Context-Adaptive Inference: Bridging Statistical and Foundation Models

Authors

- **Yue Yao**
  [ID] 0009-0000-8195-3943 · [GitHub] YueYao-stat
  Department of Statistics, University of Wisconsin-Madison

- **Caleb N. Ellington**
  [ID] 0000-0001-7029-8023 · [GitHub] cnellington · [Twitter] probablybots
  Computational Biology Department, Carnegie Mellon University

- **Jingyun Jia**
  [ID] 0009-0006-3241-3485 · [GitHub] Clouddelta
  Department of Statistics, University of Wisconsin-Madison

- **Baiheng Chen**
  [ID] 0000-0001-8554-3984 · [GitHub] BaihengChen
  Department of Statistics, University of Wisconsin-Madison

- **Dong Liu**
  [ID] 0009-0009-6815-8297 · [GitHub] NoakLiu
  Department of Computer Science, Yale University

- **Rikhil Rao**
  [ID] 0009-0003-5221-1125 · [GitHub] rikhilr
  Department of Computer Science, University of Wisconsin - Madison

- **Jiaqi Wang**
  [ID] 0009-0003-8531-9490 · [GitHub] w-jiaqi
  Paul G. Allen School of Computer Science & Engineering, University of Washington

- **Samuel Wales-McGrath**
  [ID] 0009-0008-5405-2646 · [GitHub] Samuel-WM
  Department of Computer Science and Engineering, The Ohio State University

- **Yixin Yang**
  [ID] 0009-0008-1436-5584 · [GitHub] Jessego5
  Department of Computer Sciences, University of Wisconsin-Madison

- **Zhiyuan Li**
  [ID] 0009-0006-6016-7381 · [GitHub] LZYEIL
  Department of Computer Sciences, University of Wisconsin-Madison

- **Ben Lengerich** [✉]
  [ID] 0000-0001-8690-9554 · [GitHub] blengerich · [Twitter] ben_lengerich
  Department of Statistics, University of Wisconsin-Madison

✉ — Correspondence possible via [GitHub Issues](#) or email to Ben Lengerich <lengerich@wisc.edu>.

Abstract

Context-adaptive inference enables models to adjust their behavior across individuals, environments, or tasks. Adaptation may be *explicit*, through parameterized functions of context, or *implicit*, as in foundation models that respond to prompts and support in-context learning. This review synthesizes advances in varying coefficient models and adaptive behavior in foundation models, and discusses how statistical and neural approaches converge in context-aware inference. We highlight how foundation models can serve as flexible encoders of context, and how statistical methods offer structure and interpretability. Building on these links, we propose a common perspective that places explicit and implicit mechanisms on the same spectrum of context use. We also identify open problems in identifiability, robustness under distribution shift, and efficient large-scale adaptation, outlining design principles for methods that are scalable, reliable, and transparent in real-world settings.

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. When the i.i.d. assumption breaks down, using a single global model can obscure meaningful heterogeneity.

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. Figure 1 summarizes how statistics, meta-learning, and foundation models flow into a unified **context→parameters** view, which we formalize in (★).

Statistics
- Varying-coefficient models
- Local / kernelized smoothing
- Hierarchical or multitask sharing

Meta-learning
- Bilevel training across tasks
- Fast task adaptation (e.g., MAML/adapters)
- Hypernetworks generate task parameters

Foundation models
- Prompted inference / in-context learning (ICL)
- Linear readout on prompt-induced features
- Retrieval / tools as contextual signals

Theoretical Bridge
- One view: context → parameters
- Context via support, similarity, structure

Shared Tuning Knobs
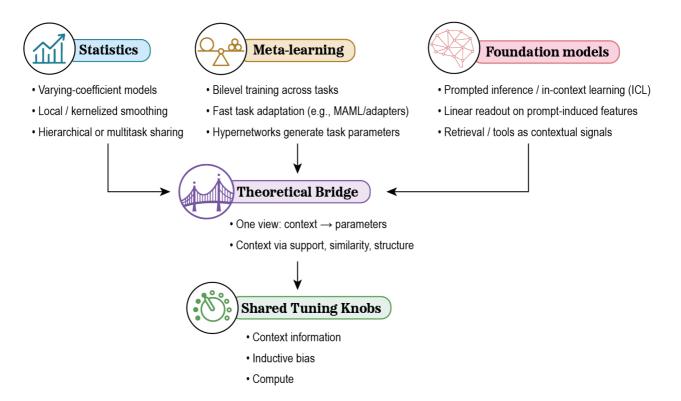- Context information
- Inductive bias
- Compute

Figure 1: Overview of the theoretical bridge. Three traditions—Statistics (varying-coefficients, local smoothing, hierarchical sharing), Meta-learning (bilevel training, fast adaptation, hypernetworks), and Foundation models (prompted inference / in-context learning)—feed into a unified context→parameters view. The bridge formalizes this connection and highlights shared tuning knobs: context information, inductive bias, and compute.

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 without structural regularization—smoothness, sparsity, shared representations, or latent grouping. And as adaptivity becomes more implicit (e.g., via neural networks or black-box inference), we need tools to recover, interpret, or constrain the underlying parameter variation.

Problem Setup and Notation

We study supervised prediction with units $i = 1, \ldots, n$. Each unit has a **context** $c_i \in \mathcal{C}$ (e.g., patient/user/site/time) and observed data $\mathcal{D}_i = \{(x_{ij}, y_{ij})\}_{j=1}^{m_i}$ with $x_{ij} \in \mathcal{X}$ and $y_{ij} \in \mathcal{Y}$. Predictions come from a model family $\mathcal{H} = \{h_\theta : \mathcal{X} \to \mathcal{Y} \mid \theta \in \Theta\}$ (e.g., linear model, neural net, probabilistic model).

In **global** (i.i.d.) models, $\theta_i \equiv \theta^\star$. In **context-adaptive** models, parameters vary with context: $\theta_i = f(c_i)$ or $\theta_i \sim P(\theta \mid c_i)$.

For a new unit with context c, we write a unified empirical objective:

$$\hat{\theta}(c) \in \arg\min_{\theta \in \Theta} \underbrace{\sum_{(i,j) \in S(c)} \ell(h_\theta(x_{ij}), y_{ij})}_{\text{context-dependent support}} + \underbrace{\mathcal{R}(\theta; c)}_{\text{context-structured regularization}}, \qquad (\bigstar)$$

where ℓ is a proper loss (e.g., squared, logistic), $S(c) \subseteq \{1, \ldots, n\} \times \mathbb{N}$ is a **support set** selected for context c, and $\mathcal{R}(\theta; c)$ encodes how parameters are allowed to vary with context (smoothness, sparsity, low-rank, hierarchy, etc.).

How context enters.
- **Explicit parameterization:** a map $f : \mathcal{C} \to \Theta$ sets $\theta_i = f(c_i)$ (e.g., varying-coefficients, hierarchical Bayes, multi-task/meta-learning). Here $\mathcal{R}(\theta; c)$ typically regularizes f (e.g., Lipschitz over $\mathcal{C}$, group lasso, low-rank). - **Implicit parameterization:** context alters optimization or internal states without exposing θ directly (e.g., mixture-of-experts with gates $g(x, c)$; retrieval where $S(c)$ is built by a retriever $R(c)$; in-context learning where a prompt map $P(c)$ conditions a foundation model).

For convenience, we use a **context encoder** $\phi : \mathcal{C} \to \mathbb{R}^d$ and a similarity/kernel $K(c, c')$. A common instance of ($\bigstar$) is kernel-weighted risk:

$$\sum_{i,j} w_{ij}(c) \ell(h_\theta(x_{ij}), y_{ij}) + \mathcal{R}(\theta), \qquad w_{ij}(c) \propto K(\phi(c), \phi(c_i)) \cdot \mathbf{1}[(i, j) \in S(c)].$$

Granularity. We refer to adaptation granularity $g \in \{\mathrm{group}, \mathrm{unit}, \mathrm{example}\}$ and to three design "knobs" we will revisit:
1) **Information** via $S(c)$ or $P(c)$ (what context is exposed),
2) **Inductive bias** via $\mathcal{R}(\theta; c)$ (how parameters may vary),
3) **Compute** via warm-starts/caching/steps (how aggressively we solve ($\star$) at test time).

Standing assumptions (used as needed).
(i) *Exchangeability within context:* conditional on (θ_i, c_i), (x_{ij}, y_{ij}) are i.i.d.
(ii) *Regularity:* either $\theta = f(c)$ with f in a regular class (e.g., Lipschitz/sparse/low-rank) or retrieval weights $w_{ij}(c)$ are bounded and locally normalized.
(iii) *Identifiability/stability:* ℓ is convex in model outputs and $\mathcal{R}$ yields a unique or stable minimizer.
(iv) *Resource tracking:* we track $|S(c)|$, optimization steps, and memory to compare **adaptation efficiency**.

With this notation in place, we now formalize the link between explicit and implicit context adaptation. This link has been contributed in pieces across recent work; here we unify those results.

Theoretical Bridge

Recent theoretical work suggests that "explicit" context models (e.g., varying-coefficients, hierarchical/multitask) and "implicit" mechanisms (e.g., in-context learning via attention) often implement the same estimator class under squared loss, differing mainly in how they encode neighborhoods and regularization. Using the notation above, we make this precise.

Proposition 1 (Explicit varying-coefficients and linear ICL coincide with kernel ridge on joint features in the linear squared-loss setting).
Assume squared loss and the regression model $y = \langle \theta(c), x \rangle + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$. Let (i) a context encoder $\phi : \mathcal{C} \to \mathbb{R}^{d_c}$, (ii) joint features $\psi(x, c) := x \otimes \phi(c) \in \mathbb{R}^{d_x d_c}$, (iii) a context-dependent support set $S(c)$ with nonnegative weights $w_{ij}(c)$.

- **(A) Explicit varying-coefficients.** Let $\theta(c) = B\,\phi(c)$ with $B \in \mathbb{R}^{d_x \times d_c}$ and ridge penalty $\lambda \|B\|_F^2$. The weighted ridge solution yields
$$\hat{y}(x, c) = k_{(x,c)}^\top \big(K + \lambda I\big)^{-1} y, \quad K_{ab} = \langle \psi_a, \psi_b \rangle = \langle x_a, x_b \rangle \cdot \langle \phi(c_a), \phi(c_b) \rangle,$$
i.e., **kernel ridge regression (KRR)** on joint features.

- **(B) Implicit adaptation via linear ICL.** Let a single linear attention layer consume the weighted support set $S(c)$ with linear $q = Q\psi$, $k = K\psi$, $v = V\psi$ and a linear readout. With attention weights proportional to $w_{ij}(c) \cdot \langle q, k_{ij} \rangle$, the induced predictor equals KRR with kernel
$$k\big((x, c), (x', c')\big) = \langle q(x, c), k(x', c') \rangle,$$
i.e., a learned **dot-product kernel** on the same joint features. If attention parameters are trained in the linearized/NTK regime, learning equals kernel regression with the network's NTK, which is again a dot-product kernel on linear transforms of ψ.

Corollary 1 (Retrieval, gating, and weighting are kernel/measure choices). Choosing $S(c)$ via a retriever $R(c)$, or gating in a mixture-of-experts, corresponds to changing the kernel and/or the empirical measure (weights $w_{ij}(c)$) used by KRR on ψ.

Proof in Appendix A.

Positioning and prior art. Proposition 1 is expository: part (A) is standard ridge⇔kernel duality on joint features; part (B) follows from (i) fixed attention + trained linear head = ridge on fixed features and (ii) NTK linearization ⇒ kernel regression with the network's NTK. Our contribution is the unified **context-aware** formulation: explicit design knobs via $S(c)$, $\mathcal{R}(\theta; c)$, and compute—and the mapping of retrieval/gating to kernel/measure choices. See transformer ICL as classical estimators [1,2,3] and NTK analyses [4,5].

Limitations This linear, squared-loss bridge captures a large class of explicit and implicit adaptors, but it abstracts away at least three realities: (i) non-quadratic losses (e.g., logistic) change the effective kernel/weighting via loss curvature; (ii) when the prediction head is nonlinear (e.g., an MLP or attention), it introduces model curvature: the head's Jacobian/Hessian depends on the input and parameters, so the effective metric and weights change with the representation; the fixed-kernel view from the linear case no longer holds; (iii) multi-modal context encoders (text, graphs, images) alter both the neighborhood definition and the regularizer. We view these as open extensions to explore beyond the scope of this review.

Scope of Review and Relation to Prior Work

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
Statistical Methods with Varying Coefficient Models[6]	Varying-coefficient modeling	Classical statistical modeling, with parameters expressed as functions of covariates	Explicit adaptivity: parameters change smoothly with context via (f(c))	Limited to explicit, parametric formulations; no connection to neural or emergent adaptation

Survey	Topic Focus	Scope	Coverage of Adaptivity	Gap Relative to This Work
A Survey of Deep Meta-Learning [7]	Meta-learning	Neural meta-learning methods for cross-task adaptation	Task-level adaptivity: models learn to generalize quickly across tasks	Focused on task switching; does not integrate explicit parameter modeling or implicit foundation model adaptation
LoRA: Low-Rank Adaptation of Large Language Models[8]	Parameter-efficient adaptation	Adaptation of large pretrained transformer models via low-rank updates while freezing base weights	Implicit adaptivity via parameter-efficient updates, enabling contextual adaptation without full fine-tuning	Strong in efficient adaptation mechanism, but narrow in scope; does not address explicit contextual structure or cross-domain generalization
Foundational Models Defining a New Era in Vision: A Survey and Outlook[9]	Vision-based foundation models	Architectures, multimodal integration, prompting, fusion in vision models	Implicit adaptivity in vision contexts, via prompt or fusion mechanisms across visual tasks	Domain-specific focus limits generalization; less discussion on theoretical adaptation across modalities
A Comprehensive Survey on Pretrained Foundation Models[10]	Pretrained foundation models	Coverage of models across modalities, training regimes, adaptation and fine-tuning strategies	Implicit adaptivity via representation transfer and generalization across tasks	Broad in scope but does not deeply analyze parameter-level adaptation or explicit–implicit alignment

Table 1: Representative surveys and key papers covering context-adaptive inference. Most works 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. The next section outlines the conceptual foundations of context-adaptive inference, preparing the ground for detailed discussions of explicit and implicit modeling approaches in later sections.

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. As data grows more complex, the failure of this assumption not only limits accuracy but also obscures causal and contextual relationships underlying modern inference.

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 2: 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 3).
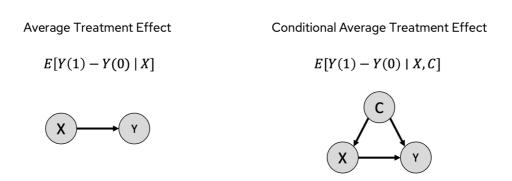
Average Treatment Effect

$$E[Y(1) - Y(0) \mid X]$$

Conditional Average Treatment Effect

$$E[Y(1) - Y(0) \mid X, C]$$

Figure 3: 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 $f(c)$.

Classical 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.

These early methods relax global homogeneity yet still rely on discrete partitions, motivating smoother and more flexible formulations explored in the next sections.

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 [11]. The estimation can be written as

$$\widehat{A} = \arg \max_A \sum_i \ell(x_i; A c_i).$$

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 [12].

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 [13,14]. 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 [15]. 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:

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

This framework has been widely applied, from machine learning toolboxes [16,17] to personalized genomics [18,19], biomedical informatics [20,21,22], and contextual feature selection [23]. 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 [24]:

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

By encouraging piecewise-constant structures, partition models get closer to personalized modeling, balancing fit and parsimony, moving closer to personalized inference, trading off flexibility for interpretability.

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 [25]. 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 [13].

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 [26].

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.
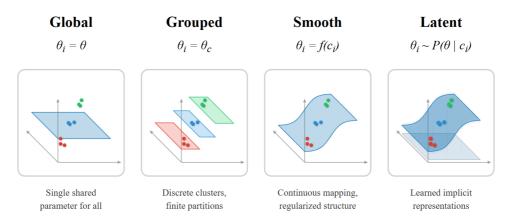

Figure 4: A spectrum of context awareness in modeling, showing global, grouped, smooth, and latent models

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 [27].

Independent and identically distributed samples

The initial research data mostly came from strictly designed experiments, such as agricultural field trials or psychological experiments. The characteristics of such data are small scale and simple structure. Researchers usually assume that each observation is independent of one another and identically distributed [28]. Under this setting, there is no dependency among the data, and researchers mainly focus on the overall average level or effect.

Linear models emerged as the fundamental approach to conduct statistical analysis for such data. One of the first methods in the development of linear models, the method of least squares, was first published in a work by Legendre [29] and later independently developed and justified by Gauss [30]. By reducing the squared deviations between predictions and results, this method offered a general framework for fitting a regression line via observed data. The estimator is expressed as follows:

$$\hat{\beta} = \arg \min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i^\top \beta \right)^2$$

which has a closed-form solution,

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

This offered a systematic procedure for predicting unknown parameters from independently and identically distributed data, which builds the foundation of regression analysis.

The establishment of generalized linear models (GLMs) expanded the concepts of linear regression to non-Gaussian outcomes. Nelder and Wedderburn [31] initiated the central concept of connecting the mean of the response variable to a linear predictor through a link function, which was later formalized into the now-standard compact notation [32]:

$$g(\mu_i) = \eta_i$$

Before this unified general framework, one of the first instances was the use of the logistic function for binary data. Berkson advocated its application to biomedical dose–response studies, illustrating how the probabilities of success or failure in experimental settings could be captured by the link formulation [33]. Based on this, logistic regression was formalized as a regression model for binary sequences, establishing the logit link [34]:

$$\log \frac{p_i}{1 - p_i} = x_i^\top \beta$$

For count data, Poisson regression was introduced within the GLM framework, employing the log link [31]:

$$\log(\mu_i) = x_i^\top \beta$$

These developments strengthened GLMs as a unifying framework for a variety of independently and identically distributed data types by extending linear modeling to categorical and count outcomes.

Alongside regression, analysis of variance (ANOVA) was another early milestone of statistical methodology. The development of ANOVA introduced the concept of splitting total variance into components related to within-group and between-group differences [35]. The central idea for ANOVA was the F ratio,

$$F = \frac{MS_{\text{Between}}}{MS_{\text{Within}}}$$

which provided a consistent framework for determining the significance of differences between group means and established the foundation of data analysis in modern experiments.

Together, these early statistical frameworks provided the foundation for analysis of independently and identically distributed data. However, as studies became more complex and observations were no longer truly independent, new methods were needed, leading to the development of hierarchical models.

Hierarchical data

With the expansion of research fields, a hierarchical structure gradually emerges in the collected empirical data. For example, bull is nested in sire and herd [36]; the respondents are nested within the population of subjects [37]; and repeated measurements of the same object at different time points also make the data show longitudinal correlation. This type of data is more complex than independent and identically distributed samples. It not only has differences among groups but also needs to take into account the changes of individuals over time simultaneously. The observed values are no longer completely independent of each other, but there exists a distinct hierarchical effect.

Early work on hierarchical dependence started with the introduction of linear models that contain both fixed and random effects to estimate genetic parameters [38]. This approach laid the statistical basis for partitioning variability into systematic and random components, which became the basis for how intraclass correlation would later be understood and applied. A general form of the linear models can be written as:

$$y_{ij} = \mu + u_j + \varepsilon_{ij}, \quad u_j \sim N(0, \sigma_u^2), \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

Around the same period, it became clear that comparative rate estimation in clinical and epidemiological studies depends on the structure of sampled subgroups, with heterogeneity in source populations affecting inference in widespread use [39]. Building on this recognition of heterogeneity, applications expanded to repeated measurements and longitudinal structures; restricted maximum likelihood estimation was introduced to advance the framework and improve variance component inference in unbalanced settings [40]. This methodological foundation enabled the formulation of linear mixed-effects models (LMMs) that combined fixed effects for population-level trends with random effects for subject- or group-specific deviations [41]. In matrix notation, these models take the form

$$y = X\beta + Zu + \varepsilon, \quad u \sim N(0, G), \quad \varepsilon \sim N(0, R)$$

and later became widely used in the field of biostatistics and social sciences. To extend mixed-effects modeling beyond Gaussian responses, practical estimation procedures for generalized linear models with random effects were proposed, enabling the application of link functions to clustered binary, count, or categorical outcomes [42]. Further methodological advances introduced penalized quasi-likelihood and approximate inference techniques that made practical application feasible in many fields, especially medical and biological fields [43]. These developments gave rise to Generalized Linear Mixed Models (GLMMs), formulated as

$$g(\mu_i) = x_i^\top \beta + z_i^\top u$$

combining link functions with both fixed (β) and random (u) effects. As mixed-effects approaches developed rapidly, the conceptual foundation of Bayesian hierarchical modeling—in which multilevel structures with prior distributions link parameters across groups—had already been articulated [44]:

$$y_{ij} \mid \theta_j \sim p(y_{ij} \mid \theta_j), \quad \theta_j \sim p(\theta_j \mid \phi), \quad \phi \sim p(\phi)$$

Nevertheless, it wasn't until later that the Bayesian hierarchical model had practical influence. In the 1980s, the demonstration of the practical use of Gibbs sampling in image analysis paved the way for adoption in hierarchical Bayesian modeling [45]. Later, advances in Markov chain Monte Carlo methods provided the computational tools necessary to fit Bayesian hierarchical models and made Bayesian hierarchical modeling practical for applied researchers [46].

Collectively, these developments shifted statistical modeling from independence assumptions to explicitly capturing correlation and hierarchical structure. As methods for nested data improved, new problems arose with functional, continuous, and high-dimensional observations. This led to the development of approaches for curves, trajectories, and large feature spaces, which in turn led to functional data analysis and high-dimensional inference.

Functional types and high-dimensional data

As data collection advanced into the 1980s and 1990s, researchers began encountering observations that are entire curves or functions (e.g., time series, spectra, images) rather than fixed-dimensional vectors. Functional Data Analysis (FDA) [47] treats each observation as a smooth function $x_i(t)$ defined over a continuum (time, wavelength, etc.). Typical FDA methods use basis expansions (e.g., Fourier or spline bases) to represent $x_i(t)$ and perform tasks such as functional principal component analysis. This emphasis on continuous predictors demanded flexible regression tools. Generalized Additive Models (GAMs) [48] were developed to capture nonlinear effects while retaining interpretability. In a GAM, the response is modeled as

$$y_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i$$

where each f_j is an estimated smooth function (often implemented with splines or kernels). GAMs then generalize linear models by allowing the data to determine the shape of each predictor's effect. For example, a GAM can automatically reveal a U-shaped or threshold effect in a covariate, which a purely linear model would miss.

As data dimensionality and complexity continued to grow (for instance with genomic or imaging data), there was a shift toward automated feature learning. Representation learning [49] generalizes classical dimension reduction (like PCA) to nonlinear, data-driven embeddings. Neural autoencoders are a prototypical example: one trains an encoder network $h_\phi(x)$ and a decoder $g_\theta(z)$ by minimizing reconstruction error,

$$(\theta^{ast}, \phi^{ast}) = \arg\min_{\theta,\phi} \sum_{i=1}^{n} |x_i - g_\theta(h_\phi(x_i))|^2$$

Here $z_i = h_\phi(x_i)$ is a low-dimensional code capturing the essential structure of x_i. With linear networks this recovers PCA, but in general the learned code can represent highly nonlinear features. Modern representation learning methods (e.g., deep autoencoders, variational autoencoders, deep embedding networks) therefore enable models to adapt their feature extraction to complex, high-

dimensional data in a context-dependent way [49]. By learning bases and transformations from the data itself, these methods built on the ideas of FDA and GAMs to handle the evolving scale and richness of scientific data.

Heterogeneous tasks and sparse data

With the proliferation of diverse tasks and domains, statistical learning shifted toward methods that transfer information across problems. Multi-task learning (MTL) [50] arose in response: it jointly models related tasks to improve performance, especially when each task has limited data. Concretely, if we have T tasks with data (X^t, Y^t) for $t = 1, \ldots, T$, one can learn task-specific models w^t by minimizing a coupled objective, for example

$$\min_W \sum_{t=1}^{T} \sum_{i=1}^{n_t} \ell(y_i^t, f(x_i^t; w^t)) + \lambda \, \Omega(W)$$

where $W = [w^1, \ldots, w^T]$ collects all task parameters and $\Omega(W)$ is a regularizer that enforces shared structure (e.g., penalizing deviations from a common parameter). This formulation lets tasks share a representation or bias: information useful for one task can aid others. Transfer learning extends this idea between domains wherein a model learned on a large source dataset is adapted to a target domain with little labeled data [51]. For instance, a convolutional neural network pretrained on ImageNet can be fine-tuned on a small medical imaging dataset, effectively reusing learned features to improve accuracy with scarce data.

Another fundamental challenge is distribution shift between training and deployment. Covariate shift occurs when the input distribution $p(x)$ changes but the conditional $p(y|x)$ remains the same. In this case one can correct the loss by importance weighting. Denote $p_{\text{train}}(x)$ and $p_{\text{test}}(x)$ the feature distributions; then

$$\mathbb{E}_{x \sim p_{\text{test}}}[\ell(f(x), y)] = \mathbb{E}_{x \sim p_{\text{train}}}\left[\frac{p_{\text{test}}(x)}{p_{\text{train}}(x)} \, \ell(f(x), y) \right], \quad w(x) = \frac{p_{\text{test}}(x)}{p_{\text{train}}(x)}$$

More generally, domain adaptation methods address cases where both $p(x)$ and $p(y|x)$ differ across source and target domains. Techniques such as kernel mean matching, adversarial domain-invariant representations, or graph-based alignment were developed to tackle these shifts. These methods emerged as data collection became distributed across different environments (e.g., new sensors, populations, or changing conditions), requiring models that adapt to new contexts beyond the original training distribution.

Finally, the combination of many tasks and very few examples per task has given rise to few-shot learning. In few-shot learning [52], the goal is to generalize to new classes or tasks from only a handful of labeled examples. Modern approaches typically leverage prior experience across tasks or classes. Metric-based methods, such as Matching Networks and Prototypical Networks [53,54], learn an embedding $\phi(x)$ so that samples from the same class cluster together. A novel class can then be represented by the mean of its few support examples in this latent space. Alternatively, meta-learning algorithms train on many simulated few-shot tasks so the model learns to adapt quickly. These few-shot and meta-learning frameworks explicitly confront data scarcity by transferring inductive biases across tasks, enabling effective learning in regimes far beyond the traditional large-sample assumptions.

Online and interactive data

In the Internet era, the way data is collected has undergone significant changes. User behavior data, sensor data, and platform experimental data exhibit the characteristics of streaming and interactivity [55]. The generation of dynamic data poses new challenges to model construction and problem-solving. Data is not collected all at once but is generated in real time and is often related to the feedback loop of the system [56]. This type of data is more dynamic and complex compared to traditional experimental or measurement data, involving both time dependence and continuous influence from the environment and interaction.

A natural starting point to address this challenge is the concept of online learning [57]. A foundational formulation is the Online Convex Optimization (OCO) framework [58]. At each round $t = 1, \dots, T$, the learner selects a decision $x_t \in F$, where $F \subset \mathbb{R}^n$ is a convex decision set. The environment then reveals a convex loss function $c_t : F \to \mathbb{R}$, and the learner incurs loss $c_t(x_t)$. The central performance measure is the regret, defined as:

$$R(T) = \sum_{t=1}^{T} c_t(x_t) - \min_{x \in F} \sum_{t=1}^{T} c_t(x)$$

which compares the learner's cumulative loss with that of the best fixed decision in hindsight. A simple and influential algorithm in this framework is Online Gradient Descent (OGD). Given a subgradient $g_t \in \partial c_t(x_t)$, OGD performs a gradient step followed by projection back onto the feasible set:

$$x_{t+1} = \Pi_F(x_t - \eta_t g_t)$$

where η_t is a step size and Π_F denotes projection onto F. With appropriately chosen step sizes, OGD achieves the classic bound $R(T) = O(\sqrt{T})$, implying vanishing average regret. This framework provides the mathematical foundation for more complex online methods, such as bandits and reinforcement learning.

While OCO provides guarantees with respect to a fixed comparator, real-world data streams rarely remain stationary. The statistical relationship between inputs x and outputs y can change over time. This phenomenon is known as concept drift [59]. To cope with drift, online adaptive learning methods augment the basic OCO paradigm by explicitly adapting to distributional changes. A representative technique for drift detection is the Drift Detection Method (DDM) [60]. It monitors the online error rate $\hat{p}\_i$ after i samples together with its standard deviation s_i. The method keeps track of the minimum values observed $(p\_\text{min}, s\_\text{min})$, and raises an alarm when

$$\hat{p}_i + s_i \geq p_\text{min} + \alpha \cdot s_\text{min}$$

for a predefined threshold α. This statistical test signals when the current error distribution significantly deviates from the past minimum, indicating concept drift. The old model is then discarded, and a new model is trained using the data accumulated since the warning level (i.e., from k_w to k_d). After reinitialization, $(p_\text{min}, s_\text{min})$ are reset, and the updated model continues processing the incoming stream. Beyond DDM, the Early Drift Detection Method (EDDM) [61] monitors the distribution of distances between errors, making it more sensitive to gradual drifts. Ensemble-based methods such as Bagging-ADWIN and Boosting-ADWIN [62], which couple classical ensemble learning with adaptive sliding windows, have also demonstrated strong performance on evolving data streams.

Though adaptive learning addresses the challenge of non-stationarity, it still operates under a full-information feedback model: after each round, the entire loss function $c_t(\cdot)$ is revealed. In many interactive systems, however, such complete feedback is unavailable. This setting motivates the study

of partial-information online learning, captured by the classic Multi-Armed Bandit (MAB) framework [63]. Within the MAB, the agent repeatedly chooses among a finite set of actions ("arms") and receives the reward of that arm. The central challenge is the exploration–exploitation trade-off. Classical strategies include ϵ-greedy [64], which with probability $1 - \epsilon$ selects the current best arm and with probability ϵ explores. Upper Confidence Bound (UCB) algorithms [65] emphasize the "Optimism in the Face of Uncertainty," where at each round t with arm a:

$$\mathrm{UCB}_a(t) = \hat{\mu}_a + \sqrt{\frac{2 \ln t}{n_a}}$$

The arm $\arg\max_a \mathrm{UCB}_a(t)$ is chosen at each round, achieving sublinear regret. Their key limitation —ignoring contextual side information—leads naturally to Contextual Bandits. Formally, at each round t, the agent now observes a context x_t and the goal is to learn a policy $\pi : X \to A$ that maximizes expected cumulative reward. A representative algorithm is LinUCB, which assumes a linear relationship between context and reward; at round t, LinUCB selects

$$a_t = \arg\max_{a \in \mathcal{A}} \left(x_t^\top \hat{\theta}_a + \alpha \sqrt{x_t^\top A_a^{-1} x_t} \right)$$

where $\hat{\theta}_a$ is the estimated weight and A_a is the design matrix. Contextual bandits improve by personalizing decisions to the observed environment. They have been widely applied in news recommendation and adaptive experimentation. However, CB assumes actions do not affect future contexts, which motivates the richer framework of reinforcement learning.

RL formalizes sequential decision-making under uncertainty through the lens of Markov Decision Processes (MDPs) [66]. An MDP is defined by a state space S, an action space A, transition dynamics $P(s' \mid s, a)$, a reward function $r(s, a)$, and a discount factor $\gamma \in [0, 1)$. At each step t, the agent observes a state s_t, selects an action $a_t \sim \pi(\cdot \mid s_t)$, receives a reward r_t, and transitions to a new state s_{t+1}. The goal is to learn a policy π that maximizes the expected cumulative discounted reward:

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t \right]$$

Classical solution approaches include value-based methods (e.g., Q-learning) [67], which estimate action-value functions and act greedily with respect to them, and policy-based methods (e.g., policy gradient, actor–critic) [68] [69], which directly optimize the policy parameters via stochastic gradient ascent. These methods establish theoretical guarantees and have enabled applications ranging from game-playing (e.g., Go, Atari) [70] to robotics.

The trajectory from online convex optimization to reinforcement learning highlights how the statistical study of streaming, interactive data has evolved into increasingly expressive frameworks for context-adaptive inference. As data collection becomes more interactive and non-stationary, the ability to learn from context and feedback loops is central to adaptive intelligence. This theme naturally connects to the next section on multimodal data, where adaptivity must also span across modalities.

Multimodal data

With the advancement of the digitalization process, research and application have begun to simultaneously involve various types of data such as images [71], audio [72], and text [73]. These data are not only high-dimensional but also show clear structural and representational differences. Text

consists of discrete symbolic sequences that contain semantic and grammatical structures; images are spatially organized pixel matrices that reflect local spatial correlations and global patterns; and audio is a continuous waveform signal that captures dynamic temporal characteristics. Compared with earlier single-type numerical or functional data, multimodal data are more heterogeneous and complex. New challenges asked researchers to explore how to establish semantic correspondences across modalities for cross-modal alignment, build joint models through shared latent spaces, and dynamically adjust inter-modal dependencies according to different tasks or contexts.

From a statistical perspective, representation learning [49] can be regarded as a generalization of traditional linear dimension reduction techniques such as Principal Component Analysis (PCA) and Factor Analysis (FA). While PCA and FA aim to identify low-dimensional subspaces that capture maximum variance or shared covariance structure through linear projections, representation learning extends this idea to nonlinear mappings that can model highly complex and structured data. Representation learning addresses this challenge by automatically learning low-dimensional embeddings that map complex observations into a shared latent space. The basic idea is to learn a nonlinear mapping

$$f_\theta : \mathcal{X} \to \mathcal{Z}$$

where $\mathcal{X}$ represents the original multimodal input space (e.g., images, text, or audio), and $\mathcal{Z}$ denotes the shared latent space. The context-adaptive nature of representation learning lies in its ability to adjust feature extraction pathways according to the input modality and to align heterogeneous data within a unified semantic space.

To infer hidden structures, traditional latent variable models in statistics such as Factor Analysis and Gaussian Mixture Models (GMM) require estimating the posterior distribution $p(z|x)$. However, when the generative process $p_\theta(x|z)$ is highly nonlinear, the posterior becomes intractable. Variational inference approximates the true posterior by introducing a tractable distribution $q(z|x)$ and minimizing their divergence, typically via the Kullback–Leibler (KL) divergence.

Amortized inference introduces a major innovation by parameterizing the approximate posterior as a learnable function $q_\phi(z|x)$ shared across all samples. This amortizes the cost of inference, enabling efficient posterior estimation via a neural network encoder. Variational Autoencoders (VAEs) [74] can be regarded as nonlinear extensions of classical latent variable models. VAEs jointly train the generative model $p_\theta(x|z)$ and the inference model $q_\phi(z|x)$ by maximizing the Evidence Lower Bound (ELBO):

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}[q_\phi(z|x) \,||\, p(z)]$$

The model can dynamically adjust posterior estimation $q_\phi(z|x)$ according to input characteristics or modality, achieving efficient and scalable probabilistic reasoning.

In many real-world scenarios, the amount of labeled data for each task may be limited. Meta-learning, or learning to learn, provides a framework for acquiring inductive knowledge that enables rapid adaptation to new tasks with few examples. The central idea is to learn a meta-model that captures transferable structures across tasks and can quickly adjust its parameters when facing a new context. It involves a two-level optimization process:

$$\min_\theta \sum_{T_i \sim p(T)} \mathcal{L}_{T_i}\big(U(\theta, T_i)\big)$$

where θ denotes the meta-parameters shared across tasks, T_i represents individual tasks sampled from a task distribution $p(T)$, and $U(\theta, T_i)$ indicates the task-specific adaptation process, such as gradient updates.

Two main paradigms have been developed. Gradient-based meta-learning, exemplified by Model-Agnostic Meta-Learning (MAML) [75], optimizes an initialization of θ that allows fast convergence on new tasks with a few gradient steps. In contrast, metric-based meta-learning [54] focuses on learning a task-invariant representation space in which samples from the same class or semantic context remain close in distance metrics.

Meta-learning can be viewed as a hierarchical model where the outer loop learns the hyperprior over tasks and the inner loop performs task-specific inference. The model internalizes the variability across tasks into shared meta-parameters and dynamically adjusts its learning strategy when exposed to new contextual distributions. Such an approach bridges the gap between multi-task learning and context-aware adaptation, providing a scalable solution for heterogeneous and data-sparse environments.

Large-scale pre-trained data

With the rapid expansion of digital ecosystems and the emergence of web-scale data collection, research has entered a new stage characterized by massive, heterogeneous, and weakly supervised datasets. In recent years, a vast amount of cross-domain data has been centrally collected—such as large-scale text corpora [76] and multimodal alignment data [77] that combine text, images, audio, and sound—representing an unprecedented scale and diversity of information sources. Unlike earlier multimodal datasets that were carefully curated for specific experimental designs, these pretraining datasets are gathered under open, non-controlled conditions and often lack explicit task labels or unified structures.

The central modeling challenge has shifted from fitting a single data distribution to extracting transferable and generalizable patterns that remain stable across heterogeneous contexts. This transition has motivated the development of foundation models and related adaptive paradigms, which aim to capture universal representations, perform context-dependent inference, and achieve robust generalization across domains.

Foundation models [78] represent a fundamental paradigm shift in machine learning toward building universal systems trained on massive and heterogeneous data sources. Unlike earlier models that were designed for specific tasks or modalities, foundation models learn general-purpose representations through large-scale pretraining, which can be adapted to downstream tasks with minimal fine-tuning. Given a massive dataset $\mathcal{D} = \{x_i\}_{i=1}^{N}$ sampled from diverse domains, the objective is to learn a parameterized function f_θ that captures broadly transferable patterns:

$$\theta^* = \arg\min_\theta \; \mathbb{E}_{x \sim \mathcal{D}} \; \ell(f_\theta(x))$$

where ℓ represents a pretraining objective such as next-token prediction or contrastive alignment.

The context-adaptive nature of foundation models arises from their ability to integrate knowledge from large and diverse data distributions, enabling emergent behaviors such as cross-modal transfer, domain generalization, and zero-shot reasoning. They thus redefine the notion of generalization—from fitting within a context to adapting across contexts—by embedding statistical invariances into large-scale learned representations.

In-context learning (ICL) [79] describes the ability of large language or multimodal models to adapt to new tasks through contextual examples provided at inference time, without updating model parameters. Given a prompt sequence $\{(x_i, y_i)\}_{i=1}^{k}$ followed by a query x_{k+1}, the model generates a prediction $\hat{y}_{k+1}$ by conditioning on the in-context information:

$$\hat{y}_{k+1} = f_\theta(x_{k+1} \mid x_1, y_1, \ldots, x_k, y_k)$$

This phenomenon suggests that pretrained models can perform implicit meta-learning within their internal representations, dynamically adjusting inference behavior according to contextual input. From a statistical perspective, ICL transforms the adaptation process from an external optimization (parameter update) to an internal inference mechanism conditioned on observed data, thus embodying a new form of context-adaptive reasoning.

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 principles 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 delineate both the potential and the pitfalls of adaptive systems.

We organize this section around six core principles: flexibility, heterogeneity signals, modularity, selectivity, data efficiency, and tradeoffs. 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 for models 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 [11]. In machine learning, meta-learning and mixture-of-experts architectures [80] 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 [11], 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 [81]. Recent work extends this further, modeling time-varying changes in continuous temporal domain generalization [82] or leveraging diversity across experts to separate stable from unstable patterns [83]. 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, transferability across tasks, and disentanglement of variation sources.

A canonical example is the mixture-of-experts framework, where a gating network routes inputs to specialized experts trained for different data regimes [80]. 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 [84], adapter libraries for large language models [85], and mixtures of LoRA experts [86]. 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 leads to overfitting, defeating the 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 [87]. Aggregation methods such as the weighted majority algorithm show how selective weighting of multiple models can improve robustness [88]. In modern machine learning, Bayesian rules can activate test-time updates only when uncertainty is manageable [89], while confidence-based strategies prevent unstable adjustments by holding back adaptation under weak signals [90]. Sparse expert models apply the same principle architecturally, activating only a few experts for easy inputs but engaging more capacity for difficult cases [91]. 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 transcends individual methods and reflects a general statistical truth.

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 [92]. Bayesian analyses of scaling laws for in-context learning formalize how the reliability of adaptation grows with data [93]. To mitigate these limits, modular reuse strategies have been developed, including adapter libraries [85] 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 brings benefits yet inevitably incurs 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 [94]: 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 [95]. 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 imply that models lack adaptivity, but that adaptation proceeds in unstable or unjustified ways.

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 [81,96]. 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 [92]. 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 [97]. Similarly, analyses of mixture-of-experts models show that mis-specified routing can cause experts to collapse or remain underutilized [95]. 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 [98]. 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.
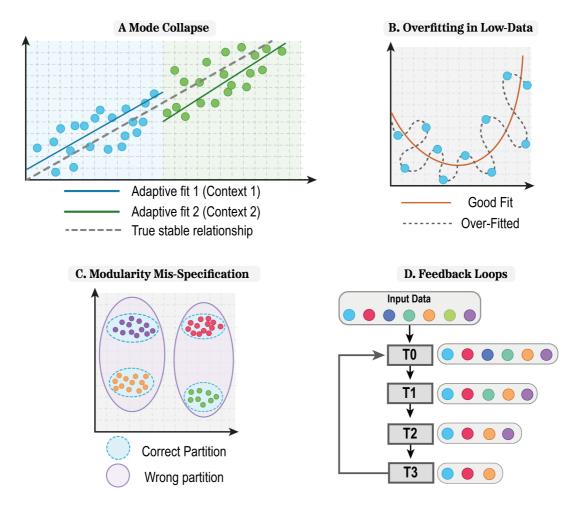

Figure 5: Failure Modes of Context-Adaptive Models. (A) **Mode Collapse**: adaptive fits diverge from the true stable relationship. (B) **Overfitting in Low-Data Contexts**: adaptation follows noise rather than signal. (C) **Modularity Mis-Specification**: incorrect partitions obscure the true structure. (D) **Feedback Loops**: adaptive decisions reshape the very data they rely on.

Having examined when and why adaptivity fails, we now synthesize these insights into a set of guiding principles for practical model design.

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 interpretable. Building on these conceptual principles, we next examine how context-adaptive inference can be made computationally and statistically efficient.

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 [99], hierarchical regularization that applies different regularization strengths to different levels of context specificity [100,101], 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 [102], while context-dependent sampling uses different sampling strategies for different contexts [103]. Caching and warm-starting leverage solutions from similar contexts to accelerate optimization, particularly effective when contexts exhibit smooth variation [104].

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 [105].

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 [106,107,108]. 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 [93,109,110]. 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 [102]. 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 $\mathcal{C}$, and suppose the per-context parameter is $\theta(c) \in \Theta$.
For observation (x, y, c), consider a conditional model $p_\theta(y \mid x, c)$ with loss $\ell(\theta; x, y, c)$.
For a context neighborhood $\mathcal{N}_\delta(c) = \{c' : d(c, c') \leq \delta\}$ under metric d, define the effective sample size available to estimate $\theta(c)$ by

$$N_{\text{eff}}(c, \delta) = \sum_{i=1}^{n} w_\delta(c_i, c), \quad w_\delta(c_i, c) \propto K\left(\frac{d(c_i, c)}{\delta}\right), \quad \sum_i w_\delta(c_i, c) = 1,$$

where K is a kernel.
A kernel-regularized estimator with smoothness penalty

$$\mathcal{R}(\theta) = \int \|\nabla_c \theta(c)\|^2 \, dc$$

solves

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; x_i, y_i, c_i) + \lambda \, \mathcal{R}(\theta).$$

Assuming local Lipschitzness in c and L-smooth, μ-strongly convex risk in θ, a standard bias–variance decomposition yields for each component j

$$\mathbb{E}\left[\|\hat{\theta}j(c) - \theta_j(c)\|^2\right] \lesssim \underbrace{\frac{\sigma^2}{N_{\text{eff}}(c, \delta)}}_{\text{variance}} + \underbrace{\delta^{2\alpha}}_{\text{approx. bias}} + \underbrace{\lambda^2}_{\text{reg. bias}}, \quad \alpha > 0,$$

which exhibits the adaptivity–data trade-off: finer locality (small δ) increases resolution but reduces N_{eff}, inflating variance.
Practical procedures pick δ and λ to balance these terms (e.g., via validation), and amortized approaches replace $\theta(c)$ by $f_\phi(c)$ with shared parameters ϕ to increase N_{eff} through parameter sharing. For computation, an early-stopped first-order method with step size η and $T(c)$ context-dependent iterations satisfies (for smooth, strongly convex risk) the bound

$$\mathcal{L}(\theta^{(T(c))}) - \mathcal{L}(\theta^\star) \leq (1 - \eta\mu)^{T(c)} \left(\mathcal{L}(\theta^{(0)}) - \mathcal{L}(\theta^\star)\right) + \frac{\eta L \sigma^2}{2\mu \, N_{\text{eff}}(c, \delta)}.$$

linking compute allocation $T(c)$ and data availability $N_{\text{eff}}(c, \delta)$ to the attainable excess risk at context c.

Formal optimization view of context-aware efficiency

Let $f_\phi : \mathcal{X} \times \mathcal{C} \to \mathcal{Y}$ be a context-conditioned predictor with shared parameters ϕ.
Given per-context compute budgets $T(c)$ and a global regularizer $\Omega(\phi)$, a resource-aware training objective is

$$\min_\phi \ \mathbb{E}_{(x,y,c)\sim\mathcal{D}}\ell\big(f_\phi(x, c), y\big) + \lambda\,\Omega(\phi) \quad \text{s.t.} \quad \mathbb{E}_c\,\mathcal{C}\big(f_\phi; T(c), c\big) \leq B,$$

where $\mathcal{C}(\cdot)$ models compute or latency.
The Lagrangian relaxation is

$$\min_\phi \ \mathbb{E}_{(x,y,c)}\ell\big(f_\phi(x, c), y\big) + \lambda\,\Omega(\phi) + \gamma\,\mathbb{E}_c\,\mathcal{C}\big(f_\phi; T(c), c\big),$$

which trades off accuracy and compute via γ.
For mixture-of-experts or sparsity-inducing designs, let $\phi = (\phi_1, \ldots, \phi_M)$ and define a gating function $\pi_\phi(m \mid c)$.
A compute-aware sparsity penalty can be written as

$$\Omega(\phi) = \sum_{m=1}^{M} \alpha_m\,\|\phi_m\|_2^2 + \tau\,\mathbb{E}_c \sum_{m=1}^{M} \pi_\phi(m \mid c),$$

encouraging few active modules per context.
Under smoothness and strong convexity, the optimality conditions yield the KKT stationarity conditions

$$\nabla_\phi\big(\mathbb{E}\,\ell + \lambda\,\Omega + \gamma\,\mathbb{E}_c\,\mathcal{C}\big) = 0, \quad \gamma\,\big(\mathbb{E}_c\,\mathcal{C} - B\big) = 0, \quad \gamma \geq 0.$$

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.

Together, these efficiency principles and formal analyses bridge conceptual foundations with implementation. In the next section, we turn to explicit adaptive models that instantiate these ideas through structured parameterization and estimation.

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 [11].

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 6).
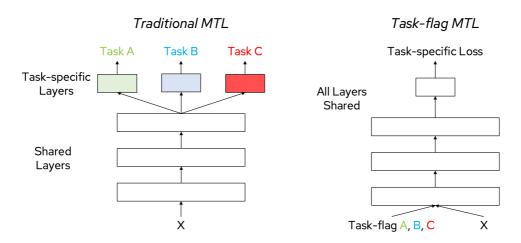

Figure 6: Multi-task learning as explicit adaptivity. In traditional MTL (left), each task has its own head on top of shared layers. In context-flagged models (right), the task identity is provided as an input, enabling a shared model to adapt across tasks.

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 [11,111]. 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. These estimators occupy the lower-capacity but more interpretable end of the explicit adaptivity spectrum, forming a conceptual baseline for more complex architectures discussed below.

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 [112]. This principle was formalized in Lauritzen's (1996) influential monograph, which systematized probabilistic graphical models and showed how independence assumptions can be embedded into estimation procedures [113]. 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 [114]. Shortly thereafter, Friedman, Hastie, and Tibshirani (2008) developed the graphical lasso, a convex penalized likelihood method that directly estimates sparse precision matrices [115]. 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 [116]. Danaher, Wang, and Witten (2014) extended this framework with the Joint Graphical Lasso, which balances shared structure and context-specific edges across multiple populations [117]. 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 explicit splits of the context space can emulate distinct tasks, clarifying where parameters should be shared or separated. 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 7).

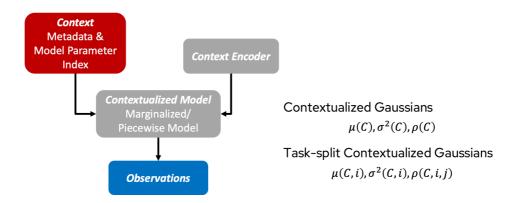## Hierarchical Encoding of Context Enables Multi–Level Adaptivity

Figure 7: 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.

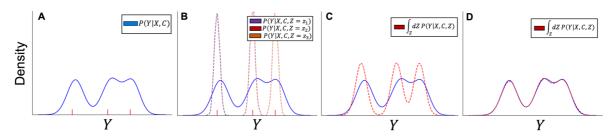## Nonparametric inference from context–adaptive parameters

Figure 8: 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 has known spatial or network structure, regularization terms can promote similarity among neighboring coefficients or nodes. 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 [118,119]. On networked data, the network VCM of [120] 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. These regularization principles can also be extended to temporal, hierarchical, or multilevel contexts, where smooth transitions or cross-level coupling may be encoded through Laplacian penalties or nested-group regularizers tailored to the structure of c.

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, [121] introduced Bayesian inference for GGMs with lattice structure, demonstrating how spatial priors can capture context-dependence across neighboring sites. Building on this idea, [122] 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, [123] 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

As context dimensionality and data complexity grow, explicit smoothness assumptions become insufficient, motivating high-capacity learners that approximate $f(c)$ directly from data. 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 [124]. 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 [125].

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 [11]. 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 [126], 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, [120] 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. This perspective reframes missingness from a nuisance into structured signal: it encodes which sources of evidence are available and how they should be combined, reflecting ideas explored in recent multimodal learning frameworks that handle missing modalities [127].
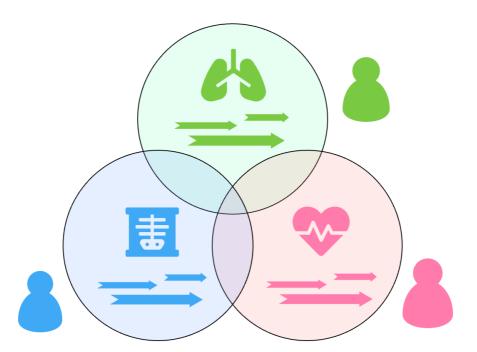

Figure 9: Patterns of missingness as context. Each dataset (e.g., cohort with labs, cohort with vitals, cohort with imaging) provides a different subset of measurements. Context-adaptive models allow integration by conditioning on measurement availability, enabling learning from fewer samples with more heterogeneous features.

Figure 9 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 [81,97]. 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 [81,128]. 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 $p(x_{\text{missing}} \mid x_{\text{observed}})$ [129,130]. For causal or estimation applications, conduct *ignorability sensitivity analyses*, contrasting MAR-based results with pattern-mixture or selection-model analyses under plausible MNAR mechanisms [131,132]. 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 [133,134]. Practical implementations of these ideas are widely available: **GRU-D** [135] and **BRITS** [136] provide mask- and time-aware sequence models, while **GAIN** [130] and **VAEAC** [129] 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 [137].

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 [99], hierarchical regularization that applies different strengths of shrinkage to varying levels of context specificity [101], 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 [102], while context-dependent sampling employs different sampling schemes across contexts [103]. Caching and warm-starting further accelerate optimization by leveraging solutions from similar contexts, particularly effective when contexts exhibit smooth variation [104].

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. These trade-offs should be evaluated in light of application-specific requirements. For example, advanced adaptive optimizers like Adam can efficiently train complex, nonlinear models, but the resulting systems may be less interpretable than simpler alternatives [105]. In practice, such context-dependent computation appears in adaptive batching, per-context learning rates, and multi-fidelity optimization pipelines that dynamically adjust compute and precision depending on context complexity.

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 [111]. 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], exhibit forms of adaptation that do not arise from any predefined mapping. Instead, their flexibility emerges from the interaction between model structure and the breadth of training data—an effect we refer to as implicit adaptivity. 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. This emergent phenomenon, referred to as implicit adaptivity, highlights how learning and inference can become intertwined within the model itself. Attention

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

In contrast to explicit parameter mapping, the simplest route to implicit adaptation is to feed context directly as part of the input. 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 [138].

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 [139].

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. By enabling dynamic, input-dependent weighting, attention supports context-aware computation without altering global parameters, thereby setting the stage for advanced on-the-fly adaptation such as 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 [140]. 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 [141].

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 [75]. 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. In this view, the capacity to adapt becomes encoded in the meta-learned parameters themselves, enabling rapid generalization from few examples without a hand-crafted map from context to coefficients and standing in clear contrast to explicit approaches.

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 [106], 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. Subsequent work has shown that the ability to generalize from few in-context examples can itself be enhanced through meta-training. MetaICL explicitly trains models across diverse meta-tasks, teaching them to infer and adapt within context at test time without gradient updates, thereby strengthening the implicit adaptability of large language models [142].

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 [107].

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 [108].

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 [143,144].

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 [145]. A complementary theoretical development interprets in-context learning as a rational adaptation process. From a Bayesian decision-theoretic standpoint,

transformers can be viewed as implicitly balancing expected loss with strategy complexity, thereby achieving near-optimal adaptation under computational constraints [146]. This rational framing connects implicit adaptivity with classical principles of statistical inference.

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 [147]. Extending this mechanistic line, Dherin et al. (2025) demonstrate that stacking self-attention and MLP layers allows transformers to implicitly update internal representations during a single forward pass, effectively realizing dynamic context-specific weight adjustments without explicit training [148]. Such implicit internal updates offer a concrete mechanistic account of how context-dependent behavior arises.

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 [79].

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 research has demonstrated that transformers trained on simple regression tasks can learn to perform ridge or kernel regression entirely within their forward pass, without any explicit supervision to do so.

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 [149].

Building on this finding, von Oswald et al. (2023) show that gradient-based training of transformers over distributions of regression tasks leads them to perform in-context gradient descent, effectively realizing kernel regression with the learned attention kernel serving as $k(c_i, c_j)$ [150]. Garg et al. (2023) further analyze which function classes can be learned in-context, demonstrating that transformers can approximate a wide family of kernel smoothers when trained on synthetic regression tasks [151].

Dai et al. (2023) provide a complementary theoretical view, arguing that transformers can implicitly implement compositional function families through their attention layers, and that in-context learning arises naturally from this functional representation [143].

Finally, Reuter et al. (2025) propose a compelling Bayesian interpretation: transformers trained under in-context learning can perform full Bayesian inference for common statistical models such as generalized linear models and latent factor models. Concretely, they train transformers to infer complex posterior distributions in context, showing that the in-context forward pass can approximate posterior sampling comparable to MCMC or variational inference methods [152].

In all these cases, the support set within the prompt plays an analogous role to the neighborhood in context space in varying-coefficient models. The query token's prediction is formed by aggregating information from the support tokens via learned similarity weights, realized by the attention mechanism rather than an explicitly defined kernel function.

Synthesis: Two Paths to the Same Estimators

Taken together, these results reveal a common form:

$$\widehat{f}\left(x^{*}, c^{*}\right) = \sum_{i=1}^{n} \alpha_{i}(c^{*})\, y_{i},$$

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.

Both perspectives yield estimators of the same functional form, with explicit kernel weighting in VCMs and learned attention weighting in transformers. This correspondence motivates a unified view of context-adaptive inference, combining the interpretability of explicit modeling with the flexibility and scale of implicit 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. This unified perspective is also extending to structured and tabular domains. TabICL introduces a foundation model architecture for large-scale tabular data, showing that in-context learning can efficiently scale to structured datasets via column-row attention mechanisms [153]. These results suggest that implicit adaptivity generalizes beyond text or vision into the broader landscape of structured scientific data.

Comparative Synthesis: Implicit versus Explicit Adaptivity

Implicit and explicit strategies reflect two complementary philosophies for modeling heterogeneity, each with distinct strengths and trade-offs. 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. Prompting strategies such as chain-of-thought demonstrate that structured context can sometimes steer internal computation, providing limited but useful handles on model behavior [154]. 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 [155], 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 [156]. 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.

Toward Explicit Modeling of Implicit Adaptivity: Local Models, Surrogates and Post Hoc Approximations

Motivation

Building on the prior discussion of implicit adaptivity, this section examines 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 obscures the underlying mechanisms of adaptation, hindering modular reuse and systematic auditing. 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 [157,158].
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**.
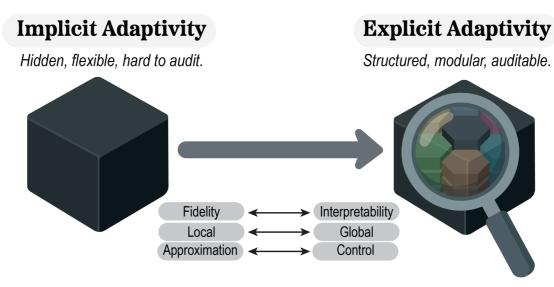
Figure 10: From Implicit to Explicit Adaptivity. A black-box model (left) represents implicit adaptation, which is hidden and opaque. Making adaptivity explicit (right) exposes structured components that can be inspected and controlled. The axes below highlight the trade-offs between fidelity and interpretability, 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

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

where $\mathcal{N}_{x_0,c_0}$ defines a locality (e.g., kernel weights), ℓ measures fidelity, and Ω controls complexity. Where $\mathcal{G}$ denotes a restricted hypothesis class, often composed of linear or other low-complexity functions chosen to enhance interpretability. 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 [159]; SHAP / DeepSHAP provide additive attributions based on Shapley values [160]. Integrated Gradients and DeepLIFT link attribution to path-integrated sensitivity or reference-based contributions [161,162]. 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 [163]. Deep k-nearest neighbors audits predictions by querying neighbors in activation space and can flag distribution shift [164]. Influence functions link a prediction to influential training points for data-centric debugging [165]. This line of work connects naturally to exemplar models and contextual bandits, where decisions are justified via comparisons to context-matched exemplars. Reports include prototype coverage and diversity, neighbor quality checks, and the effect of editing prototypes or influential examples. These prototype-based approaches make local adaptation explicit by grounding predictions in reference cases, bridging the gap between black-box models and case-based reasoning frameworks.

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 [166]. 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. Such diagnostics mirror classical checks for approximate Bayesian inference, where amortization gaps quantify the discrepancy between learned and exact posteriors.

Disentangled and Bottlenecked Representations

While amortization diagnostics target model faithfulness, disentanglement aims to expose interpretable subspaces aligned with distinct contextual factors. The aim is to expose factors that align with distinct contextual causes, making changes traceable and controllable. β-VAE encourages more factorized latents [167], while the Deep Variational Information Bottleneck promotes predictive minimality that can suppress spurious context [168]. Concept-based methods such as TCAV and ACE map latent directions to human concepts and test sensitivity at the concept level [169,170]. Fully unsupervised disentanglement is often ill-posed without inductive bias or weak supervision [171]. Quantitative evaluation of disentanglement can follow established metrics that assess factor independence, completeness, and informativeness [172]. 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 [173]; edge probing examines specific linguistic structure in contextualized representations [174]. Model editing methods such as ROME can modify stored factual associations directly in weights [175], while "knowledge neurons" seek units linked to particular facts [176]. Evaluation involves quantifying pre- and post-edit behavior, assessing locality and persistence, and documenting side effects on unrelated capabilities. Collectively, these methods transform hidden internal adaptations into analyzable modular components.

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 [177]. 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 [158].

These methodological families differ in their assumptions and computational granularity, yet they all aim to render adaptation transparent and controllable. The following sections summarize their key trade-offs and conceptual challenges.

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

$$\min_{g \in \mathcal{G}} \quad \underbrace{\phi_{\mathrm{fid}}(g; U)}_{\text{faithfulness on use set } U} + \lambda$$

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

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.

These trade-offs are not merely design choices but determine the operational boundaries within which explicit representations can remain faithful to the original adaptive system.

Open Research Directions

Reusable Modules

The challenge of isolating reusable routines parallels the quest for parameter-efficient fine-tuning in large models, where adaptation must remain modular yet composable. 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). Benchmarking under fixed capacity and identical data regimes is essential to isolate the contribution of explicit structure from mere model scaling effects. 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). Selecting the proper level of abstraction thus defines not only interpretability but also the feasible scope of control.

Evaluation and Reporting Standards for 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 [157,158]. Carmichael & Scheirer (2021) further propose a principled evaluation framework for feature-additive explainers, enabling measurement of misattribution even under known ground-truth additive models [178].

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 [157,158]. Turbé et al. (2023) demonstrate evaluation of interpretability methods on time-series models using metrics such as $\widetilde{\text{AUC}}_S$ and $\widetilde{F_{1,S}}$ to compare alignment with model internals [179].

Minimal reporting checklist

Item	Description
Data slice and context definition	Specify the subset of data and contextual variables used for generating explanations, and describe the locality or neighborhood definition.
Surrogate specification and regularization details	Report the family of surrogate models, chosen regularization strategy, and kernel or sampling parameters.
Faithfulness and robustness metrics	Include local R^2, deletion/insertion area, counterfactual validity, and robustness under perturbations.
Sensitivity and uncertainty analysis	Assess variation across baselines, random seeds, and small input perturbations, providing uncertainty estimates.
Computational constraints	Document runtime, hardware limitations, and approximation budgets that affect explanation quality.
Observed limitations and failure modes	Summarize known weaknesses, unstable regions, or interpretability failures identified during validation.

Table 2. Minimal Reporting Checklist for Post-hoc Explanations

From post hoc analysis to design

Insights from post-hoc analysis can inform proactive model design for control, auditing, and policy communication. In such cases, interpretability methods should not remain external diagnostics but serve as guides for architectures with built-in transparency. For example, Concept Bottleneck Models integrate interpretable concepts into the forward pass [180]. Similarly, Poursabzi-Sangdeh et al. (2021) conduct empirical user studies to highlight how interpretability design choices affect human use and model trust [181]. These contributions extend the vision of Doshi-Velez & Kim (2017) toward a unified

science of interpretable modeling, where explanation and model training are co-designed [157]. Taken together, these lines of work bridge black-box adaptation and structured inference and set the stage for designs where context-to-parameter mappings are specified, trained, and evaluated 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. Reinterpreting these classical estimators through the lens of explicit adaptivity situates them as early instances of structured context modeling, underscoring continuity between statistical modeling and modern machine learning.

Taken together, these strategies illustrate a gradual unification of interpretability, modularization, and adaptive modeling, paving the way toward a principled science of explicit context-aware inference.

Context-Invariant Training: A View from the Converse

While the preceding sections emphasize the importance of modeling context to tailor predictions, an equally fundamental question concerns robustness: Can we learn representations such that a single predictor performs reliably across sites, cohorts, and time, despite environmental shifts and nuisance variation? Context-invariant training aims at out-of-distribution (OOD) generalization by emphasizing features whose associations with the target remain stable across environments, while suppressing spurious correlations that vary with nuisance contexts. Standard Empirical Risk Minimization (ERM) [182] 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 seminal framework connecting modern deep learning to invariant prediction is Invariant Risk Minimization (IRM) [183], which formulates robustness as learning causally stable predictors across multiple environments. IRM seeks a representation Φ such that a shared predictor w minimizes the risk $R^e(\cdot)$ for every environment e. The original formulation is a bi-level optimization problem that is computationally intractable. To make it solvable, Arjovsky et al. propose a surrogate version, IRMv1, which introduces a penalty ensuring that the per-environment risk gradient vanishes for a shared dummy classifier $w = 1$, thereby enforcing stationarity across environments. This construction connects invariance to out-of-distribution (OOD) generalization by encouraging predictors aligned with causal mechanisms that persist across environments.

However, subsequent analyses revealed important limitations. In linear settings, IRM often fails to recover the true invariant predictor, and in nonlinear regimes, performance can deteriorate sharply when test distributions deviate from the training domains [184]. This undermines IRM's objective of handling distribution shift, where $P(X)$ changes while $P(Y|X)$ remains fixed. Thus, IRM offers no mechanism to reduce sensitivity when those shifts are amplified at test time. To mitigate these issues, Risk Extrapolation (REx) [185] extends the principle of invariance by optimizing directly over per-environment risk vectors. Two practical variants have been proposed: MM-REx and V-REx, which performs robust optimization over affine combinations of the environment risks (weights sum to 1, possibly negative), and V-REx, which minimizes the mean risk augmented by the variance of risks across environments.

Unlike IRM, which requires explicit environment labels, Beery et al. (2018) [186] propose CoRe, a method that assumes some samples share a common identifier. Features are decomposed into core

components (whose class-conditional distribution is stable across domains) and style components (e.g., brightness, pose) that vary with domains. The CoRe estimator enforces robustness by penalizing the conditional variance of the loss within groups sharing the same label–identifier pair (Y, ID).

Adversarial Robustness as Context-Invariant Training

Whereas IRM seeks robustness across discrete environments, adversarial robustness can be regarded as its infinitesimal counterpart—focusing on perturbations within a local neighborhood of each input rather than across distinct domains. 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 [187]. 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 $\|\delta\|_p \leq \varepsilon$) 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 [188] 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. Madry et al. (2018) introduce Projected Gradient Descent (PGD) as a universal first-order adversary. The generated perturbations are incorporated into the training process to improve robustness under local contextual 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. Formally, both IRM and adversarial training minimize performance variance across contexts—IRM across discrete environments, and adversarial training across continuous perturbation neighborhoods.

[189] 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 [190].

This perspective is directly applicable to the challenges faced by LLM-based Agents as surveyed in [191]. An autonomous agent does not operate in a sterile, curated dataset; it operates in the wild. These fine-grained, synthetic perturbations provide a useful abstraction for understanding the robustness challenges faced by LLM-based agents:

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.

Hence, advances in adversarial robustness directly inform the design of safer, more context-stable autonomous agents.

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 [192]. 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) [193]. 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 [194]. 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. Collectively, these approaches demonstrate a continuum from fully supervised environment-aware optimization to self-supervised environment discovery, unified under the goal of achieving context-invariant generalization. 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

This section surveys how context-adaptive methods manifest across domains, how their performance is assessed, and what tools enable them in practice.

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 [195,196]. 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 [197]. Such models often employ hierarchical priors to borrow strength across patients while maintaining individualized inference. 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 [198,199]. 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 [200]. 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. The approximate factorization reduces complexity from cubic to linear in the number of assets, making large-scale forecasting feasible.

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 based on volatility regimes and transaction costs, as studied in regime-switching portfolio models [201].

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 [202]. 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 for Evaluating Context-Aware Performance

Building on the theoretical framework introduced in earlier sections, we now formalize the evaluation criteria used to quantify context-adaptive behavior.
These metrics capture predictive accuracy, adaptation efficiency, transferability, and robustness under contextual variation.

Let $\mathcal{C}$ denote the context space and $\mathcal{D}_{\text{test}}$ a test distribution over (x, y, c).
For a predictor $\hat{f}$, define the context-conditional risk as

$$\mathcal{R}(\hat{f} \mid c) = \mathbb{E}\left[\ell\big(\hat{f}(x, c), y\big) \,\big|\, c\right], \qquad \mathcal{R}(\hat{f}) = \mathbb{E}_{c \sim \mathcal{D}_{\text{test}}}\left[\mathcal{R}(\hat{f} \mid c)\right].$$

A context-stratified evaluation reports $\mathcal{R}(\hat{f} \mid c)$ across predefined bins or via a smoothed estimate $\int \mathcal{R}(\hat{f} \mid c)\, \mathrm{d}\Pi(c)$ for a reference measure Π that weights regions of the context space.

Adaptation Efficiency

To evaluate how rapidly a model benefits from in-context examples,
let $S_k(c) = \{(x_j, y_j, c)\}_{j=1}^{k}$ denote k examples available within context c.
Define the adaptation efficiency as

$$\mathrm{AE}_k(c) = \mathcal{R}(\hat{f}_0 \mid c) - \mathcal{R}(\hat{f}_{S_k} \mid c), \qquad \mathrm{AE}_k = \mathbb{E}_c[\mathrm{AE}_k(c)],$$

where $\hat{f}_0$ is the non-adapted baseline and $\hat{f}_{S_k}$ the adapted predictor.

The function $k \mapsto \mathrm{AE}_k$ summarizes few-shot adaptation gains across different context sizes.

Transfer Performance

Transfer across source and target contexts, $\mathcal{C}_{\mathrm{src}} \to \mathcal{C}_{\mathrm{tgt}}$,
with shared representation ϕ, can be measured by

$$\mathrm{TP}(\phi) = \mathcal{R}_{\mathcal{C}_{\mathrm{tgt}}}(\hat{f}_\phi) - \mathcal{R}_{\mathcal{C}_{\mathrm{tgt}}}(\hat{f}_{\mathrm{scratch}}),$$

quantifying performance retained when transferring ϕ from source to target contexts compared with training from scratch.

Robustness to Context Shift

To assess stability under distributional perturbations,
let Q denote a family of admissible context shifts (e.g., f-divergence or Wasserstein balls over context marginals).
Then the robustness score is defined as

$$\mathrm{RS}(\hat{f}; Q) = \sup_{\widetilde{\mathcal{D}} \in Q} \left[\mathcal{R}_{\widetilde{\mathcal{D}}}(\hat{f}) - \mathcal{R}_{\mathcal{D}_{\mathrm{test}}}(\hat{f}) \right],$$

where higher values indicate greater sensitivity to contextual changes.

These metrics provide a unified quantitative view of context-aware performance.
They complement the theoretical efficiency results developed in Section 4
and serve as practical diagnostics for evaluating real-world adaptivity across diverse applications.

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 [202].

Industrial applications derive clear benefits from context-aware efficiency. In predictive maintenance, systems adapt maintenance schedules using equipment context such as age, usage history, and environmental conditions. For example, recent surveys of predictive maintenance in Industry 4.0

identify architectures that integrate sensor data, remaining-useful-life models, and context-aware scheduling policies [203,204]. In quality control, context-dependent sampling directs inspection efforts to high-risk units, reducing waste and computational cost. Inventory management likewise benefits from context-aware forecasting models that incorporate demand volatility, seasonality, and external signals; recent work shows that such approaches outperform traditional forecasts in retail settings [205].

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.

Such contextualized networks have been applied in TCGA cancer genomics to identify patient-specific driver modules.
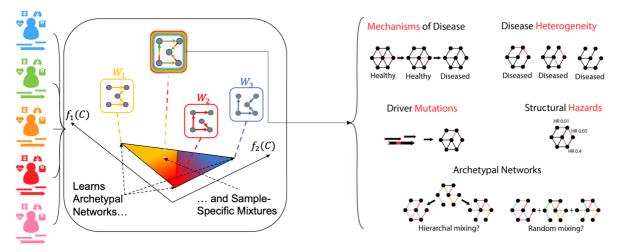
Figure 11: Contextualized networks enable inference of archetypal and sample-specific mixtures, unlocking new biological insights such as mechanisms of disease, disease heterogeneity, structural hazards, and driver mutations.

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 [206,207]. 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 [208,209]. 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 [210].

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 [211]. 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. Hence, evaluation protocols must jointly assess accuracy, calibration, and transferability under context perturbations.

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 [212]. Beyond improving factuality, RAG also underpins adaptive behavior in AI-generated content, where external retrieval reduces hallucination and provides domain-specific grounding [213]. 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 [214]. 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 [215]. 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 [216]. 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 [217,218]. 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 not only on technical choices but also on aligning model adaptivity with domain-specific interpretability and governance requirements.

Future Trends and Opportunities with Foundation Models

A New Paradigm for Context-Adaptive Inference

Recent advances in large-scale foundation models have fundamentally reshaped the landscape of context-adaptive inference. Trained on vast and diverse datasets with self-supervised objectives, these models internalize broad statistical regularities across language, vision, and multimodal data [78]. Unlike earlier approaches that relied on hand-crafted features or narrowly scoped models,

foundation models can process and structure complex, high-dimensional contexts that were previously intractable.

Their impact is clear in natural language processing, where large language models achieve strong zero-shot and few-shot generalization, and in computer vision, where multimodal encoders such as CLIP align images and text into a shared representation space [77]. These advances mark a shift from treating feature extraction and inference as separate stages toward unified systems that function simultaneously as representation learners and adaptive engines. At the same time, challenges remain, including high computational demands, the risk of amplifying societal biases, and the difficulty of interpreting learned representations [219].

To understand their contribution to context-adaptive inference, we consider three dimensions: their role as universal context encoders, the mechanisms enabling dynamic adaptation, and their integration with formal statistical and causal reasoning.

Universal Context Encoders

Foundation models act as general-purpose context encoders, transforming raw, unstructured data into meaningful representations without manual feature engineering. For textual data, models such as BERT learn embeddings that capture semantic and syntactic nuances, supporting tasks from classification to retrieval [76]. For visual and multimodal inputs, CLIP aligns images and text into a shared embedding space, enabling zero-shot classification and cross-modal retrieval [77].

These representations effectively serve as context variables—latent, structured features that can feed directly into statistical models. Classical approaches such as regression or causal inference can thus operate on data that would otherwise remain unstructured. This capacity forms the basis for integrating representation learning with formal frameworks of context-adaptive inference.

Dynamic Adaptation Mechanisms

Foundation models enable dynamic adaptation primarily at inference time, allowing models to respond to new tasks without retraining. The most prominent mechanism is in-context learning (ICL), where models adapt behavior by conditioning on examples in a prompt, enabling rapid few-shot or zero-shot generalization [151].

Scaling is supported by modular architectures such as Mixture-of-Experts (MoE), which route inputs to specialized sub-networks for sparse activation, increasing capacity without proportional compute [220]. Parameter-efficient fine-tuning (PEFT) methods such as LoRA show that models can be adapted by updating less than one percent of weights, achieving near full fine-tuning performance [8].

Together, these approaches illustrate how adaptation can be achieved both flexibly and efficiently, balancing generalization and computational constraints.

Bridging with Statistical and Causal Reasoning

An emerging research direction integrates the representational capacity of foundation models with the rigor of statistical and causal inference. Language models can already extract relational patterns from text to propose or critique causal graphs [221]. Methods such as LMPriors treat foundation models as task-specific priors, improving sample efficiency in estimation and decision making [222]. Models can also generate natural-language rationales that clarify predictions and summarize statistical findings, enhancing interpretability and transparency [177].

Consequently, foundation models serve as bridges between flexible representation learning and principled inference, offering a path toward adaptive systems that are both data-efficient and theoretically grounded.

Next-Generation Methods for Contextualized Adaptive Inference

While current foundation models already enable impressive forms of adaptivity, the next phase of research looks toward methods that will shape the future of contextualized adaptive inference. These directions point ahead, emphasizing how models may be adapted, combined, and evaluated. The aim is not only greater power, but also more transparency and reliability in high-stakes settings. We highlight three forward-looking methodological trends: modular fine tuning and compositional adaptation, mechanistic insights into in-context learning, and new frameworks for reliability and calibration.

Modular Fine-Tuning and Compositional Adaptation

Parameter-efficient fine-tuning approaches such as adapters and LoRA show that large models can be customized by updating only a small subset of parameters while preserving pretrained knowledge [8]. Future systems are expected to expand these ideas into compositional strategies, dynamically combining specialized modules optimized for different domains or contexts [223].

Recent findings suggest that merging multiple LoRA modules can even outperform full fine-tuning, signaling a paradigm where adaptation arises from modular reuse rather than retraining [224]. Compositional adaptation thus points toward building libraries of reusable context-specific skills that can be flexibly assembled for new tasks.

In-Context Learning and Mechanistic Insights

Although in-context learning has revolutionized generalization, its internal mechanisms remain partly opaque. Evidence suggests that transformers may implement optimization-like updates during forward passes, effectively performing implicit gradient descent when processing examples [150]. Other analyses interpret ICL as implicit Bayesian inference, where the prompt provides evidence that reshapes the predictive distribution [225].

Mechanistic studies further identify induction heads within transformer attention circuits as critical components for pattern induction and few-shot generalization [147]. Such insights are expected to inspire architectures that enhance both transparency and stability in adaptive learning.

Reliability, Calibration, and Context-Sensitive Evaluation

As adaptive models become more flexible, ensuring calibration and reliability across shifting contexts becomes crucial. Deep neural networks, including LLMs, are often miscalibrated, producing overconfident probabilities misaligned with true accuracy [226].

Future research will increasingly embed uncertainty quantification into adaptive pipelines through deep ensembles, Bayesian ensembling, or conformal prediction to produce valid confidence intervals [81]. Evaluation protocols must also stress robustness under distributional shifts, testing whether models can sustain performance and express uncertainty under novel or adversarial conditions [227].

By embedding calibration and robustness within design, adaptive inference can evolve toward a more trustworthy, auditable, and context-aware standard.

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 [78]. Their flexibility, scalability, and cross-domain generalization have transformed statistical modeling and data analysis.

LLMs such as GPT-4 [228] and LLaMA-3.1 [229] exemplify this progress, achieving state-of-the-art results in language understanding, summarization, and reasoning. Beyond NLP, foundation models extend to multimodal tasks [77], text embeddings [76], and even tabular and structured data [230].

Adaptivity in these systems is largely realized through prompting, which conditions responses on user-provided context without additional fine-tuning [231]. Meanwhile, Mixture-of-Experts (MoE) architectures enhance scalability by routing computation to relevant submodels for efficiency [220].

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 [234]. 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 [151]. These embeddings significantly enhance the training of both complex, black-box models [235] and simpler statistical methods like n-gram-based analyses [236], 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 [237].

Recent Innovations and Outlook

Several new architectures exemplify how foundation models advance context-sensitive inference through modularity and interpretability:

FLAN-MoE (Fine-tuned Language Model with Mixture of Experts) [238] 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) [222] 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) [222] 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.

Collectively, these directions suggest a future in which foundation models evolve from passive representation learners into active, context-sensitive inference engines that unify adaptivity, efficiency, and interpretability within a principled framework.

Open Problems

Rapid advances in context-adaptive modeling have created unprecedented opportunities while revealing fundamental challenges. This chapter identifies the central methodological questions and the broader ethical and societal challenges that will shape the future trajectory of context-adaptive inference. We begin by examining five interrelated technical questions—on modularity, the benefits of explicit structure, the level of abstraction, theoretical and practical barriers, and interpretability trade-offs—that together define the frontier of adaptive modeling research. 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 open problems span five domains: (i) modularity and reusability of adaptive components, (ii) the conditions under which explicit structure improves robustness and generalization, (iii) the appropriate level of abstraction for intervention, (iv) theoretical, computational, and data-related barriers to adoption, and (v) the tension between interpretable-by-design and post-hoc interpretability. Together, these questions delineate a research agenda that bridges theoretical statistics, machine learning, and applied modeling, combining methodological depth with practical impact.

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 [78]. 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. Promising approaches operate at different levels: (i) representation-level constraints such as concept bottlenecks enforcing human-understandable

features; (ii) memory-based mechanisms such as prototype libraries for case retrieval; and (iii) architecture-level designs such as sparse adapters or routing networks that activate context-relevant modules [86,239].

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 or instability in generalization, and poorly aligned components may propagate or amplify existing 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 [229].

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 [81]. 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 [97]. 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 [143]. Concept-level interventions provide stability and interpretability but may fail to capture the model's internal computations in full detail [240].

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 [97]. 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 [111].

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 [11]. 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 [78].

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

Emerging paradigms such as Agentic Context Engineering (ACE) push this vision further by treating the context itself as an adaptive, evolving entity. In this framework, language models continuously refine and regenerate their own contexts through feedback, reflection, and planning, enabling self-improving adaptation cycles across time [241]. 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 such as healthcare and finance must satisfy principles of interpretability, auditability, and accountability to prevent harm and maintain public trust [78]. Collaboration between regulators, practitioners, and researchers is essential to establish transparent auditing standards and verifiable documentation for adaptive decisions.

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 [98]. 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

This review established a unifying framework for understanding context-adaptive inference across both explicit statistical models and implicit adaptation in modern foundation models. By tracing how adaptation appears in parameterized functions such as varying-coefficient models and in emergent processes like in-context learning, we showed that these paradigms share a common estimator form and theoretical foundation.

Across the literature, a consistent pattern emerges: adaptivity becomes effective when context, computation, and interpretation are aligned. The principles of context-aware efficiency integrate these aspects, clarifying when adaptation enhances robustness and when it introduces instability. Within this perspective, model design choices can be connected to measurable outcomes such as data efficiency, modularity, and transferability, grounding the abstract notion of adaptivity in verifiable performance.

The unified view presented in this review connects statistical inference with ideas from machine learning and cognitive modeling, where adaptive reasoning and context-sensitive generalization are regarded as key components of intelligent behavior. Cognitive theories have long emphasized that efficient adaptation arises from internal models that balance precision and flexibility, an idea now mirrored in recent computational analyses of in-context learning [146]. By bridging these perspectives, this framework provides both a conceptual foundation and a practical guide for developing adaptive systems that are interpretable, reliable, and scalable.

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.

Recent studies also demonstrate that context-adaptive strategies can emerge spontaneously in large models trained on diverse tasks, linking computational efficiency to rational inference principles [148]. These findings suggest that implicit adaptation can serve as a computational analog of Bayesian updating, where context dynamically reweights prior knowledge to improve generalization. Similar ideas have been explored in meta-learning frameworks such as MetaICL, which meta-trains language models to acquire reusable adaptation strategies through exposure to varied task distributions [142].

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 [241].

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

Looking ahead, the evolution of context-adaptive inference will likely proceed along four interconnected paths.

Theoretical Foundations

Future research should formalize implicit adaptation within a consistent statistical framework, linking neural computation to principles of efficiency, identifiability, and invariance. Clarifying these theoretical connections will support better understanding of when implicit adaptation approximates explicit statistical reasoning and how both approaches can be integrated. Recent advances have begun to view in-context learning as an emergent form of structure induction, suggesting that large models implicitly learn compositional representations that approximate rational inference processes [146].

Modular and Compositional Methods

Progress in parameter-efficient fine-tuning, compositional adaptation, and reusable modules will make large models more flexible and controllable. Building libraries of specialized components that can be dynamically combined will promote efficient reuse and domain transfer while maintaining interpretability. Work on tabular in-context learning, such as the TabICL architecture, illustrates how these principles can scale to structured data domains while preserving modular control and generalization [153].

Evaluation and Reliability

Developing standardized benchmarks that jointly assess robustness, calibration, and interpretability is essential for advancing both theory and application. Future evaluation frameworks should emphasize context-stratified performance, long-term stability, and transparent reporting of adaptation behavior under distribution shifts. Ongoing analyses of the stability and transience of in-context strategies [148] underscore the importance of evaluating not only short-term generalization but also the persistence and reproducibility of adaptive behavior across training regimes.

Responsible and Sustainable Deployment

As adaptive systems become embedded in decision-making processes, integrating fairness auditing, human oversight, and energy efficiency into their design will be critical for ensuring public trust. Addressing the environmental cost of large-scale adaptation and developing resource-conscious algorithms will also contribute to sustainable computing practices. Emerging work on efficient foundation models and rational adaptation frameworks [241] highlights how technical design and ethical responsibility can be jointly optimized in real-world deployment.

Together, these directions outline a path toward the next generation of adaptive models that are both powerful and trustworthy. Progress will depend on combining rigorous statistical understanding with transparent design and responsible deployment, moving steadily toward the broader goal of making implicit adaptation explicit and accountable.

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. **Neural Tangent Kernel: Convergence and Generalization in Neural Networks**
   Arthur Jacot, Franck Gabriel, Clément Hongler
   arXiv (2020-02-11) https://arxiv.org/abs/1806.07572

5. **Neural Tangents: Fast and Easy Infinite Neural Networks in Python**
   Roman Novak, Lechao Xiao, Jiri Hron, Jaehoon Lee, Alexander A Alemi, Jascha Sohl-Dickstein,
   Samuel S Schoenholz
   arXiv (2019-12-06) https://arxiv.org/abs/1912.02803

6. **Statistical methods with varying coefficient models**
   Jianqing Fan, Wenyang Zhang
   Statistics and Its Interface (2008) https://doi.org/gkq3kq
   DOI: 10.4310/sii.2008.v1.n1.a15 · PMID: 18978950 · PMCID: PMC2575822

7. **A Survey of Deep Meta-Learning**
   Mike Huisman, Jan N van Rijn, Aske Plaat
   arXiv (2020) https://doi.org/g96dmh
   DOI: 10.48550/arxiv.2010.03522

8. **LoRA: Low-Rank Adaptation of Large Language Models**
   Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,
   Weizhu Chen
   arXiv (2021) https://doi.org/gthszt
   DOI: 10.48550/arxiv.2106.09685

9. **Foundational Models Defining a New Era in Vision: A Survey and Outlook**
   Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham
   Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Fahad Shahbaz Khan
   arXiv (2023) https://doi.org/g95zd3
   DOI: 10.48550/arxiv.2307.13721

10. **A Comprehensive Survey on Pretrained Foundation Models: A History from BERT to
    ChatGPT**
    Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan,
    Lifang He, … Lichao Sun
    arXiv (2023) https://doi.org/g8vjrk
    DOI: 10.48550/arxiv.2302.09419

11. **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

12. **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

13. **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

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

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

16. **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

17. **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

18. **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

19. **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

20. **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

21. **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](10.1016/j.jbi.2022.104086) · PMID: [35504543](35504543) · PMCID: [PMC9055753](PMC9055753)

22. **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](10.1101/2020.06.25.20140053)

23. **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

24. **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

25. **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

26. **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

27. **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

28. **The Design of Experiments**
Ronald A Fisher
Oliver & Boyd (1949)

29. **Nouvelles méthodes pour la détermination des orbites des comètes**
Adrien-Marie Legendre
Courcier (1805)

30. **Theoria motus corporum coelestium in sectionibus conicis solem ambientium**
Carl Friedrich Gauss
Perthes & Besser (1809)

31. **Generalized Linear Models**
JA Nelder, RWM Wedderburn
Journal of the Royal Statistical Society. Series A (General) (1972) https://doi.org/dhq253
DOI: [10.2307/2344614](10.2307/2344614)

32. **Generalized Linear Models**
P McCullagh, JA Nelder
Springer US (1989) https://doi.org/brhr
DOI: 10.1007/978-1-4899-3242-6

33. **Application of the Logistic Function to Bio-Assay**
Joseph Berkson
Journal of the American Statistical Association (1944-09) https://doi.org/gn82f3
DOI: 10.1080/01621459.1944.10500699

34. **The Regression Analysis of Binary Sequences**
DR Cox
Journal of the Royal Statistical Society Series B: Statistical Methodology (1958-07-01)
https://doi.org/gfzf66
DOI: 10.1111/j.2517-6161.1958.tb00292.x

35. **Statistical Methods for Research Workers**
Ronald A Fisher
Oliver & Boyd (1925)

36. **Herd effects on the growth of beef bulls from different sources tested together under grazing conditions**
CA Morris
New Zealand Journal of Agricultural Research (1981-01) https://doi.org/fxp28n
DOI: 10.1080/00288233.1981.10420865

37. **Construct validity in psychological tests.**
Lee J Cronbach, Paul E Meehl
Psychological Bulletin (1955-07) https://doi.org/dcsjjf
DOI: 10.1037/h0040957 · PMID: 13245896

38. **Estimation of Genetic Parameters**
Charles R Henderson
Annals of Mathematical Statistics (1950)

39. **A Method of Estimating Comparative Rates from Clinical Data. Applications to Cancer of the Lung, Breast, and Cervix**
JNCI: Journal of the National Cancer Institute
(1951-06) https://doi.org/g96wsb
DOI: 10.1093/jnci/11.6.1269

40. **Recovery of inter-block information when block sizes are unequal**
HD PATTERSON, R THOMPSON
Biometrika (1971) https://doi.org/c473hg
DOI: 10.1093/biomet/58.3.545

41. **Random-Effects Models for Longitudinal Data**
Nan M Laird, James H Ware
Biometrics (1982-12) https://doi.org/b8dmsr
DOI: 10.2307/2529876

42. **Estimation in generalized linear models with random effects**
ROBERT SCHALL
Biometrika (1991) https://doi.org/bhxfht
DOI: 10.1093/biomet/78.4.719

43. **Approximate Inference in Generalized Linear Mixed Models**
NE Breslow, DG Clayton
Journal of the American Statistical Association (1993-03) https://doi.org/ggnhwn
DOI: 10.1080/01621459.1993.10594284

44. **Bayes Estimates for the Linear Model**
DV Lindley, AFM Smith
Journal of the Royal Statistical Society Series B: Statistical Methodology (1972-09-01)
https://doi.org/gg2vv3
DOI: 10.1111/j.2517-6161.1972.tb00885.x

45. **Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images**
Stuart Geman, Donald Geman
IEEE Transactions on Pattern Analysis and Machine Intelligence (1984-11) https://doi.org/bpv4j5
DOI: 10.1109/tpami.1984.4767596 · PMID: 22499653

46. **Bayesian Data Analysis**
Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, Donald B Rubin
Chapman and Hall/CRC (2013-11-27) https://doi.org/gqfx8g
DOI: 10.1201/b16018

47. **Functional Data Analysis**
James Ramsay
Encyclopedia of Statistics in Behavioral Science (2005-04-15) https://doi.org/b9tmj6
DOI: 10.1002/0470013192.bsa239

48. **Generalized Additive Models**
Trevor Hastie, Robert Tibshirani
Statistical Science (1986-08-01) https://doi.org/cjx3mk
DOI: 10.1214/ss/1177013604

49. **Representation Learning: A Review and New Perspectives**
Y Bengio, A Courville, P Vincent
IEEE Transactions on Pattern Analysis and Machine Intelligence (2013-08)
https://doi.org/f42hw4
DOI: 10.1109/tpami.2013.50 · PMID: 23787338

50. **Multitask Learning**
Rich Caruana
Machine Learning (1997-07) https://doi.org/d3gsgj
DOI: 10.1023/a:1007379606734

51. **A Survey on Transfer Learning**
Sinno Jialin Pan, Qiang Yang
IEEE Transactions on Knowledge and Data Engineering (2010-10) https://doi.org/bc4vws
DOI: 10.1109/tkde.2009.191

52. **Generalizing from a Few Examples**
Yaqing Wang, Quanming Yao, James T Kwok, Lionel M Ni
ACM Computing Surveys (2020-06-12) https://doi.org/gg37m2
DOI: 10.1145/3386252

53. **Matching Networks for One Shot Learning**
Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, Daan Wierstra
arXiv (2016) https://doi.org/g96wsd
DOI: 10.48550/arxiv.1606.04080

54. **Prototypical Networks for Few-shot Learning**
Jake Snell, Kevin Swersky, Richard S Zemel
arXiv (2017) https://doi.org/g96wsf
DOI: 10.48550/arxiv.1703.05175

55. **A contextual-bandit approach to personalized news article recommendation**
Lihong Li, Wei Chu, John Langford, Robert E Schapire
Proceedings of the 19th international conference on World wide web (2010-04-26)
https://doi.org/bpkcx9
DOI: 10.1145/1772690.1772758

56. **A New Approach to Linear Filtering and Prediction Problems**
RE Kalman
Journal of Basic Engineering (1960-03-01) https://doi.org/dmftj3
DOI: 10.1115/1.3662552

57. **Online Learning: A Comprehensive Survey**
Steven CH Hoi, Doyen Sahoo, Jing Lu, Peilin Zhao
arXiv (2018) https://doi.org/g9643t
DOI: 10.48550/arxiv.1802.02871

58. **Online Learning and Online Convex Optimization**
Shai Shalev-Shwartz
Foundations and Trends® in Machine Learning (2011) https://doi.org/gc7rf4
DOI: 10.1561/2200000018

59. **A survey on concept drift adaptation**
João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, Abdelhamid Bouchachia
ACM Computing Surveys (2014-03) https://doi.org/gd893p
DOI: 10.1145/2523813

60. **Learning with Drift Detection**
João Gama, Pedro Medas, Gladys Castillo, Pedro Rodrigues
Lecture Notes in Computer Science (2004) https://doi.org/ckzcm4
DOI: 10.1007/978-3-540-28645-5_29

61. **Early Drift Detection Method**
Manuel Baena-García, José del Campo-Ávila, Raul Fidalgo, Albert Bifet, Ricard Gavalda, Rafael Morales-Bueno
Fourth International Workshop on Knowledge Discovery from Data Streams (2006)

62. **New ensemble methods for evolving data streams**
Albert Bifet, Geoff Holmes, Bernhard Pfahringer, Richard Kirkby, Ricard Gavaldà
Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (2009-06-28) https://doi.org/dkxcrj
DOI: 10.1145/1557019.1557041

63. **Some aspects of the sequential design of experiments**
Herbert Robbins
Bulletin of the American Mathematical Society (1952) https://doi.org/c325nh
DOI: 10.1090/s0002-9904-1952-09620-8

64. **Reinforcement Learning: An Introduction**
Richard S Sutton, Andrew G Barto
MIT Press (1998)
ISBN: 978-0-262-19398-6

65. **Finite-time Analysis of the Multiarmed Bandit Problem**
Peter Auer, Nicolò Cesa-Bianchi, Paul Fischer
Machine Learning (2002-05) https://doi.org/dxkxgv
DOI: 10.1023/a:1013689704352

66. **A Markovian Decision Process**
Richard Bellman
Journal of Mathematics and Mechanics (1957)

67. **Learning from Delayed Rewards**
Christopher John Cornish Hellaby Watkins
King's College, University of Cambridge (1989)

68. **Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning**
Ronald J Williams
Machine Learning (1992-05) https://doi.org/b762gd
DOI: 10.1023/a:1022672621406

69. **Neuronlike adaptive elements that can solve difficult learning control problems**
Andrew G Barto, Richard S Sutton, Charles W Anderson
IEEE Transactions on Systems, Man, and Cybernetics (1983-09) https://doi.org/gddhk6
DOI: 10.1109/tsmc.1983.6313077

70. **Human-level control through deep reinforcement learning**
Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, … Demis Hassabis
Nature (2015-02-25) https://doi.org/gc3h75
DOI: 10.1038/nature14236 · PMID: 25719670

71. **Compression, restoration, resampling, 'compressive sensing': fast transforms in digital imaging**
LP Yaroslavsky
Journal of Optics (2015-07-01) https://doi.org/g96wr9
DOI: 10.1088/2040-8978/17/7/073001

72. **Speech Analysis and Synthesis by Linear Prediction of the Speech Wave**
BS Atal, Suzanne L Hanauer
The Journal of the Acoustical Society of America (1971-08-01) https://doi.org/cc657m
DOI: 10.1121/1.1912679 · PMID: 4106390

73. **A vector space model for automatic indexing**
G Salton, A Wong, CS Yang
Communications of the ACM (1975-11) https://doi.org/fw8vv8
DOI: 10.1145/361219.361220

74. **Auto-Encoding Variational Bayes**
Diederik P Kingma, Max Welling
arXiv (2022-12-13) https://arxiv.org/abs/1312.6114

75. **Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks**
Chelsea Finn, Pieter Abbeel, Sergey Levine
arXiv (2017) https://doi.org/g5v2js
DOI: 10.48550/arxiv.1703.03400

76. **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

77. **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

78. **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

79. **A Survey on In-context Learning**
Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, … Zhifang Sui
arXiv (2023) https://doi.org/gsv9x4
DOI: 10.48550/arxiv.2301.00234

80. **Adaptive Mixtures of Local Experts**
Robert A Jacobs, Michael I Jordan, Steven J Nowlan, Geoffrey E Hinton
Neural Computation (1991-02) https://doi.org/cnsnqg
DOI: 10.1162/neco.1991.3.1.79 · PMID: 31141872

81. **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

82. **Continuous Temporal Domain Generalization**
Zekun Cai, Guangji Bai, Renhe Jiang, Xuan Song, Liang Zhao
arXiv (2024) https://doi.org/g9582d
DOI: 10.48550/arxiv.2405.16075

83. **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://doi.org/g9582n
DOI: 10.48550/arxiv.2410.17020

84. **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://doi.org/g9582k
DOI: 10.48550/arxiv.2410.10181

85. **Towards Modular LLMs by Building and Reusing a Library of LoRAs**
Oleksiy Ostapenko, Zhan Su, Edoardo Maria Ponti, Laurent Charlin, Nicolas Le Roux, Matheus Pereira, Lucas Caccia, Alessandro Sordoni
arXiv (2024) https://doi.org/g9582c
DOI: 10.48550/arxiv.2405.11157

86. **Mixture of LoRA Experts**

Xun Wu, Shaohan Huang, Furu Wei
arXiv (2024) https://doi.org/g93rnr
DOI: 10.48550/arxiv.2404.13628

87. **Optimal pointwise adaptive methods in nonparametric estimation**
OV Lepski, VG Spokoiny
The Annals of Statistics (1997-12-01) https://doi.org/fwzw52
DOI: 10.1214/aos/1030741083

88. **The Weighted Majority Algorithm**
N Littlestone, MK Warmuth
Information and Computation (1994-02) https://doi.org/c8hw9h
DOI: 10.1006/inco.1994.1009

89. **Selective Test-Time Adaptation for Unsupervised Anomaly Detection using Neural Implicit Representations**
Sameer Ambekar, Julia A Schnabel, Cosmin I Bercea
arXiv (2024) https://doi.org/g9582h
DOI: 10.48550/arxiv.2410.03306

90. **Test-Time Adaptation Induces Stronger Accuracy and Agreement-on-the-Line**
Eungyeup Kim, Mingjie Sun, Christina Baek, Aditi Raghunathan, JZico Kolter
arXiv (2023) https://doi.org/g958z7
DOI: 10.48550/arxiv.2310.04941

91. **Harder Tasks Need More Experts: Dynamic Routing in MoE Models**
Quzhe Huang, Zhenwei An, Nan Zhuang, Mingxu Tao, Chen Zhang, Yang Jin, Kun Xu, Kun Xu, Liwei Chen, Songfang Huang, Yansong Feng
arXiv (2024) https://doi.org/g9582b
DOI: 10.48550/arxiv.2403.07652

92. **Unsupervised Learning via Meta-Learning**
Kyle Hsu, Sergey Levine, Chelsea Finn
arXiv (2018) https://doi.org/g958zs
DOI: 10.48550/arxiv.1810.02334

93. **Bayesian scaling laws for in-context learning**
Aryaman Arora, Dan Jurafsky, Christopher Potts, Noah D Goodman
arXiv (2024) https://doi.org/g9582m
DOI: 10.48550/arxiv.2410.16531

94. **An overview of statistical learning theory**
VN Vapnik
IEEE Transactions on Neural Networks (1999) https://doi.org/fdvhmd
DOI: 10.1109/72.788640 · PMID: 18252602

95. **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://doi.org/g9582f
DOI: 10.48550/arxiv.2406.18219

96. **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

97. **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

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

99. **Model Selection and Estimation in Regression with Grouped Variables**
Ming Yuan, Yi Lin
Journal of the Royal Statistical Society Series B: Statistical Methodology (2005-12-21)
https://doi.org/fvntrn
DOI: 10.1111/j.1467-9868.2005.00532.x

100. **Regression Shrinkage and Selection Via the Lasso**
Robert Tibshirani
Journal of the Royal Statistical Society Series B: Statistical Methodology (1996-01-01)
https://doi.org/gfn45m
DOI: 10.1111/j.2517-6161.1996.tb02080.x

101. **Data Analysis Using Regression and Multilevel/Hierarchical Models**
Andrew Gelman, Jennifer Hill
Cambridge University Press (2006-12-18) https://doi.org/dbrqk6
DOI: 10.1017/cbo9780511790942

102. **Optimization Methods for Large-Scale Machine Learning**
Léon Bottou, Frank E Curtis, Jorge Nocedal
arXiv (2016) https://doi.org/g958zf
DOI: 10.48550/arxiv.1606.04838

103. **Contextual Bandits with Cross-learning**
Santiago Balseiro, Negin Golrezaei, Mohammad Mahdian, Vahab Mirrokni, Jon Schneider
arXiv (2018) https://doi.org/g958zr
DOI: 10.48550/arxiv.1809.09582

104. **Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers**
Stephen Boyd
Foundations and Trends® in Machine Learning (2010) https://doi.org/d3kztk
DOI: 10.1561/2200000016

105. **Adam: A Method for Stochastic Optimization**
Diederik P Kingma, Jimmy Ba
arXiv (2014) https://doi.org/hnkr
DOI: 10.48550/arxiv.1412.6980

106. **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://doi.org/gpmv43
DOI: 10.48550/arxiv.2005.14165

107. **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://doi.org/jpr3
DOI: 10.48550/arxiv.2206.07682

108. **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://doi.org/gtkkf4
DOI: 10.48550/arxiv.2202.12837

109. **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
arXiv (2020) https://doi.org/gtb96w
DOI: 10.48550/arxiv.2001.08361

110. **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
arXiv (2022) https://doi.org/gthszs
DOI: 10.48550/arxiv.2203.15556

111. **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

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

113. **Graphical Models**
Steffen L Lauritzen
Oxford University PressOxford (1996-05-02) https://doi.org/g958zb
DOI: 10.1093/oso/9780198522195.001.0001

114. **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

115. **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

116. **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

117. **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

118. **Fast spatio-temporally varying coefficient modeling with reluctant interaction selection**
Daisuke Murakami, Shinichiro Shirota, Seiji Kajita, Mami Kajita
arXiv (2024) https://doi.org/g9582j
DOI: 10.48550/arxiv.2410.07229

119. **Spatially Varying Coefficient Models for Estimating Heterogeneous Mixture Effects**
Jacob Englert, Howard Chang
arXiv (2025) https://doi.org/g9582q
DOI: 10.48550/arxiv.2502.14651

120. **Network Varying Coefficient Model**
Xinyan Fan, Kuangnan Fang, Wei Lan, Chih-Ling Tsai
Journal of the American Statistical Association (2025-04-11) https://doi.org/g9t2rm
DOI: 10.1080/01621459.2025.2470481

121. **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

122. **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

123. **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

124. **Tree Boosted Varying Coefficient Models**
Yichen Zhou, Giles Hooker
arXiv (2019) https://doi.org/g958zt
DOI: 10.48550/arxiv.1904.01058

125. **A tree-based varying coefficient model**
Henning Zakrisson, Mathias Lindholm
arXiv (2024) https://doi.org/g958z8
DOI: 10.48550/arxiv.2401.05982

126. **Penalized Spline Estimation for Varying-Coefficient Models**
Yiqiang Lu, Riquan Zhang, Liping Zhu
Communications in Statistics - Theory and Methods (2008-05-27) https://doi.org/fpj5gm
DOI: 10.1080/03610920801931887

127. **Deep Multimodal Learning with Missing Modality: A Survey**
Renjie Wu, Hu Wang, Hsiang-Ting Chen, Gustavo Carneiro
arXiv (2024) https://doi.org/g9582g
DOI: 10.48550/arxiv.2409.07825

128. **Domain Adaptation under Missingness Shift**
Helen Zhou, Sivaraman Balakrishnan, Zachary C Lipton
arXiv (2022) https://doi.org/g958z3
DOI: 10.48550/arxiv.2211.02093

129. **Variational Autoencoder with Arbitrary Conditioning**
Oleg Ivanov, Michael Figurnov, Dmitry Vetrov
arXiv (2018) https://doi.org/g958zp
DOI: 10.48550/arxiv.1806.02382

130. **GAIN: Missing Data Imputation using Generative Adversarial Nets**
Jinsung Yoon, James Jordon, Mihaela van der Schaar
arXiv (2018) https://doi.org/g958zq
DOI: 10.48550/arxiv.1806.02920

131. **A class of pattern-mixture models for normal incomplete data**
RODERICK JA LITTLE
Biometrika (1994) https://doi.org/bqw3x9
DOI: 10.2307/2337120

132. **Multiple imputation of incomplete multilevel data using Heckman selection models**
Johanna Muñoz, Matthias Egger, Orestis Efthimiou, Vincent Audigier, Valentijn MT de Jong, ThomasPA Debray
arXiv (2023) https://doi.org/g958z5
DOI: 10.48550/arxiv.2301.05043

133. **XGBoost: A Scalable Tree Boosting System**
Tianqi Chen, Carlos Guestrin
arXiv (2016) https://doi.org/g958zc
DOI: 10.48550/arxiv.1603.02754

134. **The Missing Indicator Method: From Low to High Dimensions**
Mike Van Ness, Tomas M Bosschieter, Roberto Halpin-Gregorio, Madeleine Udell
arXiv (2022) https://doi.org/g958z4
DOI: 10.48550/arxiv.2211.09259

135. **Recurrent Neural Networks for Multivariate Time Series with Missing Values**
Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, Yan Liu
arXiv (2016) https://doi.org/g958zd
DOI: 10.48550/arxiv.1606.01865

136. **BRITS: Bidirectional Recurrent Imputation for Time Series**
Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, Yitan Li
arXiv (2018) https://doi.org/g958zn
DOI: 10.48550/arxiv.1805.10572

137. **XGBoost**
Tianqi Chen, Carlos Guestrin
Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016-08-13) https://doi.org/gdp84q
DOI: 10.1145/2939672.2939785

138. **An Overview of Multi-Task Learning in Deep Neural Networks**
Sebastian Ruder
arXiv (2017) https://doi.org/g958zh
DOI: 10.48550/arxiv.1706.05098

139. **Attention Is All You Need**
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, Illia Polosukhin
arXiv (2017) https://doi.org/gpnmtv
DOI: 10.48550/arxiv.1706.03762

140. **Auto-Encoding Variational Bayes**
Diederik P Kingma, Max Welling
arXiv (2013) https://doi.org/gpp5xv
DOI: 10.48550/arxiv.1312.6114

141. **Meta-Learning in Neural Networks: A Survey**
Timothy Hospedales, Antreas Antoniou, Paul Micaelli, Amos Storkey
arXiv (2020) https://doi.org/g958zx
DOI: 10.48550/arxiv.2004.05439

142. **MetaICL: Learning to Learn In Context**
Sewon Min, Mike Lewis, Luke Zettlemoyer, Hannaneh Hajishirzi
arXiv (2021) https://doi.org/g96dmj
DOI: 10.48550/arxiv.2110.15943

143. **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

144. **Transformers as Support Vector Machines**
Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, Samet Oymak
arXiv (2023) https://doi.org/g958z6
DOI: 10.48550/arxiv.2308.16898

145. **An Explanation of In-context Learning as Implicit Bayesian Inference**
Sang Michael Xie, Aditi Raghunathan, Percy Liang, Tengyu Ma
arXiv (2021) https://doi.org/gtkkfs
DOI: 10.48550/arxiv.2111.02080

146. **In-Context Learning Strategies Emerge Rationally**
Daniel Wurgaft, Ekdeep Singh Lubana, Core Francisco Park, Hidenori Tanaka, Gautam Reddy, Noah D Goodman
arXiv (2025) https://doi.org/g96dmp
DOI: 10.48550/arxiv.2506.17859

147. **In-context Learning and Induction Heads**
Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, … Chris Olah
arXiv (2022) https://doi.org/g95xv8
DOI: 10.48550/arxiv.2209.11895

148. **Learning without training: The implicit dynamics of in-context learning**
Benoit Dherin, Michael Munn, Hanna Mazzawi, Michael Wunder, Javier Gonzalvo
arXiv (2025) https://doi.org/g96dmq
DOI: 10.48550/arxiv.2507.16003

149. **What learning algorithm is in-context learning? Investigations with linear models**
Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, Denny Zhou

arXiv (2022) https://doi.org/grq582
DOI: 10.48550/arxiv.2211.15661

150. **Transformers learn in-context by gradient descent**
Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander
Mordvintsev, Andrey Zhmoginov, Max Vladymyrov
arXiv (2022) https://doi.org/gshbsq
DOI: 10.48550/arxiv.2212.07677

151. **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

152. **Can Transformers Learn Full Bayesian Inference in Context?**
Arik Reuter, Tim GJ Rudner, Vincent Fortuin, David Rügamer
arXiv (2025) https://doi.org/g9582p
DOI: 10.48550/arxiv.2501.16825

153. **TabICL: A Tabular Foundation Model for In-Context Learning on Large Data**
Jingang Qu, David Holzmüller, Gaël Varoquaux, Marine Le Morvan
arXiv (2025) https://doi.org/g96dmn
DOI: 10.48550/arxiv.2502.05564

154. **Chain-of-Thought Prompting Elicits Reasoning in Large Language Models**
Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc
Le, Denny Zhou
arXiv (2022) https://doi.org/gr263w
DOI: 10.48550/arxiv.2201.11903

155. **Explainable AI: A Review of Machine Learning Interpretability Methods**
Pantelis Linardatos, Vasilis Papastefanopoulos, Sotiris Kotsiantis
Entropy (2020-12-25) https://doi.org/gktm9k
DOI: 10.3390/e23010018 · PMID: 33375658 · PMCID: PMC7824368

156. **Language Models as Knowledge Bases?**
Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller,
Sebastian Riedel
arXiv (2019) https://doi.org/g958zw
DOI: 10.48550/arxiv.1909.01066

157. **Towards A Rigorous Science of Interpretable Machine Learning**
Finale Doshi-Velez, Been Kim
arXiv (2017) https://doi.org/h3cz
DOI: 10.48550/arxiv.1702.08608

158. **Rethinking Explainable Machine Learning as Applied Statistics**
Sebastian Bordt, Eric Raidl, Ulrike von Luxburg
arXiv (2024) https://doi.org/g958z9
DOI: 10.48550/arxiv.2402.02870

159. **"Why Should I Trust You?": Explaining the Predictions of Any Classifier**
Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin
arXiv (2016) https://doi.org/hsmk
DOI: 10.48550/arxiv.1602.04938

160. **A Unified Approach to Interpreting Model Predictions**
Scott Lundberg, Su-In Lee
arXiv (2017) https://doi.org/gp6hf4
DOI: 10.48550/arxiv.1705.07874

161. **Axiomatic Attribution for Deep Networks**
Mukund Sundararajan, Ankur Taly, Qiqi Yan
arXiv (2017) https://doi.org/grx4kq
DOI: 10.48550/arxiv.1703.01365

162. **Learning Important Features Through Propagating Activation Differences**
Avanti Shrikumar, Peyton Greenside, Anshul Kundaje
arXiv (2017) https://doi.org/g958zg
DOI: 10.48550/arxiv.1704.02685

163. **This Looks Like That: Deep Learning for Interpretable Image Recognition**
Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, Cynthia Rudin
arXiv (2018) https://doi.org/kg6p
DOI: 10.48550/arxiv.1806.10574

164. **Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning**
Nicolas Papernot, Patrick McDaniel
arXiv (2018) https://doi.org/g958zm
DOI: 10.48550/arxiv.1803.04765

165. **Understanding Black-box Predictions via Influence Functions**
Pang Wei Koh, Percy Liang
arXiv (2017) https://doi.org/mcsx
DOI: 10.48550/arxiv.1703.04730

166. **Inference Suboptimality in Variational Autoencoders**
Chris Cremer, Xuechen Li, David Duvenaud
arXiv (2018) https://doi.org/g958zj
DOI: 10.48550/arxiv.1801.03558

167. **beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework**
Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, Alexander Lerchner
International Conference on Learning Representations (2017) https://openreview.net/forum?id=Sy2fzU9gl

168. **Deep Variational Information Bottleneck**
Alexander A Alemi, Ian Fischer, Joshua V Dillon, Kevin Murphy
arXiv (2016) https://doi.org/gq9mrm
DOI: 10.48550/arxiv.1612.00410

169. **Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)**
Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, Rory Sayres
arXiv (2017) https://doi.org/khk9
DOI: 10.48550/arxiv.1711.11279

170. **Towards Automatic Concept-based Explanations**
Amirata Ghorbani, James Wexler, James Zou, Been Kim
arXiv (2019) https://doi.org/kf7w

DOI: 10.48550/arxiv.1902.03129

171. **Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations**
Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, Olivier Bachem
arXiv (2018) https://doi.org/grx79c
DOI: 10.48550/arxiv.1811.12359

172. **A Framework for the Quantitative Evaluation of Disentangled Representations**
Cian Eastwood, Christopher KI Williams
International Conference on Learning Representations (2018) https://openreview.net/forum?id=By-7dz-AZ

173. **Understanding intermediate layers using linear classifier probes**
Guillaume Alain, Yoshua Bengio
arXiv (2016) https://doi.org/khmg
DOI: 10.48550/arxiv.1610.01644

174. **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://doi.org/g958zv
DOI: 10.48550/arxiv.1905.06316

175. **Locating and Editing Factual Associations in GPT**
Kevin Meng, David Bau, Alex Andonian, Yonatan Belinkov
arXiv (2022) https://doi.org/g958z2
DOI: 10.48550/arxiv.2202.05262

176. **Knowledge Neurons in Pretrained Transformers**
Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, Furu Wei
arXiv (2021) https://doi.org/gs8cqf
DOI: 10.48550/arxiv.2104.08696

177. **In-Context Explainers: Harnessing LLMs for Explaining Black Box Models**
Nicholas Kroeger, Dan Ley, Satyapriya Krishna, Chirag Agarwal, Himabindu Lakkaraju
arXiv (2023) https://doi.org/g95xwd
DOI: 10.48550/arxiv.2310.05797

178. **A Framework for Evaluating Post Hoc Feature-Additive Explainers**
Zachariah Carmichael, Walter J Scheirer
arXiv (2021) https://doi.org/g958zz
DOI: 10.48550/arxiv.2106.08376

179. **Evaluation of post-hoc interpretability methods in time-series classification**
Hugues Turbé, Mina Bjelogrlic, Christian Lovis, Gianmarco Mengaldo
Nature Machine Intelligence (2023-03-13) https://doi.org/grzdnk
DOI: [10.1038/s42256-023-00620-w](10.1038/s42256-023-00620-w)

180. **Concept Bottleneck Models**
Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, Percy Liang
arXiv (2020) https://doi.org/khz6
DOI: 10.48550/arxiv.2007.04612

181. **Manipulating and Measuring Model Interpretability**
Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, Hanna Wallach
arXiv (2018) https://doi.org/g958zk
DOI: 10.48550/arxiv.1802.07810

182. **Principles of risk minimization for learning theory**
V Vapnik
Advances in Neural Information Processing Systems 5 (NIPS'91) (1991)
https://dl.acm.org/doi/10.5555/2986916.2987018
ISBN: 1558602224

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

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

185. **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

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

187. **Causal inference using invariant prediction: identification and confidence intervals**
Jonas Peters, Peter BÃ¼hlmann, Nicolai Meinshausen
arXiv (2024-04-26) https://arxiv.org/abs/1501.01332

188. **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

189. **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

190. **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

191. **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

192. **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

193. **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

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

195. **Multilevel Statistical Models**
Harvey Goldstein
Wiley Series in Probability and Statistics (2010-10-29) https://doi.org/cj8tgk
DOI: 10.1002/9780470973394

196. **Multilevel growth curve models that incorporate a random coefficient model for the level 1 variance function**
Harvey Goldstein, George Leckie, Christopher Charlton, Kate Tilling, William J Browne
Statistical Methods in Medical Research (2017-05-01) https://doi.org/f95xmx
DOI: 10.1177/0962280217706728 · PMID: 28459180

197. **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) https://doi.org/g96dmg
DOI: 10.1002/sim.9582 · PMID: 36181392 · PMCID: PMC9931182

198. **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

199. **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 (2022) https://doi.org/g96dmk
DOI: 10.48550/arxiv.2212.03471

200. **Bayesian Dynamic Factor Models for High-dimensional Matrix-valued Time Series**
Wei Zhang
arXiv (2024) https://doi.org/g96dmm
DOI: 10.48550/arxiv.2409.08354

201. **International Asset Allocation With Regime Shifts**
Andrew Ang, Geert Bekaert
Review of Financial Studies (2002-07) https://doi.org/b535qr
DOI: 10.1093/rfs/15.4.1137

202. **A Model-Based Method for Remaining Useful Life Prediction of Machinery**
Yaguo Lei, Naipeng Li, Szymon Gontarz, Jing Lin, Stanislaw Radkowski, Jacek Dybala
IEEE Transactions on Reliability (2016-09) https://doi.org/f82bw2
DOI: 10.1109/tr.2016.2570568

203. **Predictive maintenance in the Industry 4.0: A systematic literature review**
Tiago Zonta, Cristiano André da Costa, Rodrigo da Rosa Righi, Miromar José de Lima, Eduardo Silveira da Trindade, Guann Pyng Li
Computers & Industrial Engineering (2020-12) https://doi.org/ghtwvv

DOI: [10.1016/j.cie.2020.106889](https://doi.org/10.1016/j.cie.2020.106889)

204. **Predictive Maintenance Approaches in Industry 4.0: A Systematic Literature Review**
Fidma Mohamed Abdelillah, Hamour Nora, Ouchani Samir, Sidi Mohamed Benslimane
2023 IEEE International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE) (2023-12-14) https://doi.org/g96qvf
DOI: [10.1109/wetice57085.2023.10477842](https://doi.org/10.1109/wetice57085.2023.10477842)

205. **A data-driven and context-aware approach for demand forecasting in the beverage industry**
Benedict Jun Ma, Ilya Jackson, Maggie Huang, Sebastian Villegas, Jaime Macias-Aguayo
International Journal of Logistics Research and Applications (2025-10-10) https://doi.org/g96qvd
DOI: [10.1080/13675567.2025.2566806](https://doi.org/10.1080/13675567.2025.2566806)

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

207. **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

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

209. **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

210. **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

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

212. **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

213. **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

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

215. **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

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

217. **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

218. **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

219. **On the Dangers of Stochastic Parrots**
Emily M Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell
Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (2021-03) https://doi.org/gh677h
DOI: 10.1145/3442188.3445922

220. **Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer**
Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, Jeff Dean
arXiv (2017) https://doi.org/g95xv5
DOI: 10.48550/arxiv.1701.06538

221. **Curvature-Torsion Entropy for Twisted Curves under Curve Shortening Flow**
Gabriel Khan
arXiv (2023) https://doi.org/g95xv9
DOI: 10.48550/arxiv.2305.07171

222. **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

223. **AdapterFusion: Non-Destructive Task Composition for Transfer Learning**
Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, Iryna Gurevych
arXiv (2020) https://doi.org/g95xv7
DOI: 10.48550/arxiv.2005.00247

224. **Does Combining Parameter-efficient Modules Improve Few-shot Transfer Accuracy?**
Nader Asadi, Mahdi Beitollahi, Yasser Khalil, Yinchuan Li, Guojun Zhang, Xi Chen
arXiv (2024) https://doi.org/g95xwf
DOI: 10.48550/arxiv.2402.15414

225. **In-Context Learning through the Bayesian Prism**
Madhur Panwar, Kabir Ahuja, Navin Goyal
arXiv (2023) https://doi.org/g95xwb
DOI: 10.48550/arxiv.2306.04891

226. **On Calibration of Modern Neural Networks**
Chuan Guo, Geoff Pleiss, Yu Sun, Kilian Q Weinberger

arXiv (2017) https://doi.org/g95xv6
DOI: 10.48550/arxiv.1706.04599

227. **Holistic Evaluation of Language Models**
Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, … Yuta Koreeda
arXiv (2022) https://doi.org/kh33
DOI: 10.48550/arxiv.2211.09110

228. **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

229. **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

230. **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

231. **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

232. **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

233. **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

234. **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

235. **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

236. **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

237. **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

238. **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

239. **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

240. **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

241. **Agentic Context Engineering: Evolving Contexts for Self-Improving Language Models**
Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, … Kunle Olukotun
arXiv (2025) https://doi.org/g96dmr
DOI: 10.48550/arxiv.2510.04618

Appendix A

This appendix gives full proofs for Proposition 1 and Corollary 1. We keep the weighted support-set notation from the Introduction and make all linear-algebra steps explicit.

A.0 Preliminaries and identities

- **Joint features.** For any pair (x, c), define
$$\psi(x, c) := x \otimes \phi(c) \in \mathbb{R}^{d_x d_c}.$$
For each indexed training example a (standing in for (i, j)), write $\psi_a := \psi(x_a, c_a)$.

- **Design/labels/weights.** Stack $N = \sum_i m_i$ training rows:
$$Z \in \mathbb{R}^{N \times d_x d_c} \text{ with rows } Z_a = \psi_a^T, \qquad y \in \mathbb{R}^N, \qquad W = \operatorname{diag}(w_a) \in \mathbb{R}^{N \times N}, \ w_a \geq 0.$$
Define the (unweighted) Gram matrix $K := ZZ^\top$ and the weighted Gram
$$K_W := W^{1/2} K W^{1/2} = W^{1/2} ZZ^\top W^{1/2}.$$
For a query (x, c), let $k(\cdot, (x, c)) := Z \psi(x, c) \in \mathbb{R}^N$ and $k_{(x,c)} := W^{1/2} k(\cdot, (x, c))$.

- **Vectorization identity.** For conformable matrices A, B, C,
$$\operatorname{vec}(ABC) = (C^\top \otimes A) \operatorname{vec}(B), \quad \langle \operatorname{vec}(B), x \otimes z \rangle = x^\top B z.$$

- **Weighted ridge solution.** For any $X \in \mathbb{R}^{N \times p}$, ridge objective
$$\min_\beta \ \|W^{1/2}(y - X\beta)\|_2^2 + \lambda \|\beta\|_2^2$$
has unique minimizer $\widehat{\beta} = (X^\top W X + \lambda I)^{-1} X^\top W y$ and equivalent dual form
$$\widehat{\beta} = X^\top W^{1/2} (W^{1/2} XX^\top W^{1/2} + \lambda I)^{-1} W^{1/2} y.$$
Predictions for a new feature vector $x_\star$ equal
$$\widehat{f}(x_\star) = x_\star^\top \widehat{\beta} = \underbrace{(W^{1/2} X x_\star)^\top}_{k_\star^\top} (W^{1/2} XX^\top W^{1/2} + \lambda I)^{-1} W^{1/2} y.$$

  This is **kernel ridge regression** (KRR) with kernel $K_W = W^{1/2} XX^\top W^{1/2}$ and query vector $k_\star = W^{1/2} X x_\star$.

A.1 Proof of Proposition 1(A): explicit varying-coefficients ⇔ weighted KRR on joint features

Assume the linear, squared-loss setting with $y = \langle \theta(c), x \rangle + \varepsilon$ and $\mathbb{E}[\varepsilon] = 0$.
Let the varying-coefficients model be $\theta(c) = B \phi(c)$ with $B \in \mathbb{R}^{d_x \times d_c}$ and ridge penalty $\lambda \|B\|_F^2$.

Step 1 (reduce to ridge in joint-feature space).
Vectorize B as $\beta = \operatorname{vec}(B) \in \mathbb{R}^{d_x d_c}$.
By the identity above,
$$x_a^T B \phi(c_a) = \langle \beta, x_a \otimes \phi(c_a) \rangle = \langle \beta, \psi_a \rangle.$$
Thus the weighted objective specialized from (★) is

$$\min_{\beta \in \mathbb{R}^{d_x d_c}} \|W^{1/2}(y - Z\beta)\|_2^2 + \lambda\|\beta\|_2^2,$$

which is exactly weighted ridge with design $X \equiv Z$.

Step 2 (closed form and prediction).
By the ridge solution,
$$\widehat{\beta} = (Z^T W Z + \lambda I)^{-1} Z^T W y,$$
and the prediction at a query (x, c) with joint feature $\psi(x, c)$ is

$$\hat{y}(x, c) = \psi(x, c)^T \widehat{\beta} = \underbrace{\left(W^{1/2} Z \, \psi(x, c)\right)^\top}_{k_{(x,c)}} \left(W^{1/2} Z Z^\top W^{1/2} + \lambda I\right)^{-1} W^{1/2} y.$$

Step 3 (kernel form).
Since $K := Z Z^T$ and $K_W := W^{1/2} K W^{1/2}$,

$$\boxed{\hat{y}(x, c) \;=\; k_{(x,c)}^T \left(K_W + \lambda I\right)^{-1} W^{1/2} y}.$$

Moreover, the (a, b)-th entry of the kernel matrix K is
$$K_{ab} = \langle \psi_a, \psi_b \rangle = \langle x_a \otimes \phi(c_a), \, x_b \otimes \phi(c_b) \rangle = \langle x_a, x_b \rangle \cdot \langle \phi(c_a), \phi(c_b) \rangle,$$
so (A) is precisely **KRR on joint features** with sample weights W.
This proves part (A). ∎

A.2 Proof of Proposition 1(B): linear ICL ⇒ kernel regression

We analyze a single attention layer operating on the weighted support set $S(c)$, using **linear** maps for queries, keys, and values:
$$q(x, c) = Q \, \psi(x, c), \qquad k_a = K \, \psi_a, \qquad v_a = V \, \psi_a,$$
with $Q \in \mathbb{R}^{d_q \times d_\psi}$, $K \in \mathbb{R}^{d_k \times d_\psi}$, $V \in \mathbb{R}^{d_v \times d_\psi}$, $d_\psi = d_x d_c$. Let the **unnormalized** attention score for index a be
$$s_a(x, c) := w_a \langle q(x, c), k_a \rangle \;=\; w_a \psi(x, c)^T Q^T K \, \psi_a.$$
Define normalized weights $\alpha_a(x, c) := s_a(x, c) / \sum_b s_b(x, c)$ (or any fixed positive normalization; the form below is pointwise in $\{\alpha_a\}$). The context representation and scalar prediction are
$$z(x, c) = \sum_a \alpha_a(x, c) \, v_a, \qquad \hat{y}(x, c) = u^T z(x, c).$$

We prove two statements: **(B1)** exact KRR if the attention maps are fixed and only the readout is trained, and **(B2)** kernel regression with the NTK if the attention parameters are trained in the linearized regime.

A.2.1 (B1) Fixed attention, trained linear head ⇒ exact KRR

Assume Q, K, V are fixed functions (pretrained or chosen a priori), hence $\alpha_a(x, c)$ are **deterministic** functions of (x, c) and the support set. Define the induced **feature map**
$$\varphi(x, c) := \sum_a \alpha_a(x, c) \, v_a \; \in \; \mathbb{R}^{d_v}.$$
Stack $\varphi_a := \varphi(x_a, c_a)$ row-wise into $\Phi \in \mathbb{R}^{N \times d_v}$. Training only the readout u with weighted ridge,
$$\widehat{u} \in \arg\min_u \; \|W^{1/2}(y - \Phi u)\|_2^2 + \lambda\|u\|_2^2$$
yields $\widehat{u} = (\Phi^T W \Phi + \lambda I)^{-1} \Phi^T W y$ and predictions

$$\hat{y}(x,c) = \varphi(x,c)^T\hat{u} = \underbrace{\left(W^{1/2}\Phi\,\varphi(x,c)\right)^T}_{k_{(x,c)}}\left(W^{1/2}\Phi\Phi^TW^{1/2} + \lambda I\right)^{-1}W^{1/2}y.$$

Therefore,

$$\boxed{\hat{y}(x,c) = k_{(x,c)}^T\left(K_W + \lambda I\right)^{-1}W^{1/2}y}, \quad K_W := W^{1/2}\underbrace{(\Phi\Phi^T)}_{=:K}W^{1/2},$$

which is exactly **kernel ridge regression** with kernel

$$k\big((x,c),(x',c')\big) = \langle\varphi(x,c),\varphi(x',c')\rangle.$$

Because $v_a = V\psi_a$ and $\alpha_a(x,c) \propto w_a\,\psi(x,c)^TQ^TK\psi_a$, φ is a linear transform of a **weighted average of joint features**; hence the kernel is a dot-product on linear transforms of $\{\psi_a\}$. This proves (B1). ∎

A.2.2 (B2) Training attention in the linearized/NTK regime ⇒ kernel regression with NTK

Now let $\theta = (Q,K,V,u)$ be trainable, and suppose training uses squared loss with gradient flow (or sufficiently small steps) starting from initialization θ_0. The **linearized model** around θ_0 is the first-order Taylor expansion

$$\hat{y}_\theta(x,c) \approx \hat{y}_{\theta_0}(x,c) + \nabla_\theta\hat{y}_{\theta_0}(x,c)^T(\theta - \theta_0) =: \hat{y}_{\theta_0}(x,c) + \phi_{\mathrm{NTK}}(x,c)^T(\theta - \theta_0),$$

where $\phi_{\mathrm{NTK}}(x,c) := \nabla_\theta\hat{y}_{\theta_0}(x,c)$ are the **tangent features**. Standard NTK results (for squared loss, gradient flow, and linearization-validity conditions) imply that the learned function equals **kernel regression with the NTK**:

$$k_{\mathrm{NTK}}\big((x,c),(x',c')\big) := \big\langle\phi_{\mathrm{NTK}}(x,c),\,\phi_{\mathrm{NTK}}(x',c')\big\rangle,$$

i.e., predictions have the KRR form with kernel K_{NTK} on the training set (and explicit ridge if used, or implicit regularization via early stopping).

It remains to identify the structure of ϕ_{NTK} for our **linear attention** block and show it lies in the span of **linear transforms of joint features**. Differentiating $\hat{y}(x,c) = u^T\sum_a \alpha_a(x,c)\,V\psi_a$ at θ_0 yields four groups of terms:

- **Readout path (u).** $\partial\hat{y}/\partial u = \sum_a \alpha_a(x,c)\,V\psi_a = \varphi_0(x,c)$. This is linear in $\{\psi_a\}$.

- **Value path (V).** $\partial\hat{y}/\partial V = \sum_a \alpha_a(x,c)\,u\,\psi_a^T$. This contributes terms of the form $(u \otimes I)\sum_a \alpha_a(x,c)\psi_a$, i.e., linear in $\{\psi_a\}$.

- **Query/key paths (Q,K).** For linear attention with scores $s_a = w_a\,\psi(x,c)^TQ^TK\psi_a$ and normalized $\alpha_a = s_a/\sum_b s_b$, derivatives of α_a w.r.t. Q and K are linear combinations of $\psi(x,c)$ and $\{\psi_a\}$:

$$\frac{\partial\alpha_a}{\partial Q} \propto \sum_b \left[\delta_{ab} - \alpha_b(x,c)\right]w_aw_b\big(K\psi_a\,\psi(x,c)^T\big),$$

$$\frac{\partial\alpha_a}{\partial K} \propto \sum_b \left[\delta_{ab} - \alpha_b(x,c)\right]w_aw_b\big(\psi(x,c)\,\psi_a^TQ^T\big),$$

and hence $\partial\hat{y}/\partial Q$, $\partial\hat{y}/\partial K$ are finite linear combinations of tensors each bilinear in $\psi(x,c)$ and some ψ_a. Contracting with u and V produces terms *linear* in $\psi(x,c)$ and linear in the set $\{\psi_a\}$.

Collecting all components, the tangent feature map can be written as

$$\phi_{\mathrm{NTK}}(x, c) = \mathcal{L}\big(\psi(x, c), \{\psi_a\}\big),$$

where $\mathcal{L}$ is a fixed linear operator determined by θ_0, W, and the normalization rule for attention. Consequently, the NTK takes the **dot-product** form

$$k_{\mathrm{NTK}}\big((x, c), (x', c')\big) = \Psi(x, c)^T \mathcal{M} \, \Psi(x', c'),$$

for some positive semidefinite matrix $\mathcal{M}$ and a finite-dimensional feature stack Ψ that concatenates linear transforms of $\psi(x, c)$ and of the support-set $\{\psi_a\}$. In particular, k_{NTK} is a dot-product kernel on **linear transforms of the joint features** (possibly augmented by normalization-dependent combinations). Therefore, training the linear-attention ICL model in the linearized regime equals kernel regression with such a kernel—completing (B2). ∎

Assumptions for A.2.2. Squared loss; gradient flow (or sufficiently small steps); initialization independent of the data; and a regime where the linearization error stays controlled over training (e.g., small learning rate, sufficient width/depth so that the NTK remains close to its initialization).

A.3 Proof of Corollary 1: retrieval/gating/weighting as kernel/measure choices

In both A.1 and A.2, predictions have the KRR form

$$\hat{y}(x, c) = k_{(x,c)}^T \big(K^{\sharp} + \lambda I\big)^{-1} \mu,$$

where $K^{\sharp}$ is a positive semidefinite kernel matrix computed over the support set (e.g., $K_W = W^{1/2} Z Z^T W^{1/2}$ in A.1 or $W^{1/2} \Phi \Phi^T W^{1/2}$ / K_{NTK} in A.2), $k_{(x,c)}$ is the associated query vector, and $\mu = W^{1/2} y$ (or an equivalent reweighting).

- **Retrieval $R(c)$ / gating.** Changing the support set $S(c)$ (e.g., via a retriever or a gating policy) **removes or adds rows/columns** in $K^{\sharp}$ and entries in $k_{(x,c)}$. This is equivalent to changing the **empirical measure** over which the kernel smoother is computed (i.e., which samples contribute and how).

- **Weights $w_{ij}(c)$.** Changing the weights modifies W and hence replaces K by $K_W = W^{1/2} K W^{1/2}$ and k by $k_{(x,c)} = W^{1/2} k$. This is standard **importance weighting** in kernel regression.

- **Induced kernels.** Attention, value projections, or learned encoders change the **feature map** (e.g., $\psi \mapsto V\psi$ or $\psi \mapsto Q\psi$), thereby changing the kernel $k((x, c), (x', c')) = \langle \Phi(x, c), \Phi(x', c') \rangle$.

Thus retrieval/gating instantiate **neighborhood selection** (measure choice), and value/query/key processing instantiate **kernel choice**. ∎

A.4 Remarks

- **No Gaussianity is required.** Part (A) only uses squared loss and linear algebra; the noise model $y = f(x, c) + \varepsilon$ with $\mathbb{E}[\varepsilon] = 0$ suffices.
- **Early stopping vs. explicit ridge.** If training uses early stopping rather than explicit λ, the resulting predictor is still a kernel regressor with an *implicit* regularization parameter controlled by stopping time (for gradient flow on squared loss).
- **Multiple layers / nonlinear value stacks.** With deeper nonlinear stacks, the exact identities above become local/first-order (linearized) approximations; the NTK statement continues to apply under its usual conditions.