Nicolai Meinshausen, Peter Bühlmann
The Annals of Statistics (2006-06-01) https://doi.org/fwt5kt
DOI: 10.1214/009053606000000281

55. **Sparse inverse covariance estimation with the graphical lasso**
Jerome Friedman, Trevor Hastie, Robert Tibshirani
Biostatistics (2007-12-12) https://doi.org/db7svr
DOI: 10.1093/biostatistics/kxm045 · PMID: 18079126 · PMCID: PMC3019769

56. **Joint estimation of multiple graphical models**
J Guo, E Levina, G Michailidis, J Zhu
Biometrika (2011-02-09) https://doi.org/fqvbh2
DOI: 10.1093/biomet/asq060 · PMID: 23049124 · PMCID: PMC3412604

57. **The Joint Graphical Lasso for Inverse Covariance Estimation Across Multiple Classes**
Patrick Danaher, Pei Wang, Daniela M Witten

Journal of the Royal Statistical Society Series B: Statistical Methodology (2013-08-12)
https://doi.org/f5sj9g
DOI: 10.1111/rssb.12033 · PMID: 24817823 · PMCID: PMC4012833

58. **Fast Spatio-Temporally Varying Coefficient Modeling With Reluctant Interaction Selection**
Daisuke Murakami, Shinichiro Shirota, Seiji Kajita, Mami Kajita
Geographical Analysis (2024) https://api.semanticscholar.org/CorpusID:273233361

59. **Spatially Varying Coefficient Models for Estimating Heterogeneous Mixture Effects**
Jacob R Englert, Howard Chang
(2025) https://api.semanticscholar.org/CorpusID:276482233

60. **Network Varying Coefficient Model**
Xinyan Fan, Kuangnan Fang, Wei Lan, Chih-Ling Tsai
Journal of the American Statistical Association (2025)
https://api.semanticscholar.org/CorpusID:276599564

61. **Bayesian Inference for General Gaussian Graphical Models With Application to Multivariate Lattice Data**
Adrian Dobra, Alex Lenkoski, Abel Rodriguez
Journal of the American Statistical Association (2011-12) https://doi.org/fxq5wh
DOI: 10.1198/jasa.2011.tm10465 · PMID: 26924867 · PMCID: PMC4767185

62. **Bayesian Inference of Multiple Gaussian Graphical Models**
Christine Peterson, Francesco C Stingo, Marina Vannucci
Journal of the American Statistical Association (2015-01-02) https://doi.org/f69dnj
DOI: 10.1080/01621459.2014.896806 · PMID: 26078481 · PMCID: PMC4465207

63. **Bayesian covariate-dependent graph learning with a dual group spike-and-slab prior**
Zijian Zeng, Meng Li, Marina Vannucci
Biometrics (2025-04-02) https://doi.org/g95bkg
DOI: 10.1093/biomtc/ujaf053 · PMID: 40322851

64. **Tree Boosted Varying Coefficient Models**
Yichen Zhou, Giles Hooker
arXiv: Methodology (2019) https://api.semanticscholar.org/CorpusID:91184310

65. **A tree-based varying coefficient model**
Henning Zakrisson, Mathias Lindholm
arXiv (2024) https://api.semanticscholar.org/CorpusID:266933390

66. **Penalized Spline Estimation for Varying-Coefficient Models**
Yiqiang Lu, Riquan Zhang, Liping Zhu
Communications in Statistics - Theory and Methods (2008)
https://api.semanticscholar.org/CorpusID:121809483

67. **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 (2019) https://arxiv.org/abs/1911.08731

68. **WILDS: A Benchmark of in-the-Wild Distribution Shifts**
Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Sergey Levine, Chelsea Finn, Percy Liang
Proceedings of the 38th International Conference on Machine Learning (2021)
https://proceedings.mlr.press/v139/koh21a.html

69. **Domain Adaptation under Missingness Shift**
    Helen Zhou, Sivaraman Balakrishnan, Zachary C Lipton
    Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (2023)
    https://proceedings.mlr.press/v206/zhou23b.html

70. **Variational Autoencoder with Arbitrary Conditioning**
    Oleg Ivanov, Michael Figurnov, Dmitry Vetrov
    arXiv (2019) https://arxiv.org/abs/1806.02382

71. **GAIN: Missing Data Imputation using Generative Adversarial Nets**
    Jinsung Yoon, James Jordon, Mihaela van der Schaar
    Proceedings of the 35th International Conference on Machine Learning (2018)
    https://proceedings.mlr.press/v80/yoon18a.html

72. **A Class of Pattern-Mixture Models for Normal Incomplete Data**
    Roderick JA Little
    Biometrika (1994) https://academic.oup.com/biomet/article-abstract/81/3/471/256979
    DOI: 10.1093/biomet/81.3.471

73. **Multiple imputation of incomplete multilevel data using Heckman selection models**
    Johanna Muñoz, Matthieu Audibert, Melina Pavlou, Richard D Riley, Thomas PA Debray
    arXiv (2023) https://arxiv.org/abs/2301.05043

74. **XGBoost: A Scalable Tree Boosting System**
    Tianqi Chen, Carlos Guestrin
    Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016) https://dl.acm.org/doi/10.1145/2939672.2939785
    DOI: 10.1145/2939672.2939785

75. **The Missing Indicator Method: From Low to High Dimensions**
    Mike Van Ness, Tomas M Bosschieter, Roberto Halpin-Gregorio, Madeleine Udell
    arXiv (2022) https://arxiv.org/abs/2211.09259

76. **Recurrent Neural Networks for Multivariate Time Series with Missing Values**
    Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, Yan Liu
    Scientific Reports (2018) https://www.nature.com/articles/s41598-018-24271-9
    DOI: 10.1038/s41598-018-24271-9

77. **BRITS: Bidirectional Recurrent Imputation for Time Series**
    Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, Yitan Li
    Advances in Neural Information Processing Systems 31 (NeurIPS 2018) (2018)
    https://proceedings.neurips.cc/paper/2018/hash/734e6bfcd358e25ac1db0a4241b95651-Abstract.html

78. **On the Opportunities and Risks of Foundation Models**
    Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora

79. **An Overview of Multi-Task Learning in Deep Neural Networks**
    Sebastian Ruder
    ArXiv (2017) https://api.semanticscholar.org/CorpusID:10175374

80. **Attention is All you Need**
    Ashish Vaswani, Noam M Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, Illia Polosukhin
    Neural Information Processing Systems (2017)
    https://api.semanticscholar.org/CorpusID:13756489

81. **Auto-Encoding Variational Bayes**
Diederik P Kingma, Max Welling
CoRR (2013) https://api.semanticscholar.org/CorpusID:216078090

82. **Meta-Learning in Neural Networks: A Survey**
Timothy M Hospedales, Antreas Antoniou, Paul Micaelli, Amos J Storkey
IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
https://api.semanticscholar.org/CorpusID:215744839

83. **Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks**
Chelsea Finn, P Abbeel, Sergey Levine
International Conference on Machine Learning (2017)
https://api.semanticscholar.org/CorpusID:6719686

84. **Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers**
Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, Furu Wei
(2022) https://api.semanticscholar.org/CorpusID:258686544

85. **Transformers as Support Vector Machines**
Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, Samet Oymak
ArXiv (2023) https://api.semanticscholar.org/CorpusID:261395206

86. **An Explanation of In-context Learning as Implicit Bayesian Inference**
Sang Michael Xie, Aditi Raghunathan, Percy Liang, Tengyu Ma
ArXiv (2021) https://api.semanticscholar.org/CorpusID:241035330

87. **In-context Learning and Induction Heads**
Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova Dassarma, TJ Henighan, Benjamin Mann, Amanda Askell, Yuntao Bai, Anna Chen, … Chris Olah
ArXiv (2022) https://api.semanticscholar.org/CorpusID:252532078

88. **A Survey on In-context Learning**
Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, Zhifang Sui
Conference on Empirical Methods in Natural Language Processing (2022)
https://api.semanticscholar.org/CorpusID:255372865

89. **Chain of Thought Prompting Elicits Reasoning in Large Language Models**
Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H Chi, F Xia, Quoc Le, Denny Zhou

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

91. **Language Models as Knowledge Bases?**
Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, Sebastian Riedel

92. **Towards A Rigorous Science of Interpretable Machine Learning**
Finale Doshi-Velez, Been Kim
arXiv: Machine Learning (2017) https://api.semanticscholar.org/CorpusID:11319376

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

(2024) https://api.semanticscholar.org/CorpusID:267412927

94. **"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 (2016) https://api.semanticscholar.org/CorpusID:13029170

95. **A Unified Approach to Interpreting Model Predictions**
Scott M Lundberg, Su-In Lee
Neural Information Processing Systems (2017)
https://api.semanticscholar.org/CorpusID:21889700

96. **Axiomatic Attribution for Deep Networks**
Mukund Sundararajan, Ankur Taly, Qiqi Yan
International Conference on Machine Learning (2017)
https://api.semanticscholar.org/CorpusID:16747630

97. **Learning Important Features Through Propagating Activation Differences**
Avanti Shrikumar, Peyton Greenside, Anshul Kundaje
International Conference on Machine Learning (2017)
https://api.semanticscholar.org/CorpusID:3385018

98. **This looks like that: deep learning for interpretable image recognition**
Chaofan Chen, Oscar Li, Alina Jade Barnett, Jonathan Su, Cynthia Rudin
Neural Information Processing Systems (2018)
https://api.semanticscholar.org/CorpusID:49482223

99. **Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning**
Nicolas Papernot, Patrick Mcdaniel
ArXiv (2018) https://api.semanticscholar.org/CorpusID:3882460

100. **Understanding Black-box Predictions via Influence Functions**
Pang Wei Koh, Percy Liang
International Conference on Machine Learning (2017)
https://api.semanticscholar.org/CorpusID:13193974

101. **Inference Suboptimality in Variational Autoencoders**
Chris Cremer, Xuechen Li, David Kristjanson Duvenaud
International Conference on Machine Learning (2018)
https://api.semanticscholar.org/CorpusID:3524184

102. **beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework**
Irina Higgins, Loïc Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, Alexander Lerchner
International Conference on Learning Representations (2016)
https://api.semanticscholar.org/CorpusID:46798026

103. **Deep Variational Information Bottleneck**
Alexander A Alemi, Ian Fischer, Joshua V Dillon
ArXiv (2017) https://api.semanticscholar.org/CorpusID:204922497

104. **Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)**
Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J Cai, James Wexler, Fernanda B Viégas, Rory Sayres

International Conference on Machine Learning (2017)
https://api.semanticscholar.org/CorpusID:51737170

105. **Towards Automatic Concept-based Explanations**
Amirata Ghorbani, James Wexler, James Y Zou, Been Kim
Neural Information Processing Systems (2019)
https://api.semanticscholar.org/CorpusID:184487319

106. **Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations**
Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Scholkopf, Olivier Bachem
International Conference on Machine Learning (2018)
https://api.semanticscholar.org/CorpusID:54089884

107. **Understanding intermediate layers using linear classifier probes**
Guillaume Alain, Yoshua Bengio
ArXiv (2016) https://api.semanticscholar.org/CorpusID:9794990

108. **What do you learn from context? Probing for sentence structure in contextualized word representations**
Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, RThomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, Ellie Pavlick
ArXiv (2019) https://api.semanticscholar.org/CorpusID:108300988

109. **Locating and Editing Factual Associations in GPT**
Kevin Meng, David Bau, Alex Andonian, Yonatan Belinkov
Neural Information Processing Systems (2022)
https://api.semanticscholar.org/CorpusID:255825985

110. **Knowledge Neurons in Pretrained Transformers**
Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Furu Wei
ArXiv (2021) https://api.semanticscholar.org/CorpusID:233296761

111. **In-Context Explainers: Harnessing LLMs for Explaining Black Box Models**
Nicholas Kroeger, Dan Ley, Satyapriya Krishna, Chirag Agarwal, Himabindu Lakkaraju
(2023) https://api.semanticscholar.org/CorpusID:263829193

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

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

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

115. **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

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

117. **Causal inference using invariant prediction: identification and confidence intervals**
Jonas Peters, Peter Bühlmann, Nicolai Meinshausen
arXiv (2024-04-26) https://arxiv.org/abs/1501.01332

118. **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

119. **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

120. **Adversarial Examples Are Not Bugs, They Are Features**
Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, Aleksander Madry
arXiv (2019-08-13) https://arxiv.org/abs/1905.02175

121. **The Rise and Potential of Large Language Model Based Agents: A Survey**
Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, … Tao Gui
arXiv (2023-09-20) https://arxiv.org/abs/2309.07864

122. **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

123. **Just Train Twice: Improving Group Robustness without Training Group Information**
Evan Zheran Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, Chelsea Finn
arXiv (2021-09-28) https://arxiv.org/abs/2107.09044

124. **Environment Inference for Invariant Learning**
Elliot Creager, Jörn-Henrik Jacobsen, Richard Zemel
arXiv (2021-07-16) https://arxiv.org/abs/2010.07249

125. **A review of multivariate longitudinal data analysis.**
S Bandyopadhyay, B Ganguli, A Chatterjee
Statistical methods in medical research (2010-03-08)
https://www.ncbi.nlm.nih.gov/pubmed/20212072
DOI: 10.1177/0962280209340191 · PMID: 20212072

126. **[A field study: piroxicam in arthrosis and soft tissue rheumatism. Results of a multicenter study].**
O Knüsel, G Hinz
Schweizerische Rundschau fur Medizin Praxis = Revue suisse de medecine Praxis (1983-09-13)
https://www.ncbi.nlm.nih.gov/pubmed/6356121
PMID: 6356121

127. **A Bayesian multilevel time-varying framework for joint modeling of hospitalization and survival in patients on dialysis.**
Esra Kürüm, Danh V Nguyen, Sudipto Banerjee, Yihao Li, Connie M Rhee, Damla Şentürk

Statistics in medicine (2022-10-01) https://www.ncbi.nlm.nih.gov/pubmed/36181392
DOI: 10.1002/sim.9582 · PMID: 36181392 · PMCID: PMC9931182

128. **Dynamic effects of increasing heterogeneity in financial markets**
Ahmad K Naimzada, Giorgio Ricchiuti
Chaos, Solitons & Fractals (2009-08) https://doi.org/bfbqxn
DOI: 10.1016/j.chaos.2008.07.022

129. **Bayesian Forecasting in Economics and Finance: A Modern Review**
Gael M Martin, David T Frazier, Worapree Maneesoonthorn, Ruben Loaiza-Maya, Florian Huber, Gary Koop, John Maheu, Didier Nibbering, Anastasios Panagiotelis
arXiv (2023-08-01) https://arxiv.org/abs/2212.03471

130. **Bayesian Dynamic Factor Models for High-dimensional Matrix-valued Time Series**
Wei Zhang
arXiv (2025-08-11) https://arxiv.org/abs/2409.08354

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

132. **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

133. **Chaotic Bayesian Inference: Strange Attractors as Risk Models for Black Swan Events**
Crystal Rust
arXiv (2025-09-11) https://arxiv.org/abs/2509.08183

134. **A Multi-target Bayesian Transformer Framework for Predicting Cardiovascular Disease Biomarkers during Pandemics**
Trusting Inekwe, Emmanuel Agu, Winnie Mkandawire, Andres Colubri
arXiv (2025-09-03) https://arxiv.org/abs/2509.01794

135. **Bayesian Models for Joint Selection of Features and Auto-Regressive Lags: Theory and Applications in Environmental and Financial Forecasting**
Alokesh Manna, Sujit K Ghosh
arXiv (2025-08-18) https://arxiv.org/abs/2508.10055

136. **SafeInfer: Context Adaptive Decoding Time Safety Alignment for Large Language Models**
Somnath Banerjee, Sayan Layek, Soham Tripathy, Shanu Kumar, Animesh Mukherjee, Rima Hazra
arXiv (2024-12-17) https://arxiv.org/abs/2406.12274

137. **Robustness, Evaluation and Adaptation of Machine Learning Models in the Wild**
Vihari Piratla
arXiv (2023-03-05) https://arxiv.org/abs/2303.02781v1

138. **A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions**
Shailja Gupta, Rajesh Ranjan, Surya Narayan Singh
arXiv (2024-10-18) https://arxiv.org/abs/2410.12837

139. **Retrieval-Augmented Generation for AI-Generated Content: A Survey**
Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, Bin Cui

arXiv (2024-06-24) https://arxiv.org/abs/2402.19473

140. **Billion-scale similarity search with GPUs**
Jeff Johnson, Matthijs Douze, Hervé Jégou
arXiv (2018-06-07) https://arxiv.org/abs/1702.08734

141. **From Human Memory to AI Memory: A Survey on Memory Mechanisms in the Era of LLMs**
Yaxiong Wu, Sheng Liang, Chen Zhang, Yichao Wang, Yongyue Zhang, Huifeng Guo, Ruiming Tang, Yong Liu
arXiv (2025-04-24) https://arxiv.org/abs/2504.15965

142. **Memory OS of AI Agent**
Jiazheng Kang, Mingming Ji, Zhe Zhao, Ting Bai
arXiv (2025-06-10) https://arxiv.org/abs/2506.06326

143. **Efficient Streaming Language Models with Attention Sinks**
Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, Mike Lewis
arXiv (2024-04-09) https://arxiv.org/abs/2309.17453

144. **Efficient Memory Management for Large Language Model Serving with PagedAttention**
Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E Gonzalez, Hao Zhang, Ion Stoica
arXiv (2023-09-13) https://arxiv.org/abs/2309.06180

145. **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

146. **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

147. **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

148. **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

149. **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

150. **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](#)

151. **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](#)

152. **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](#)

153. **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](#)

154. **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](#)

155. **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](#)

156. **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](#)

157. **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](#)

158. **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](#)

159. **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](#)

160. **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

161. **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

162. **Mixture of LoRA Experts**
Xun Wu, Shaohan Huang, Furu Wei
arXiv (2024) https://doi.org/g93rnr
DOI: 10.48550/arxiv.2404.13628

163. **Investigating Lane-Free Traffic with a Dynamic Driving Simulator**
Maya Sekeran, Arslan Ali Syed, Johannes Lindner, Martin Margreiter, Klaus Bogenberger
arXiv (2023) https://doi.org/g93rnq
DOI: 10.48550/arxiv.2311.16142

164. **WILDS: A Benchmark of in-the-Wild Distribution Shifts**
Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, … Percy Liang
arXiv (2020) https://doi.org/g93rnp
DOI: 10.48550/arxiv.2012.07421

165. **Shortcut Learning in Deep Neural Networks**
Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, Felix A Wichmann
arXiv (2020) https://doi.org/g93rnn
DOI: 10.48550/arxiv.2004.07780

166. **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 (2019) https://doi.org/g93rnm
DOI: 10.48550/arxiv.1911.08731

167. **Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers**
Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, Furu Wei
arXiv (2022) https://doi.org/gtkkf9
DOI: 10.48550/arxiv.2212.10559

168. **An Overview of Perception Methods for Horticultural Robots: From Pollination to Harvest**
Ho Seok Ahn, Feras Dayoub, Marija Popovic, Bruce MacDonald, Roland Siegwart, Inkyu Sa
arXiv (2018) https://doi.org/g93rnk
DOI: 10.48550/arxiv.1807.03124

169. **Publication Trends on the Varying Coefficients Model: Estimating the Actual (Under)Utilization of a Highly Acclaimed Method for Studying Statistical Interactions**
Assaf Botzer
Publications (2025-04-07) https://doi.org/g9t2rq
DOI: 10.3390/publications13020019

170. **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:

171. **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