AIMS Mathematics, 11(8): 24650–24666.
DOI:10.3934/math.2026992
Received: 17 May 2026
Revised: 10 July 2026
Accepted: 31 July 2026
Published: 11 August 2026

https://www.aimspress.com/journal/Math

Research article

Information criteria as statistical complexity measures for AI models

Mohieddine Rahmouni*

Applied College, King Faisal University, Al-Ahsa, Saudi Arabia

* **Correspondence:** Email: mrahmouni@kfu.edu.sa.

Abstract: Artificial intelligence (AI) models are often described as complex because they contain many parameters, use nonlinear architectures, or require substantial computational resources. Architectural size, however, is not the same as statistical complexity. A large model can generalize well when regularization or optimization restricts the functions it effectively uses, while a smaller model may overfit if its selected structure is unstable. This paper develops a focused framework for interpreting information criteria as measures of effective statistical complexity in AI model evaluation. It connects Akaike's information criterion, the Bayesian information criterion, Takeuchi's information criterion, the minimum description length principle, and the widely applicable information criterion to distinct sources of model complexity: explicit parameters, shrinkage, structural partitioning, and singular or distributed representations. The framework is positioned relative to distribution-free capacity measures such as VC dimension and Rademacher complexity, and its applicability conditions are stated explicitly. The paper argues against a single universal complexity score and instead recommends reporting three separable quantities: goodness of fit, complexity cost, and residual uncertainty. A numerical illustration on a simulated regression problem shows how the three-part decomposition separates architectural size from generalization behavior across linear, penalized, and tree-based models. This decomposition provides a practical basis for comparing models, assessing parsimony, and avoiding unsupported claims based only on raw parameter counts or in-sample fit.

Keywords: statistical complexity; information criteria; AIC; BIC; MDL; WAIC; effective degrees of freedom; machine learning; model selection; artificial intelligence; regularization; VC dimension; Rademacher complexity; generalization
Mathematics Subject Classification: 62F07; 62J02; 62F15; 62B10

1. Introduction

Artificial intelligence (AI) has intensified a long-standing problem in statistical modeling: How should model complexity be measured? In applied work, complexity is frequently described through

visible architectural features such as the number of parameters, the number of layers, the number of trees, or the depth of a decision rule. These quantities are easy to report, but are often poor measures of the statistical behavior of a model. A model may be large in nominal terms and yet generalize well because regularization, optimization, data augmentation, or implicit constraints restrict the functions it effectively uses. Conversely, a relatively small model may behave as highly complex if it is selected through unstable data-driven search, if it adapts too closely to sample noise, or if its apparent simplicity conceals a large model-selection space.

Throughout the paper, "AI models" refers to the broad class of data-driven predictive models used in modern machine learning practice: regularized linear and generalized linear models, tree ensembles (random forests, gradient boosting), and neural networks, together with the classical statistical models from which they descend. The paper does not treat "AI" as a monolithic object with one governing theory. Instead, it treats AI systems as a family of estimation procedures that share a common evaluation problem, distinguishing genuine signal from flexibility that merely fits sample noise, while generating complexity through different mechanisms depending on model class. This working definition is deliberately narrower than public or policy usage of the term, and it is adopted so that the statistical arguments that follow apply to concretely specified estimators rather than to AI as a general concept.

This distinction matters for research and practice. In economics, finance, medicine, public policy, and legal decision-making, models are not evaluated only by predictive accuracy. They must also be interpretable, auditable, robust, and proportionate to the information contained in the data. A model that improves in-sample fit by exploiting excessive flexibility may perform poorly out of sample, generate unstable explanations, and create governance risks. For these reasons, statistical complexity is not simply a technical concern. It is part of the broader problem of responsible model evaluation.

Econometrics and statistics offer a mature vocabulary for addressing this issue. Information criteria such as Akaike's information criterion (AIC), the Bayesian information criterion (BIC), Takeuchi's information criterion (TIC), the minimum description length (MDL) principle, and the widely applicable information criterion (WAIC) were developed to formalize the trade-off between fit and parsimony [1, 2]. Their usual role is model selection, but their deeper significance is that they attach an explicit cost to flexibility. They ask whether the additional structure captured by a model is worth the complexity required to obtain it.

This paper argues that information criteria provide a disciplined foundation for measuring effective statistical complexity in AI models. The argument is not that AIC, BIC, MDL, TIC, and WAIC are interchangeable or that one universal criterion applies to every AI system. On the contrary, the paper emphasizes that different model classes generate complexity through different mechanisms. Linear models generate complexity primarily through explicit parameters. Penalized regressions generate complexity through shrinkage and sparsity. Tree ensembles generate complexity through data-dependent partitioning of the input space. Neural networks generate complexity through distributed representations, singular parameterizations, and optimization-induced regularization. Because these mechanisms differ, the relevant complexity diagnostic must also differ.

Existing treatments of model complexity generally fall into two separate literatures that rarely engage with each other. Econometric and statistical reviews of AIC, BIC, and related criteria typically remain within regular parametric settings and do not address singular models such as neural networks. Machine-learning treatments of capacity, by contrast, are usually organized around distribution-free

measures such as VC dimension and Rademacher complexity, or around empirical generalization phenomena such as double descent, and rarely connect back to the information-criterion vocabulary that applied econometricians and biostatisticians already use. This paper's contribution is not a new criterion, but an explicit bridge between these two literatures and a diagnosis of when each family of tools is applicable.

Building on this bridge, the paper makes four contributions. First, it clarifies the conceptual difference between architectural complexity and statistical complexity. Architectural complexity refers to the formal size or design of a model. Statistical complexity refers to the amount of effective flexibility the model uses to fit regularities in the data. Second, it organizes major information criteria according to the type of complexity they measure and states the regularity, computability, or Bayesian conditions each one requires, so that the choice of diagnostic is tied to a testable applicability condition rather than to convention: AIC and BIC are most natural for regular parametric models, TIC is useful under misspecification, MDL is appropriate when the cost of describing the model structure is central, and WAIC is relevant for Bayesian and singular models. Third, it relates this information-criterion perspective to distribution-free complexity measures from statistical learning theory, VC dimension and Rademacher complexity, clarifying what each family of tools can and cannot say about a fitted model. Fourth, it proposes a practical decomposition of model evaluation into three quantities: goodness of fit, complexity cost, and residual uncertainty, illustrated numerically on a simulated regression problem. This avoids the problems created by forcing all complexity information into one scalar index.

The remainder of the paper is organized as follows: Section 2 defines statistical complexity, distinguishes it from architectural complexity, and relates it to distribution-free capacity measures. Section 3 reviews the relevant information criteria, explains their complexity interpretations and applicability conditions, and discusses cases in which they agree or disagree. Section 4 develops a taxonomy of complexity sources in AI model classes. Section 5 proposes an operational framework for comparing models and illustrates it numerically. Section 6 concludes with implications for model governance.

2. Statistical complexity and model evaluation

2.1. Architectural and statistical complexity

A useful starting point is to distinguish between two notions of complexity. The first is *architectural complexity*: the complexity visible from the design of a model, such as the number of predictors in a regression, the number of trees in a forest, the depth of a boosted tree, or the number of weights in a neural network. Architectural complexity is directly observable before the model is trained.

The second is *statistical complexity*: the amount of flexibility the fitted model effectively uses when learning from data. Statistical complexity depends not only on the architecture, but also on the sample size, the signal-to-noise ratio, the estimation procedure, the regularization scheme, the optimization path, and the distribution of the data. It is therefore a property of the fitted model and its relationship to the data-generating process, not merely of the model blueprint.

The distinction is central to modern machine learning. A neural network with millions of parameters may have high architectural complexity, but its effective statistical complexity may be constrained by weight decay, dropout, early stopping, stochastic optimization, or the geometry of the loss surface [3,4].

A stepwise regression with a modest number of final predictors may appear simple, but the search procedure used to select those predictors can inflate effective degrees of freedom. A random forest may contain hundreds of trees, but bootstrap aggregation can stabilize predictions and reduce variance relative to a single deep tree [5].

The purpose of complexity measurement is therefore not to punish large models mechanically. It is to ask whether the model's flexibility is warranted by the regularities in the data. This requires measuring complexity in relation to fit, uncertainty, and predictive performance.

2.2. Fit, flexibility, and generalization

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ denote a dataset, and let M denote a fitted model with predictive density $p_M(y \mid x)$. A model is statistically useful when it captures stable regularities in the conditional relationship between Y and X. A model is statistically excessive when it uses flexibility to fit sample-specific noise. The practical difficulty is that both behaviors may improve the in-sample likelihood.

Information criteria address this difficulty by decomposing model evaluation into two parts:

$$\text{Information criterion} = \text{lack of fit} + \text{complexity penalty}. \tag{2.1}$$

For likelihood-based models, the lack-of-fit component is usually $-2\ell(\hat{\theta})$, where $\ell(\hat{\theta})$ is the maximized log-likelihood. The penalty component depends on the criterion and on the assumed model class. The penalty is not arbitrary. It represents a theoretical approximation to the optimism, description length, or Bayesian evidence cost associated with additional flexibility.

This logic is closely related to the bias-variance trade-off. A model that is too simple may have high bias because it cannot represent the relevant structure. A model that is too flexible may have high variance because it adapts too strongly to sample noise. Information criteria formalize this trade-off by asking whether the improvement in fit justifies the increase in flexibility.

2.3. Effective degrees of freedom

In regular parametric models, complexity is often measured by the number of freely estimated parameters k. In many machine learning models, however, k is an inadequate measure. A penalized regression with p candidate predictors may have effective degrees of freedom far below p. A neural network may have a very large parameter count, but many parameter configurations can represent nearly identical functions. A decision tree may have relatively few numerical parameters but a large combinatorial search space.

The concept of effective degrees of freedom provides a more general language. In linear smoothing problems, if the fitted values can be written as

$$\hat{y} = S_\lambda y, \tag{2.2}$$

where S_λ is a smoothing matrix depending on a regularization parameter λ, then the effective degrees of freedom are

$$\mathrm{df}(\lambda) = \mathrm{tr}(S_\lambda). \tag{2.3}$$

For ordinary least squares, this reduces to the number of estimated regression coefficients. For ridge regression, the effective degrees of freedom decline as the penalty increases. For non-linear or non-smooth procedures, analogous quantities may be obtained through covariance penalties, active-set

approximations, cross-validation, bootstrap methods, Bayesian posterior variance terms, or algorithm-specific measures [6, 7].

The key point is that complexity should be measured by the flexibility actually exercised by the fitted procedure, not only by the number of parameters written in the model specification.

2.4. Relation to distribution-free capacity measures

Information criteria are not the only tools developed to formalize model complexity. Statistical learning theory offers distribution-free capacity measures, most notably the Vapnik-Chervonenkis (VC) dimension and Rademacher complexity, which bound the gap between empirical and expected risk uniformly over a hypothesis class [8, 9]. These measures answer a different question from the one addressed by information criteria, and the difference is worth making explicit.

VC dimension and Rademacher complexity are properties of a *hypothesis class*, not of a fitted model. They yield worst-case, distribution-free generalization bounds: For any data-generating process, the gap between training and test risk is bounded in terms of the capacity of the class from which the model is drawn. This generality is their main strength. It is also their main limitation for the purposes of this paper: the resulting bounds are frequently loose in practice, are difficult to compute for deep non-linear architectures, and by construction ignore the specific data-generating process and the specific fitted model, both of which matter for applied model comparison.

Information criteria take the opposite stance. AIC, BIC, TIC, MDL, and WAIC are properties of a *fitted model evaluated against a specific sample*. They condition on the data-generating process (or an approximation to it) and on the realized sample size, and they attach a numerical complexity cost to one particular fitted object rather than to the class it was drawn from. This is not a defect. A complexity measure intended to support a concrete modeling decision (is this fitted model justified by this dataset?) should depend on the sample size and the signal actually present in the data, because the same architecture can be adequately supported by a large, informative sample and be statistically unjustified on a small or noisy one. Distribution-free capacity measures, precisely because they abstract away from the sample, are not designed to answer that question; they are designed to certify worst-case behavior across all possible samples of a given size, which is a different and complementary objective.

The two families are therefore best viewed as serving different purposes rather than competing on the same criterion. VC dimension and Rademacher complexity are appropriate when the goal is a distribution-free theoretical guarantee on generalization for an entire hypothesis class, for instance in deriving learning-theoretic sample-complexity bounds. Information criteria are appropriate when the goal is to compare specific fitted models on a specific dataset and to report an interpretable decomposition of fit versus complexity. Recent work connecting generalization bounds, algorithmic stability, and the double-descent phenomenon in over-parameterized models shows that neither family alone gives a complete account of generalization in modern AI systems [10–12]; the operational framework in Section 5 is deliberately compatible with reporting learning-theoretic capacity measures alongside information criteria rather than in place of them.

3. Information criteria as complexity penalties

3.1. AIC and predictive efficiency

AIC was derived as an approximately unbiased estimator of expected Kullback-Leibler predictive risk [1]. For a regular parametric model with k estimated parameters and maximized log-likelihood $\ell(\hat{\theta})$, AIC is

$$\mathrm{AIC} = -2\ell(\hat{\theta}) + 2k. \tag{3.1}$$

The first term rewards fit; the second term penalizes the optimism created by estimating parameters from the same data used to assess fit. AIC does not aim primarily to identify the true model. Its target is predictive efficiency: among candidate models, it tends to favor the model with low expected information loss when the true data-generating process is approximated by the fitted model.

From a complexity perspective, AIC treats each additional regular parameter as adding approximately two penalty units on the -2 log-likelihood scale. This makes AIC attractive when the objective is predictive performance and when the candidate models are regular enough for the approximation to be meaningful. However, AIC can select models that are too rich if the goal is consistent identification of a finite true model.

3.2. AICc and finite-sample correction

When the sample size is small relative to the number of estimated parameters, AIC may under-penalize complexity. The corrected AIC, usually denoted AICc, adds a finite-sample adjustment [13]:

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1}. \tag{3.2}$$

This correction becomes negligible as n grows, but it can be important when n is modest. In artificial intelligence applications with high-dimensional predictors but limited sample sizes, the rationale underlying AICc is particularly relevant: the complexity cost is not independent of sample size.

3.3. BIC and Bayesian evidence

BIC arises from a large-sample approximation to the marginal likelihood under regularity conditions [2]. It is given by

$$\mathrm{BIC} = -2\ell(\hat{\theta}) + k\log n. \tag{3.3}$$

BIC penalizes complexity more heavily than AIC when $n > e^2$. The penalty reflects the Bayesian evidence cost of spreading prior mass over a larger parameter space. Under suitable conditions, and if the true model is included among the candidates, BIC is consistent for model identification. This property should not be generalized without qualification: it depends on regularity, correct specification, and the structure of the candidate set.

In complexity terms, BIC is most appropriate when the analyst wants a parsimonious finite-dimensional explanation rather than the model with the lowest estimated predictive risk. It is therefore closer to a model-identification criterion than a pure forecasting criterion.

3.4. TIC and misspecification

AIC assumes a regular parametric setting in which the fitted model is used to approximate the data-generating process. In many applied settings, and especially in artificial intelligence, all candidate models are misspecified. Takeuchi's Information Criterion extends the AIC logic to misspecified models. It replaces the simple parameter count with a trace term involving the expected Hessian and the variance of the score:

$$\mathrm{TIC} = -2\ell(\hat{\theta}) + 2\,\mathrm{tr}\left\{J(\hat{\theta})^{-1}K(\hat{\theta})\right\}, \tag{3.4}$$

where J is the negative expected Hessian of the log-likelihood and K is the covariance matrix of the score. Under correct specification, $J = K$, and the penalty reduces to $2k$. Under misspecification, the effective penalty may differ from the nominal parameter count.

TIC is conceptually important for artificial intelligence because misspecification is the norm rather than the exception. In practice, estimating the trace term may be difficult, but the criterion clarifies why raw parameter counts can be misleading: the complexity cost depends on the curvature and variability of the fitted model under the data-generating process.

3.5. MDL and description length

The MDL principle evaluates a model by the length of the shortest code needed to describe the model and the data given the model [14, 15]. In a simplified two-part formulation,

$$\mathrm{MDL} = L(M) + L(\mathcal{D} \mid M), \tag{3.5}$$

where $L(M)$ is the code length of the model and $L(\mathcal{D} \mid M)$ is the code length of the data conditional on the model. The best model is the one that compresses the data most effectively without requiring an unnecessarily long description of the model itself.

MDL is especially useful when complexity is structural rather than purely parametric. Decision trees, rule lists, model-selection procedures, and ensembles can be evaluated by asking how many bits are needed to describe their structure, splitting rules, thresholds, and terminal predictions. Unlike AIC and BIC, which are naturally expressed for regular parametric likelihoods, MDL provides a more general language for models whose complexity lies in a data-dependent architecture.

3.6. WAIC and singular learning

WAIC was developed in singular learning theory and Bayesian statistics [16]. Singular models are models in which the Fisher information matrix may be degenerate, parameters may be non-identifiable, and the usual regular asymptotic approximations fail. Neural networks, mixture models, hidden Markov models, and many hierarchical Bayesian models are examples.

WAIC estimates Bayesian predictive accuracy using the log pointwise predictive density and a penalty term based on the posterior variance of the log-likelihood. In one common notation,

$$\mathrm{WAIC} = -2\left(\mathrm{lppd} - p_{\mathrm{WAIC}}\right), \tag{3.6}$$

where

$$\mathrm{lppd} = \sum_{i=1}^{n} \log E_{\theta|\mathcal{D}}\left[p(y_i \mid x_i, \theta)\right] \tag{3.7}$$

and

$$p_{\mathrm{WAIC}} = \sum_{i=1}^{n} \mathrm{Var}_{\theta|\mathcal{D}} \left[\log p(y_i \mid x_i, \theta) \right]. \tag{3.8}$$

WAIC should not be described as simply "consistent" in the same sense as BIC. Its central role is different: it approximates Bayesian leave-one-out predictive performance and remains meaningful in singular settings where standard parameter-count penalties are not theoretically reliable.

Practical applicability to neural networks. WAIC requires posterior samples of the parameters, $\theta^{(1)}, \ldots, \theta^{(S)} \sim p(\theta \mid \mathcal{D})$, from which lppd and p_{WAIC} are estimated by Monte Carlo averages over $s = 1, \ldots, S$. For a standard deterministically trained neural network, this posterior is not directly available, and WAIC is therefore not a routine diagnostic for a single point-estimate network fitted by ordinary stochastic gradient descent. It becomes practically applicable in three settings that are increasingly common in AI practice: (i) genuinely Bayesian neural networks fitted by variational inference or Markov chain Monte Carlo, where the variational posterior or MCMC draws supply $\theta^{(s)}$ directly; (ii) ensembles of independently-trained networks or Monte Carlo dropout at test time, which are commonly used as a tractable approximation to posterior sampling, so that each ensemble member or dropout mask plays the role of $\theta^{(s)}$; and (iii) stochastic-weight-averaging or checkpoint ensembles along a single optimization trajectory, which approximate local posterior mass around a minimum. Outside these settings, norm-based capacity measures, PAC-Bayesian bounds, and cross-validated predictive error remain the more readily computable diagnostics for a single trained network, and WAIC is best treated as the target that these more tractable diagnostics approximate rather than as a routinely computed number.

3.7. Applicability conditions, agreement, and disagreement

The criteria in Table 1 are not applicable under a common set of assumptions, and treating them as substitutes for one another is a common source of misuse. AIC and BIC require a regular parametric likelihood: a fixed, finite-dimensional parameter vector, a non-degenerate Fisher information matrix, and standard asymptotic conditions on the maximum likelihood estimator. TIC relaxes correct specification but still requires a regular, differentiable likelihood so that the score covariance $K(\hat{\theta})$ and Hessian $J(\hat{\theta})$ are well-defined. WAIC relaxes regularity itself and only requires that the model be a well-defined Bayesian probability model with a computable (or Monte Carlo approximable) posterior; it does not require a parametric likelihood to be regular or even identifiable. MDL is the least restrictive in one specific sense: it requires only that model and data be *describable in a code*, that is, that a coding scheme exists, and does not require a probabilistic likelihood at all. This is why MDL extends naturally to rule lists, tree structures, and other objects for which a likelihood-based criterion is awkward to define, while AIC, BIC, and TIC do not extend to such objects without first embedding them in a probabilistic model.

Because the criteria answer related but distinct questions, they can disagree even when applied to the same candidate models. The following three cases illustrate this point.

(1) *Sample size shifts the AIC/BIC ranking.* Since the BIC penalty $k \log n$ grows with n while the AIC penalty $2k$ does not, a richer model can be preferred by AIC and simultaneously rejected by BIC once $n > e^2$, even for the same fitted likelihoods. This is expected, not anomalous: AIC targets

predictive risk, while BIC targets identification of a parsimonious model, and there is no reason these two targets should always coincide.

(2) *TIC and AIC diverge under misspecification.* When the fitted likelihood is correctly specified, $J = K$ and the TIC penalty collapses to $2k$, matching AIC. Under misspecification, $\mathrm{tr}(J^{-1}K)$ can exceed or fall below k, so TIC can impose a heavier or lighter penalty than AIC on the same model; a persistent gap between the two is itself diagnostic of misspecification.

(3) *MDL and likelihood-based criteria can rank structural models oppositely.* A decision tree with many shallow, low-information splits may have a short two-part MDL code (few bits to describe uninformative splits and a simple residual) while still returning a poor pseudo-likelihood-based AIC if node predictions are treated as if they were independently estimated regular parameters. This mismatch is precisely the motivation for using structural description length, rather than a naive parameter count, for tree-based and ensemble models (Section 4).

In each case the appropriate response is not to average or reconcile the criteria into a single number, but to ask which target, predictive risk, Bayesian evidence, misspecification-robust predictive risk, or description length, matches the modeling question at hand, and to report the corresponding criterion together with its applicability conditions.

Table 1. Information criteria and their complexity interpretations.

Criterion	Penalty term	Complexity interpretation	Typical application
AIC	$2k$	Predictive optimism from estimating regular parameters	Regular parametric models; prediction-oriented selection
AICc	AIC plus finite-sample correction	Stronger penalty when n is small relative to k	Small and moderate samples
BIC	$k \log n$	Bayesian evidence cost of a larger parameter space	Regular finite-dimensional models; parsimonious identification
TIC	$2\,\mathrm{tr}(J^{-1}K)$	Misspecification-adjusted effective parameter cost	Misspecified likelihood models
MDL	Code length	Cost of describing both structure and residual information	Trees, rules, ensembles, and structural models
WAIC	Posterior variance penalty	Bayesian effective complexity in regular or singular models	Bayesian models, hierarchical models, and neural networks

Note: The criteria have different theoretical targets. They should not be interpreted as interchangeable measures of one universal quantity.

4. A taxonomy of complexity sources in artificial intelligence

The mechanisms that generate statistical complexity differ systematically across model classes, and the diagnostic appropriate to each differs accordingly. This section organizes those mechanisms into four groups (parametric, shrinkage-induced, structural, and distributed or singular complexity) and Table 2 summarizes, for each group, the source of flexibility, the diagnostic best suited to measuring it, and the principal caution in applying that diagnostic.

Table 2. Sources of complexity across model classes.

Model class	Complexity source	Useful diagnostic	Main caution
OLS/GLM	Explicit parameters	AIC, BIC, AICc, TIC	Search and selection can add hidden complexity
Ridge	Shrinkage of coefficients	Effective degrees of freedom; AIC with df_λ	Degrees of freedom depend on the penalty and design matrix
Lasso	Sparsity and shrinkage	Active-set size; covariance penalties; cross-validation	Active-set instability can understate selection complexity
Decision trees	Recursive partitions	MDL-style structure length; pruning path	Simple parameter counts are misleading
Forests	Averaged partitions	Out-of-bag error; structural description length; stability	Ensemble size and depth affect complexity differently
Boosted trees	Sequential additive structure	Validation error; learning rate; depth; number of iterations	Overfitting depends on shrinkage and iteration path
Neural nets	Distributed and singular representations	WAIC, cross-validation, PAC-Bayes, compression, norms, stability	Raw parameter count is not enough

Note: The table summarizes diagnostic orientation rather than prescribing one universal measure.

4.1. Parametric complexity

Parametric complexity arises when model flexibility is driven primarily by the number of estimated coefficients. Ordinary least squares is the canonical example. If an OLS regression includes p regressors and an intercept, its nominal degrees of freedom are directly tied to the number of estimated parameters. Under Gaussian errors, the AIC can be written up to an additive constant as

$$\mathrm{AIC}_{\mathrm{OLS}} = n \log(\widehat{\sigma}^2) + 2k, \tag{4.1}$$

where $\widehat{\sigma}^2 = \mathrm{RSS}/n$ under maximum likelihood and k includes the intercept and any additional variance parameter when appropriate.

In this setting, architectural complexity and statistical complexity are relatively close. Adding a predictor generally adds one degree of freedom. This is why classical information criteria work well for comparing nested and non-nested linear models under regular assumptions. However, the equivalence breaks down when variables are selected through repeated search. The final selected model may have k coefficients, but the procedure that found it may have explored a much larger space.

4.2. Shrinkage-induced complexity

Penalized regression methods such as ridge regression, Lasso, and elastic net introduce a deliberate distinction between nominal and effective complexity [17–19]. The objective function typically takes the form

$$\min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda P(\beta) \right\}, \tag{4.2}$$

where $P(\beta)$ is a penalty function and λ controls the strength of regularization. Ridge uses an L_2 penalty, Lasso uses an L_1 penalty, and elastic net combines both.

For ridge regression, the fitted values can be written as $\hat{y} = S_\lambda y$, where

$$S_\lambda = X(X'X + \lambda I)^{-1}X'. \tag{4.3}$$

The effective degrees of freedom are then

$$\mathrm{df}_\lambda = \mathrm{tr}(S_\lambda). \tag{4.4}$$

For Lasso, the effective degrees of freedom can often be approximated by the number of active variables under suitable conditions [7], although this approximation should be interpreted with caution when predictors are highly correlated or when model selection is unstable.

The central point is that penalization reduces effective complexity without necessarily reducing architectural dimension. A penalized model may begin with many candidate predictors but behave statistically like a lower-dimensional model.

4.3. Structural complexity

Tree-based models generate complexity through data-dependent partitions of the feature space. A decision tree recursively splits the predictor space into regions and assigns predictions to terminal nodes. Random forests average many trees built from bootstrap samples and randomized split candidates [5]. Gradient boosting builds an additive ensemble of trees, each attempting to correct the errors of the previous ensemble [20, 21].

The complexity of such models is not adequately described by a simple coefficient count. It depends on the number of terminal nodes, tree depth, split variables, split thresholds, number of trees, learning rate, subsampling, and pruning. A formal MDL analysis would require specifying a code for the tree structure, variables, thresholds, and leaf predictions. A schematic description length for one tree might include terms such as

$$L(T) = L(\text{tree topology}) + L(\text{split variables}) + L(\text{thresholds}) + L(\text{leaf predictions}). \tag{4.5}$$

For an ensemble, this cost must be aggregated across trees and adjusted for the ensemble mechanism. This approach is more defensible than treating a forest as if it were a regular parametric model with a simple parameter count.

To make (4.5) operational, each term can be assigned a concrete code. The topology term L(tree topology) can be encoded by a binary string that marks, for each node visited in a fixed traversal order, whether it is a leaf or an internal split; for a binary tree with m internal nodes this requires on the order of $2m + 1$ bits. The variable-selection term L(split variables) requires $\log_2 p$ bits per internal node to identify which of p candidate predictors is split on. The threshold term L(thresholds) requires a code for the cut point, which can be taken as $\log_2 n_v$ bits if the threshold is chosen among n_v order statistics of the splitting variable observed at that node, reflecting the fact that a split selected from more candidate locations carries a higher description cost. The leaf-prediction term L(leaf predictions) is the cost of encoding the terminal value or class distribution at each leaf, which for regression leaves can be taken as a normal or Laplace code for the residuals within that leaf, and for classification leaves as the Shannon code length implied by the empirical class frequencies. Summing these terms over all nodes gives a single number, in bits, that can be compared across trees of different depth and across trees versus their pruned subtrees, exactly as $k \log n$ is compared across nested regressions under BIC. For an ensemble of B trees, the aggregate structural cost is $\sum_{b=1}^{B} L(T_b)$ plus, where relevant, the cost of describing the ensemble mechanism itself (for example, the learning rate and number of boosting rounds), and this aggregate can be directly compared against the ensemble's fit term to assess whether additional trees are earning their structural cost.

4.4. Distributed and singular complexity

Neural networks generate complexity through distributed representations. Their behavior is not usually interpretable as the sum of independent parameter effects. Many different parameter configurations may represent the same function, and the loss surface can contain flat directions, symmetries, and non-identifiable regions. These properties make neural networks singular statistical models in the sense of singular learning theory [16, 22].

In this setting, classical BIC approximations based on a non-singular Fisher information matrix may fail. Effective complexity is better approached through predictive criteria, Bayesian posterior variance terms, PAC-Bayesian bounds, compression arguments, norm-based capacity measures, stability, or empirical generalization diagnostics [3, 23, 24]. No single measure resolves all issues. The important point is that raw parameter count is not a sufficient statistic for neural-network complexity.

5. An operational framework for complexity comparison

5.1. Avoiding a single forced scalar

A tempting strategy is to define a single scalar measure of effective complexity that combines information criteria, uncertainty reduction, and predictive performance. Such measures can be useful for visualization, but they can also obscure interpretation. AIC and BIC are usually minimized, uncertainty reduction is usually desirable, prediction error is also minimized, and interpretability may increase as complexity decreases. Combining these objects into a single scalar measure requires arbitrary scaling and may reverse the meaning of the index.

A more transparent approach is to report three quantities separately:

(1) **Fit**: how well the model explains or predicts the observed data.
(2) **Complexity cost**: the effective flexibility used by the model.
(3) **Residual uncertainty**: the uncertainty or unexplained variation remaining after fitting.

This decomposition keeps the direction of interpretation clear. A good model should fit well, use no more complexity than necessary, and leave residual uncertainty that is consistent with irreducible noise rather than systematic missed structure.

5.2. Fit

For likelihood-based models, fit can be summarized by $-2\ell(\hat{\theta})$, log predictive density, deviance, or negative log-likelihood on held-out data. For regression models, root mean squared error (RMSE), mean absolute error (MAE), or out-of-sample R^2 may also be used. For classification, log loss, Brier score, calibration error, and accuracy may be relevant. The metric should match the purpose of the model.

In time-series and forecasting settings, random K-fold cross-validation may be inappropriate because it can leak future information into the training set. Rolling-window or expanding-window validation is usually more defensible. In cross-sectional settings, grouped or clustered validation may be needed when observations are not independent.

5.3. Complexity cost

Complexity cost should be measured using the diagnostic appropriate to the model class. In regular parametric models, this can be k, AIC penalty, BIC penalty, or TIC trace penalty. In penalized regression, effective degrees of freedom or active-set size may be used. In tree models, structural quantities such as number of leaves, depth, and description length are more informative. In Bayesian and singular models, WAIC penalty terms, leave-one-out effective number of parameters, posterior concentration, or compression-based measures may be more meaningful.

For comparability, the paper recommends reporting both model-specific complexity diagnostics and standardized ranks within an empirical application. Direct numerical comparison across criteria should be performed only when the criteria are defined on compatible likelihoods and samples.

5.4. Residual uncertainty

Residual uncertainty measures what remains unexplained. For continuous outcomes, residual variance and calibrated predictive uncertainty are natural candidates. For probabilistic classification, predictive entropy and calibration measures are useful. For structured outputs, task-specific uncertainty measures may be required.

Residual uncertainty is not always undesirable. If the outcome contains irreducible randomness, a model that honestly represents uncertainty may be preferable to one that produces overconfident predictions. Thus, residual uncertainty should be interpreted alongside calibration and out-of-sample performance.

5.5. Complexity-performance frontiers

The recommended empirical visualization is a complexity-performance frontier. Let $E(M)$ denote out-of-sample error and $C(M)$ denote an appropriate complexity cost. A model M_a dominates model M_b if

$$E(M_a) \leq E(M_b), \qquad C(M_a) \leq C(M_b), \tag{5.1}$$

with at least one strict inequality. Models that are not dominated form the frontier. This approach avoids the need to impose arbitrary scalar weights on complexity and error.

When models come from different classes, the frontier should be interpreted carefully. The complexity axis may need to use standardized or class-specific measures. For example, one figure may show each model's prediction error against an internally standardized complexity score, while a companion table reports the raw complexity diagnostic used for each class.

5.6. A numerical illustration

The following simulated example is intended purely to illustrate the three-part decomposition of Section 5; it is not a benchmark comparison and no claim of general superiority for any model class should be inferred from it. Data were generated from $y_i = \sin(1.6x_i) + 0.3x_i + \varepsilon_i$, with $x_i \sim \text{Uniform}(-3, 3)$, $\varepsilon_i \sim N(0, 0.5^2)$, and $n = 150$ observations split evenly into training and held-out test sets, so that $n_{\text{train}} = n_{\text{test}} = 75$. The true regression function belongs to no candidate class, so every fitted model is misspecified in the sense of Section 3.4. Six models were compared: an underparameterized linear fit; an overparameterized degree-15 polynomial fitted by ordinary least

squares; ridge and Lasso fits using the same degree-15 polynomial features, standardized before penalization, with $\lambda = 8.0$ for ridge and $\alpha = 0.04$ for Lasso; a shallow (depth-3) regression tree; and an unpruned regression tree grown to full depth. For each model, Table 3 reports the nominal or effective complexity, the in-sample and held-out RMSE, and a Gaussian AIC computed on the training half as

$$\mathrm{AIC} = n_{\mathrm{train}} \log\left(2\pi\widehat{\sigma}^2\right) + n_{\mathrm{train}} + 2k, \tag{5.2}$$

where $\widehat{\sigma}^2 = \mathrm{RSS}/n_{\mathrm{train}}$ and k counts the estimated coefficients and the intercept.

Table 3. Illustrative decomposition of fit, complexity cost, and residual uncertainty

Model	Complexity (k or df_λ)	RMSE (train)	RMSE (test)	AIC
Linear (degree 1)	2.00	0.8272	0.8135	188.4
Polynomial OLS (degree 15)	16.00	0.4258	0.4444	116.8
Ridge (degree 15)	5.06	0.6423	0.6952	156.6
Lasso (degree 15)	4.00	0.5862	0.6229	140.7
Regression tree (depth 3)	8.00	0.4333	0.5519	103.4[†]
Regression tree (unpruned)	75.00	0.0000	0.5523	undefined[†]

Note: the complexity column reports the diagnostic appropriate to each model class, and its entries are not mutually commensurable: the number of estimated coefficients for the linear and polynomial fits, $\mathrm{df}_\lambda = \mathrm{tr}(S_\lambda)$ as in (2.3) for ridge, the active-set size for Lasso, and the number of terminal nodes for the trees. RMSE is reported to four decimal places because the comparison between the two trees turns on the fourth.

[†] Tree AIC values substitute the terminal-node count for a regular parameter count in (5.2) and are reported only to illustrate that this substitution is inappropriate; see the discussion below and Section 4.3. AIC is undefined for the unpruned tree because its training residual sum of squares is exactly zero; a numerical implementation that floors $\widehat{\sigma}^2$ away from zero will return a large negative value here, which should not be read as a criterion value.

Three points from Table 3 support the argument of the paper. First, architectural size does not order generalization performance. The unpruned tree has by far the largest nominal complexity (75 terminal nodes, one per training observation) and fits the training data perfectly, yet its held-out RMSE (0.5523) is no better than that of the depth-3 tree (0.5519), which uses fewer than a ninth as many terminal nodes; a ninefold increase in structural size buys nothing out of sample, and in fact costs a negligible amount. Both trees are also outperformed out of sample by the degree-15 polynomial (0.4444). The two trees are nearly indistinguishable in held-out error, but they are far apart on the other two quantities of Section 5, and in opposite ways. On complexity cost they differ by a factor of nine. On residual uncertainty, the depth-3 tree leaves an in-sample residual spread of 0.4333 against a true noise level of 0.5, while the unpruned tree leaves none at all, so no predictive interval can be derived from its training residuals. Held-out error alone therefore ranks these two models as equivalent; only the complexity and residual uncertainty columns reveal that one of them has absorbed the entire sample into its structure and retains no honest estimate of the noise it was fitted to. Second, shrinkage genuinely reduces effective complexity below the architecture's nominal size: ridge and Lasso operate on the same degree-15 feature set as the unpenalized polynomial fit, yet their effective degrees of freedom (5.06 and 4.00, respectively) are far below the 16 nominal coefficients, exactly the mechanism described in Section 4.2. Third, the table shows where likelihood-based penalties cease to apply. Within the parametric family, the AIC ordering is informative. Once the terminal-node count is substituted for a parameter count, it is

not: the depth-3 tree receives the lowest AIC in the table although its held-out error exceeds that of the polynomial fit, and the criterion is undefined for the interpolating tree. Neither outcome is a paradox; both follow from applying a criterion derived under regularity conditions to a recursively partitioned, data-dependent fit that does not satisfy them, and both motivate the structural description length of Section 4.3. Taken together, the table shows that neither the coefficient count nor the leaf count orders held-out performance, and that a likelihood penalty built on a parameter count is informative within the parametric family but breaks down once the same count is applied to a data-dependent partition.

6. Conclusions

This paper has developed a focused framework for understanding information criteria as measures of effective statistical complexity in artificial intelligence. The central argument is that model complexity should not be reduced to architectural size. Statistical complexity concerns the flexibility a fitted model uses to extract structure from data, and this flexibility must be assessed relative to fit, uncertainty, and generalization.

AIC, BIC, TIC, MDL, and WAIC each provide a different lens on this problem. AIC emphasizes predictive efficiency in regular models. BIC emphasizes Bayesian evidence and parsimonious identification under regular conditions. TIC adjusts the penalty logic for misspecification. MDL treats complexity as description length and is especially useful for structural models. WAIC provides a Bayesian predictive criterion applicable to singular settings. These criteria are not interchangeable, but together they provide a disciplined vocabulary for complexity assessment.

The paper has also argued that model evaluation should keep three quantities separate: goodness of fit, complexity cost, and residual uncertainty. This separation is more transparent than forcing all aspects of complexity into one scalar index. It allows researchers to ask whether a model's additional flexibility is justified by meaningful improvements in predictive or explanatory performance. The numerical illustration in Section 5.6 showed concretely that this three-part decomposition, and not the raw parameter or leaf count, is what tracks generalization behavior across parametric and structural model families. Section 2.4 further situated this decomposition relative to distribution-free capacity measures from statistical learning theory, arguing that information criteria and measures such as VC dimension or Rademacher complexity answer complementary rather than competing questions.

These results have practical implications beyond model selection in the narrow sense. Regulatory and governance frameworks for AI increasingly ask model developers to justify model complexity relative to the problem being solved, to document the basis for believing a model generalizes, and to distinguish genuine predictive structure from artifacts of flexible fitting. A three-part report of fit, complexity cost, and residual uncertainty, together with an explicit statement of which information criterion was used and why its applicability conditions hold, gives auditors and reviewers a concrete, criterion-referenced basis for such judgments, in place of an unexplained parameter count or an aggregate accuracy figure that cannot be decomposed. This is directly relevant to model risk management in finance, to clinical model validation in medicine, and to algorithmic accountability requirements in public-sector AI deployment, where a documented and reproducible complexity assessment is often as important as predictive accuracy itself.

The framework proposed here also points to natural extensions. Formal treatment of model-selection procedures that combine several criteria, extension of the applicability analysis of Section 3.7

to criteria not covered here (for example, focused information criteria or cross-validation-based selection procedures), and systematic empirical comparison of information-criterion-based complexity accounting against learning-theoretic and double-descent-based diagnostics on real AI benchmarks are left for future work.

The principal conclusion is that the econometric and statistical tradition of balancing fit against parsimony remains highly relevant in the age of artificial intelligence. Modern models may be larger and more flexible than classical models, but the underlying question is unchanged: Does the structure learned by the model justify the complexity used to obtain it?

Use of Generative-AI tools declaration

The author declares that the use of AI tools was limited to language editing, checking the logical consistency of selected mathematical arguments, and assisting with the code for the numerical illustration. All scientific content, proofs, results, and conclusions were developed and independently verified by the author, who takes full responsibility for the work.

Acknowledgments

This work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU264210].

Conflict of interest

The author declares no conflicts of interest in this paper.

References

1. H. Akaike, A new look at the statistical model identification, *IEEE Trans. Automat. Contr.*, **19** (1974), 716–723. https://doi.org/10.1109/TAC.1974.1100705
2. G. Schwarz, Estimating the dimension of a model, *Ann. Stat.*, **6** (1978), 461–464. https://doi.org/10.1214/aos/1176344136
3. S. Hochreiter, J. Schmidhuber, Flat minima, *Neural Comput.*, **9** (1997), 1–42. https://doi.org/10.1162/neco.1997.9.1.1
4. P. Bartlett, A. Montanari, A. Rakhlin, Deep learning: a statistical viewpoint, *Acta Numer.*, **30** (2021), 87–201. https://doi.org/10.1017/S0962492921000027
5. L. Breiman, Random forests, *Mach. Learn.*, **45** (2001), 5–32. https://doi.org/10.1023/A:1010933404324
6. B. Efron, The estimation of prediction error: covariance penalties and cross-validation, *J. Am. Stat. Assoc.*, **99** (2004), 619–632. https://doi.org/10.1198/016214504000000692
7. H. Zou, T. Hastie, R. Tibshirani, On the "degrees of freedom" of the Lasso, *Ann. Stat.*, **35** (2007), 2173–2192. https://doi.org/10.1214/009053607000000127
8. V. N. Vapnik, *Statistical learning theory*, New York: Wiley, 1998.

9. P. Bartlett, S. Mendelson, Rademacher and Gaussian complexities: risk bounds and structural results, *J. Mach. Learn. Res.*, **3** (2002), 463–482.
10. M. Belkin, D. Hsu, S. Ma, S. Mandal, Reconciling modern machine-learning practice and the classical bias-variance trade-off, *Proc. Nat. Acad. Sci. U.S.A.*, **116** (2019), 15849–15854. https://doi.org/10.1073/pnas.1903070116
11. O. Bousquet, A. Elisseeff, Stability and generalization, *J. Mach. Learn. Res.*, **2** (2002), 499–526.
12. Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, S. Bengio, Fantastic generalization measures and where to find them, *Proceedings of International Conference on Learning Representations 2020*, 2020, 1–33.
13. C. Hurvich, C. Tsai, Regression and time series model selection in small samples, *Biometrika*, **76** (1989), 297–307. https://doi.org/10.1093/biomet/76.2.297
14. J. Rissanen, Modeling by shortest data description, *Automatica*, **14** (1978), 465–471. https://doi.org/10.1016/0005-1098(78)90005-5
15. P. Grünwald, *The minimum description length principle*, Cambridge: MIT Press, 2007. https://doi.org/10.7551/mitpress/4643.001.0001
16. S. Watanabe, Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory, *J. Mach. Learn. Res.*, **11** (2010), 3571–3594.
17. A. Hoerl, R. Kennard, Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, **12** (1970), 55–67. https://doi.org/10.1080/00401706.1970.10488634
18. R. Tibshirani, Regression shrinkage and selection via the Lasso, *J. R. Stat. Soc. B*, **58** (1996), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
19. H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *J. R. Stat. Soc. B*, **67** (2005), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
20. J. Friedman, Greedy function approximation: a gradient boosting machine, *Ann. Stat.*, **29** (2001), 1189–1232. https://doi.org/10.1214/aos/1013203451
21. T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, 785–794. https://doi.org/10.1145/2939672.2939785
22. S. Watanabe, *Algebraic geometry and statistical learning theory*, Cambridge: Cambridge University Press, 2009. https://doi.org/10.1017/CBO9780511800474
23. D. McAllester, Some PAC-Bayesian theorems, *Mach. Learn.*, **37** (1999), 355–363. https://doi.org/10.1023/A:1007618624809
24. C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning requires rethinking generalization, *Proceedings of International Conference on Learning Representations 2017*, 2017, 1–15.