Submitted to Statistical Science

Statistics and AI: New Challenges, Enduring Principles

T. Tony Cai

Abstract. Artificial intelligence (AI) is fundamentally transforming how data are generated, models are built, and scientific discoveries are made. These transformations make statistical decision theory an indispensable framework for evaluating procedures, assessing their consequences, and identifying fundamental limits under explicit constraints. In practice, this means asking how much a better procedure can achieve with the evidence at hand and when further gains require changing how that evidence is obtained. The answer can guide what procedures to build and what evidence to collect, both in developing AI systems and in using them for scientific discovery.

Statistical methods help shape AI systems, and AI in turn raises new statistical problems. We illustrate this relationship through text watermarking, differential privacy, federated learning, transfer learning, and AI evaluation. Across these examples, what a procedure can achieve depends on how evidence is generated, what can be observed or shared, and which errors matter. We also discuss how AI predictions can guide the choice of measurements and how independent evidence can test the resulting scientific claims.

Key words and phrases: Artificial intelligence, statistical decision theory, text watermarking, differential privacy, federated learning, transfer learning, AI evaluation.

1. INTRODUCTION

Statistics, built on more than a century of foundational work, faces a new period of opportunity, challenge, and responsibility. Artificial intelligence (AI) is changing how data are generated, curated, analyzed, and used in decisions. Its effects reach across science and society, reshaping how researchers pursue discoveries, how people learn and work, and how institutions allocate resources. A study may combine direct measurements with AI-generated records or examine outcomes shaped by actions taken in response to a model's predictions.

The predictive power of modern AI can make classical statistical principles seem less necessary. Those principles matter more than ever as people increasingly act on AI outputs and need to judge whether the evidence supports their decisions. An AI tool can produce plausible conclusions even when the sample is unrepresentative, the model is misspecified, or a key assumption is unjustified. A benchmark ranking can also reverse in deployment if the task mix or error costs differ, even when both systems' error rates stay fixed within each task type (Section 3.5). Statisticians therefore ask how the evidence was obtained, what the analysis is meant to establish, and where the result will be used.

Statistical decision theory uses the consequences of decisions to guide the design of procedures, the choice of evidence to collect, and the comparison of competing methods. The standard of comparison is the best performance attainable with the available evidence under stated assumptions and constraints [8, 11, 67]. That benchmark can show how much a better procedure could achieve and when progress requires different measurements, different access to data, or stronger assumptions that the science can justify.

Statistics has a distinctive role in this transformation because its methods help build the AI systems that are changing scientific practice [14]. In text watermarking, developers can design the signal and the test together to improve detection while controlling false alarms [44]. Differential privacy gives developers a way to train models with a formal guarantee about how much changing one person's record can affect the distribution of the released model [1]. Using such guarantees in federated learning has spurred research on what accuracy is attainable from private site summaries [15].

T. Tony Cai is Daniel H. Silberberg Professor of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania, USA (e-mail: tcai@wharton.upenn.edu).

Transfer learning is central to the development of large language models. BERT and T5 showed how a pretrained model could be adapted to tasks such as question answering and text classification [30, 64]. Reusing a model is especially valuable when labeled examples for a new task are scarce. The statistical question is when reuse improves on learning from target data alone and how much target evidence is needed to verify the gain. When hospitals adapt a shared model, patient records may have to stay at each site, and the updates they exchange may have to satisfy a privacy guarantee. Determining what the hospitals can learn under those conditions is a statistical problem that brings together transfer learning, federated learning, and private inference.

Tukey treated data analysis as an empirical science that includes planning how to collect data and judging what they establish [65]. He saw Wald's decision theory as a way to clarify the choices among procedures, with scientific judgment guiding the selection. Breiman put performance on new data at the center of model comparison [13], and Donoho called for empirical studies of entire analysis workflows [31]. Such studies show how procedures perform in practice, and decision theory helps specify what to measure to judge that performance. In AI-assisted discovery, the study design determines which of a model's suggestions the evidence can support.

In a study of the human nuclear pore, researchers used methods including AlphaFold [47] to predict protein structures and fitted them into maps obtained by cryo-electron microscopy. Their model accounted for more than 90% of the scaffold's molecular mass. The fit showed how well the model described the maps used to assemble it. For one scaffold protein, NUP155, the team obtained a separate cryo-electron microscopy structure and found that it closely matched the local AlphaFold prediction [58].

Section 2 connects the comparison of procedures with the design of studies. Section 3 illustrates the framework through five problems in AI development and use. We then discuss what these changes mean for statistical education and how AI can help researchers choose which measurements to make.

2. DECISION THEORY FOR COMPARING AND DESIGNING PROCEDURES

We can improve a decision by making better use of the available observations or by changing what we observe. Statistical decision theory provides a general framework for both. Holding fixed what can be observed, which procedures are allowed, and how errors are valued gives us an objective basis for comparing procedures and helps clarify what a new procedure should achieve. To compare study designs, we keep the possible states, actions, and costs of error fixed and ask what each design can achieve with the same budget for gathering evidence.

Consider an engineer who uses AI to analyze drone photographs and must decide whether to close a bridge. A better classifier may detect surface cracks that the current system misses. An inspection that measures internal damage changes what the engineer can observe.

In this example, the *state* is the condition of the bridge. The *statistical experiment* is the family of observation distributions indexed by state. A *procedure* maps the observations to a recommendation to close the bridge or keep it open, possibly at random. The *loss* assigns a cost to each action in each state, such as the disruption from closing a safe bridge or the harm from leaving an unsafe bridge open. The *risk* is expected loss in a given state, averaged over the observations and any randomization in the procedure [8, 67]. Figure 1 connects evidence to actions and their consequences.

The *worst-case risk* of a procedure is the supremum of its risk over the states under consideration. The *minimax risk* is the infimum of worst-case risk over admissible procedures. For the bridge, let P_0 and P_1 be the distributions of the photographs when the bridge is safe and when it is unsafe, respectively. Their total variation distance, $\mathrm{TV}(P_0, P_1) = \sup_A |P_0(A) - P_1(A)|$, is the largest difference in probability they assign to a measurable event. Under zero-one loss, an incorrect decision incurs loss one and a correct decision incurs loss zero. For any rule, the probability of closing the bridge when it is safe plus the probability of leaving it open when it is unsafe is at least $1 - \mathrm{TV}(P_0, P_1)$. The worst-case error is at least half that sum, so the minimax risk is at least $(1 - \mathrm{TV}(P_0, P_1))/2$.

If the photographs have similar distributions in the two states, even the best rule based on them faces a substantial worst-case error. With the two states and loss fixed, attaining a worst-case error below the bound requires evidence beyond the photographs, such as an inspection for internal damage. More generally, matching upper and lower bounds establish the best attainable rate as resources grow. Section 3.2 shows how changing access to raw data can change such a limit.

The decision-theoretical framework also guides the construction of procedures. The loss identifies which errors matter, and the observation model suggests where to look for evidence. In some models of edited text, a few unusually large scores carry much of the surviving watermark signal. This motivates a test that looks for an excess of large scores and is calibrated to control false alarms (Section 3.1).

These comparisons are conditional on the states, sampling assumptions, and losses used to define the problem. Choosing them requires scientific knowledge and judgment about which consequences matter and how they

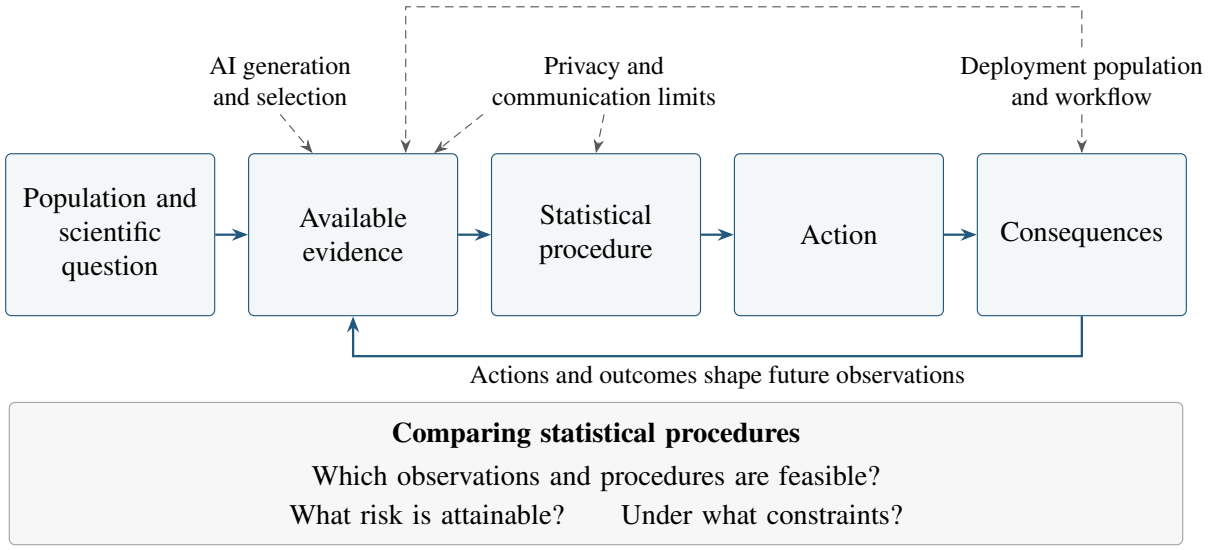


FIG 1. *Comparing procedures begins with the available evidence and the consequences of acting on it. AI can generate or select observations. Privacy and communication rules limit what a procedure can use, and the deployment population and workflow can change both the evidence needed and the consequences of an action. Actions and outcomes can also shape future observations.*

should be weighed. If the population, measurement process, or costs of error change, the model and its guarantees need to be reassessed.

The distinction between improving a procedure and obtaining additional information also helps assess the value of synthetic data. Suppose the analyst keeps the original sample and uses a fixed generator to produce synthetic records. Once the original sample is given, assume that the distribution of the synthetic records does not depend on the state. Given enough computing power and a way to simulate those records, the analyst can reproduce any procedure that uses them. The analyst can also ignore the synthetic records. Thus each procedure using one set of observations has a counterpart using the other with the same risk in every state [10, 11], although adding synthetic records can improve a particular procedure. A generator trained on external data can bring additional information when those data bear on the target population.

Predictions offer a different opportunity when the inputs to a model are available for additional cases and their outcomes are costly to measure. Prediction-powered inference combines predictions on those cases with a smaller sample of measured outcomes [3]. Consider two independent samples from the same population and a fixed predictor trained independently of both. The predictor's inputs are observed in both samples, and outcomes are measured in the smaller sample. Add the mean prediction on the larger sample to the average prediction residual on the smaller sample, where a residual is the measured outcome minus its prediction. This correction yields an unbiased estimate of the population mean. The two averages are independent, so their variances add. If the prediction errors are less variable than the outcomes, sufficiently many predictions can make this estimate more precise than the mean of the measured outcomes alone.

Galaxy Zoo 2 is a citizen science project in which volunteers classify visible features of galaxies in telescope images [69]. To estimate the fraction of galaxies with spiral arms, researchers hid the volunteer classifications for most images to mimic a study with many images and few human labels. Combining model predictions with the retained classifications gave narrower confidence intervals than using those classifications alone. Intervals based on treating predictions as labels missed the reference value obtained from the full set of volunteer classifications [3]. Section 3.5 applies the same idea to AI evaluation and shows how the precision of the correction can depend on which cases receive human review.

3. FIVE STATISTICAL PROBLEMS IN THE AGE OF AI

The framework in Section 2 helps us see how choices about what to measure, generate, or share shape what can be learned. Statisticians can compare these choices by asking what each design can achieve for the same decision problem. We use five examples to show how practical demands in AI raise new statistical questions and how the answers guide the design and evaluation of AI systems. The examples move from designing evidence in watermarking, through limits on access and communication in privacy and federated learning, to using evidence across settings in transfer learning and AI evaluation. Table 1 summarizes the decisions, criteria, and design choices.

3.1 Watermarks and evidence of origin

The widespread use of language models makes the origin of written text a practical question for publishers, plat-

TABLE 1
Decisions and design choices in the five applications.

Problem	Decision and criterion	Design choice
Watermarks	Flag a passage or locate marked positions. Control false flags and missed marks.	Adjust the strength of the mark, calibrate the detector, or seek more text.
Differential privacy	Release an estimate or model with low error under a formal privacy guarantee.	Define the protected unit and who can see raw records before choosing how to privatize the release.
Federated learning	Estimate or train from site messages under privacy and communication limits.	Choose what sites send, which exchanges are allowed, and how to combine the messages.
Transfer learning	Estimate or predict in the target population. Measure gains over using target data alone.	Select sources, check overlap, and collect target data where source and target may differ.
AI evaluation	Choose what to improve, release, or deploy. Compare expected loss and assess rare harms.	Sample from the intended setting, allocate human reviews, and decide which harms need separate measurement.

forms, and readers. An editor who receives a disputed passage needs evidence about how it was produced. A text watermark supplies one kind of evidence by embedding a statistical signal during generation. The verifier tests whether the passage carries a particular mark. A false flag can wrongly cast doubt on the passage's origin, while a missed mark leaves the editor without that evidence. Localization asks which positions carry the mark. The developer must make the mark informative for the intended use while preserving the quality and variety of the text.

Article 50(2) of the European Union (EU) AI Act requires providers to mark covered AI-generated content in a machine-readable form, subject to technical feasibility and specified exceptions [37]. Google DeepMind deployed SynthID-Text in Gemini in 2024 [29], and Anthropic reports watermarking text from supported Claude models [4]. In October 2026, OpenAI announced textGrain for eligible ChatGPT text in the EU and as an opt-in for selected models through its application programming interface [59, 60].

A keyed watermark uses a secret key to influence which tokens, or units of text, a model generates [49]. In textGrain, the key and a window of preceding tokens determine which group each possible next token belongs to and which column of a sampling table the model uses. Each column gives the probabilities of choosing among the groups. Within each group, the model preserves the original relative probabilities. For a fixed next-token distribution, averaging over the randomly selected column recovers the model's original next-token probabilities. An explicit budget limits the average loss of sampling randomness, helping retain variety when the key is reused [54].

A design can choose the embedding and test jointly to improve detection at a given false-positive limit [44]. For a specified watermark, statistical testing also guides the choice of score and detector [52]. textGrain illustrates these choices with an explicit constraint on token sampling [54].

The detector reconstructs scores from the text and key, then flags the passage if their sum exceeds a threshold. To calibrate the test, the idealized calculation assumes that unmarked text is independent of the random key and that distinct context windows use independent randomness. Scoring only the first occurrence of each window then gives a reference distribution for controlling the false-positive rate, the probability of flagging unmarked text. The guarantee averages over keys [54]. With a fixed deployed key, a separate sample of unmarked text can help assess the false-positive rate if it reflects how passages are selected, excerpted, or edited in practice. If scores share randomness, modeling their dependence can improve detection at the same false-positive rate [17].

At a target false-positive rate of 1%, OpenAI reported detecting about 80% of watermarks in 200-token answers to psychology questions and 95% in 400-token answers, with lower rates for mathematics [59]. To see why prevalence matters, consider 1,000 passages, 10 of them marked. A detector with an 80% detection rate and a 1% false-positive rate would yield about eight correct flags and ten false flags on average. At this prevalence and these error rates, the probability that a flagged passage carries the mark is about 45%.

After editing, a few positions may carry most of the surviving watermark evidence. When scores are summed over the whole passage, variation at unmarked positions can mask that evidence. This is a sparse-signal detection problem [16, 26]. A detection boundary describes how many signals must survive, and how strong they must be, for reliable detection in a given model [53]. In one model of textGrain, a test based on the sum of the scores attains the boundary when the number of columns in the sampling table stays fixed as the passage grows. When that number grows and the fraction of sampling randomness the watermark may remove stays bounded away from

zero, the evidence can concentrate in the upper tail of the score distribution. A test sensitive to an excess of large scores can then detect signals the sum misses [9]. The boundary helps distinguish cases where a better test can help from those that require a stronger mark or more text.

Detection and localization serve different purposes in mixed human–AI text. In one model of edited text with a Gumbel-max watermark, passage-level detection can be reliable even when a rule judging each position by its own score cannot both flag at least one position with probability tending to one and drive the marginal false discovery rate to zero as the text gets longer. That rate is the expected number of false flags divided by the expected total number of flags, when the latter is positive [12]. The comparison shows why evidence adequate for flagging an entire passage may give an editor little basis for identifying particular excerpts.

Text watermarking illustrates the two-way relationship between statistics and AI. Statistical testing guides the design of the watermark, and the need to verify edited and mixed human–AI text raises new statistical questions about the limits of detection and localization.

3.2 Differential privacy

Protecting individual privacy is one of the most important challenges in the age of AI. A published statistic or a trained model can reveal information about the people whose records were used, even when the records themselves remain confidential. The European Union's General Data Protection Regulation, for example, requires a lawful basis for processing personal data, including disclosure [36]. Differential privacy limits how much changing one person's data can affect the distribution of a released estimate or model. The statistical question is how much accuracy that protection costs and how the answer depends on who can see the original records.

Suppose a health agency wants to estimate the probability of readmission for patients in a given population. The agency has one binary record for each of n patients, indicating whether the patient was readmitted. The goal is to minimize the worst-case mean squared error over all possible readmission probabilities, subject to a specified privacy guarantee for the released estimate. To isolate the accuracy cost of privacy, compare the smallest attainable error with and without the requirement, keeping the sample size, sampling assumptions, data-access rules, and measure of error fixed.

To define differential privacy, first specify the protected unit and which datasets count as neighbors. In the readmission example, the unit is a patient, and neighboring datasets have the same size and differ by one record replacement. Take any such pair and any measurable set of outputs, where the output includes everything the observer can see. The probability of that set under either dataset is at most e^{ε} times its probability under the other, plus δ, where $\varepsilon \geq 0$ and $0 \leq \delta < 1$. Smaller values give a stronger guarantee. Pure differential privacy sets $\delta = 0$ [35]. When patients have repeated visits, the definition of neighboring datasets must say whether the guarantee covers a single visit or a patient's full history.

Suppose the n records are independent draws with the same readmission probability. Under central privacy, a trusted curator sees all the records and releases the sample mean plus independent, mean-zero Laplace noise of scale $1/(n\varepsilon)$. Replacing one record changes the mean by at most $1/n$, so this release gives pure differential privacy for $\varepsilon > 0$ [34]. The ordinary sample proportion has worst-case mean squared error $1/(4n)$. The added noise increases that error by $2/(n^2\varepsilon^2)$. This extra variance quantifies the cost of privacy for this release. When ε is proportional to $n^{-1/2}$, both terms are of order $1/n$, so the total error tends to zero.

Under local privacy, each patient randomizes their response before sharing it. One rule reports the true response with probability $e^{\varepsilon}/(1+e^{\varepsilon})$ and flips it otherwise. As ε shrinks, the report becomes closer to a coin toss, and correcting for the randomization amplifies sampling variation. For $0 < \varepsilon \leq 1$, the minimax mean squared error for the readmission probability under pure local privacy is of order $\min\{1, 1/(n\varepsilon^2)\}$ with one report per patient, even if each patient's randomization can depend on earlier reports [32]. When ε is proportional to $n^{-1/2}$, as in the central example, the local minimax error stays bounded away from zero. At the same privacy level, who can see the raw records can therefore determine whether estimation error tends to zero.

In private AI training, a trusted trainer can clip each record's contribution to a model update, add noise to the aggregate, and account for privacy across training steps [1]. Studies of means and regression coefficients show how the accuracy cost of privacy can depend on the number of parameters and the structure of the model [19, 21]. Lower bounds identify fundamental limits on how accurately one can learn from the data under a given privacy guarantee [20, 68]. The gap between a procedure's worst-case error and such a bound sets a ceiling on how much a better procedure could gain under the same assumptions and privacy requirement.

The readmission example asks what accuracy a privacy guarantee permits. A privacy audit can test whether a reported guarantee for a trained model is too strong. In a study of private fine-tuning of a language model, the training analysis certified that $\varepsilon = 4$ suffices at a specified δ. At that same δ, an audit using probabilities from the final model gave a 95% lower confidence bound of 0.96 on the smallest valid ε [61, Table 4]. The training analysis gives an upper bound on the smallest valid privacy parameter. The audit gives a lower bound based on what the released model reveals.

The need to learn from sensitive data has made statistical inference under differential privacy an important area of research. Bounds on attainable accuracy can show agencies and developers when the performance they need is out of reach under a proposed privacy guarantee. Audits address a different question by looking for evidence that the reported guarantee is too strong.

3.3 Federated learning across sites

Institutions often hold complementary data and face different limits on what they can share. Federated learning lets them learn from data held across sites while keeping raw records where they are. A hospital with few patients in a group of interest may be able to improve its estimates by working with other hospitals that serve comparable patients. Sites can share summaries or model updates within privacy and communication limits. How much each site can contribute depends on how many records it holds, whom those records represent, and what information it can share. The statistical question is how to organize those exchanges and use the messages to estimate or predict accurately in the population of interest. Choosing the protocol is therefore part of designing the statistical experiment.

Keeping the records separate can add an accuracy cost even at a fixed privacy level. To isolate that cost, compare separate site releases with a single pooled release using the same records and guarantee. If the analyst sees each site's message, the guarantee must cover the joint release.

Return to the readmission example in Section 3.2 and divide the n independent binary records evenly among m hospitals. Exact counts of patients and readmissions suffice to recover the pooled sample proportion. Under pure differential privacy, a trusted service can add Laplace noise to the pooled proportion after combining exact counts, or each hospital can release its own mean with independent Laplace noise. The pooled guarantee assumes that the exact counts are visible only to the trusted service. In the separate-release design, replacing a record at a fixed hospital affects only that hospital's message, so the analyst can see all the noisy messages under the same privacy guarantee. With only n/m patients in each hospital's sample, one record can change its mean by m times as much as the pooled mean. Each hospital therefore adds more noise for the same protection. For this one-release-per-hospital mechanism, averaging the m noisy means preserves the sampling variance and yields m times the privacy-noise variance of the pooled private release.

When privacy requirements differ, sample size alone can give a misleading picture of a site's precision. If sites provide independent, unbiased estimates of the same mean with known variances, inverse-variance weighting gives the smallest variance among unbiased weighted averages. A larger site can then receive less weight if its stricter privacy requirement adds enough noise. If the site means differ, the weights also determine which population mean the estimate represents.

Estimating a curve adds a choice about how much detail to retain. Narrower age bands can show finer changes in readmission risk. With fewer patients in each band, a change to one record can have a larger effect on the estimate, so protecting it may require more noise. One federated regression method shares private summaries of broad trends and finer features of a curve. The weight given to each summary depends on site size, the privacy budget, and the level of detail. The method attains the minimax rates up to logarithmic factors in the settings studied [15]. In an analysis of heart disease data from four hospitals, the researchers used private estimates of the sites' age distributions to weight their contributions to a common curve. The fitted curve declined with age [15]. To estimate the curve at older ages, a smaller hospital with more observations in that age range may be more informative than a larger hospital with few older patients, depending on how much privacy noise each site must add.

Communication can be scarce even when sites are allowed to share model updates. In a study of next-word prediction for Google's Gboard keyboard, phones could participate in federated training only while idle, charging, and connected to an unmetered network [42]. Slow uploads and limited opportunities to connect restrict both how much a phone can send and how often it can exchange updates with the server [56].

Federated averaging addresses the cost of repeated exchanges. A server sends the current model to selected devices, which update it on their local data. The server then averages their updates using sample-size weights. Taking several local steps between exchanges reduced the number of communication rounds needed to reach a target accuracy in the original experiments [56].

A bit budget forces sites to choose between sending more features and describing a few more precisely. Matching bounds for distributed means and functions show when extra bits improve accuracy and when sampling variation limits the gain [24, 25]. With independent site messages, adapting to a curve's unknown smoothness can require more bits to attain the optimal rate than when smoothness is known [24].

The number of bits needed can also depend on how the exchange is organized. In distributed Gaussian mean estimation with unknown variance, early messages can give the analyst starting estimates of the mean and variance. Later sites can then send a single bit indicating whether their observation exceeds the current estimate of the mean. In the model studied, this feedback lets the analyst match the pooled-data rate with fewer bits than independent messages require [23].

Training a shared AI model without pooling records makes communication a statistical design problem. Developers must decide which summaries or updates to request and whether later messages can respond to earlier ones. Comparing protocols at a fixed communication budget can show when a different choice of messages or the use of feedback improves the resulting model.

3.4 Transfer learning across populations

Transfer learning seeks to improve estimation or prediction in a target task or population by using data or a model from a related source. It is especially valuable when outcomes are costly to measure in the target setting. In a study of poverty in five African countries, researchers trained an image model to predict nighttime light levels from daytime satellite images. They then used the resulting image features and household survey data from a much smaller number of locations to predict local spending and wealth [46].

Transfer learning is also central to the development and use of large language models. A single pretrained model can be adapted to many tasks and domains, including those with few labeled examples. To build BioBERT, for example, researchers continued pretraining BERT on biomedical literature and then adapted the model to tasks such as identifying terms and extracting relationships in biomedical text [50]. Developers must choose what data to learn from and how much of the model to change. Continued pretraining on text from the intended domain or task improved performance in a range of applications [41]. Instruction tuning trains a model on examples presented as instructions and can improve performance on tasks left out of that training [27]. Low-rank adaptation (LoRA) fits an update with relatively few trainable parameters while keeping the pretrained weights fixed [43]. The challenge is to limit training costs while allowing enough flexibility to capture the differences that matter for the target task.

Across these applications, what carries over depends on whether populations differ in composition or in how predictors relate to outcomes. The heart-rate analysis in Section 3.3 assumed a shared relationship across hospitals. A hospital adapting a diagnosis model trained elsewhere may have few labeled clinical notes of its own. If the two hospitals describe the same finding differently, words that were informative in the source notes may be less reliable in the target notes. Candidate procedures differ in what they borrow and how they use target observations to correct for discrepancies. Their performance is measured by estimation error or prediction loss in the target population. For classification, bounds on target error depend on source error, differences in the distributions of inputs, and how well a single classifier can work in both settings [7].

If the populations differ in composition and the mean outcome carries over within each group, estimate the target mean by weighting the source group means by the target group proportions. This requires source observations from every group represented in the target. To generalize a randomized trial, the probability of trial participation within each target group must be positive. Among trial participants in each group, the probability of assignment to each treatment being compared must also be positive. The treatments must be comparable across settings, and the mean outcome under each treatment must carry over within groups [28].

A change in how predictors relate to outcomes calls for a different correction. Target outcomes can help estimate that change. In high-dimensional linear regression, Oracle Trans-Lasso fits an initial model using data from the target and sources known to be informative, then estimates a correction from the target data. The gains depend on the amount of source data and on how small or concentrated the coefficient differences are. The method attains minimax rates for estimation and prediction under the conditions studied [51]. When the informative sources are unknown, Trans-Lasso reserves part of the target sample to combine candidate fits, including a fit based on target data alone. The bound on its prediction error reflects how well the best candidate predicts and the cost of combining the fits [51].

In a gene expression study, researchers used data from other tissues to help predict expression of the gene JAM2 in each target brain tissue. Relative to a target-only Lasso, Trans-Lasso reduced fivefold cross-validated prediction error by 22% on average across those tissues. Treating all source tissues as informative performed worse than the target-only Lasso in some tissues [51].

A source can be useful even when it contains fewer observations than the target. In one model for estimating a regression curve, a smaller source study can improve the best attainable accuracy when its curve is smoother and the difference between the curves is well approximated by a low-degree polynomial [18]. For classification, minimax results quantify when source data can improve target accuracy and how privacy may reduce that gain [5, 22].

Evaluate candidate procedures on target cases set aside before fitting or selecting an adaptation. Compare the adapted model with the unadapted source model to estimate the gain from adaptation. To assess the gain from borrowing, also compare it with a model fitted on the same target cases alone, using the same target-label budget for fitting and tuning. Report how many target labels each procedure uses for fitting, tuning, and evaluation. When comparing update designs, holding the number of trainable parameters fixed can help isolate the effect of the update itself.

Synthetic records from an externally trained generator pose a related transfer problem. When the origins of the

records are unknown and the relationship between the two sources is unspecified, the same mixture can be consistent with different distributions of real target records. If the origins are known, evaluate predictions from analyses with and without synthetic records on fresh target cases. Check whether the added records improve accuracy in the target groups where errors matter most. For estimation, check how conclusions change under plausible discrepancies between the two sources. Where such discrepancies are consequential, measuring more real outcomes may be more informative than generating more records.

The success of transfer learning in language models, remote sensing, and biomedicine has made it increasingly important to understand what carries over across tasks and populations. Statistical research helps identify which similarities make a source useful. At a fixed budget for target data, comparisons with a target-only procedure help developers and researchers decide what to reuse and what still needs to be learned from new observations.

3.5 Comparing AI systems for deployment

Evaluating an AI system begins with specifying the tasks it will face and the consequences of its errors. For two fixed systems, the statistical problem is to estimate the difference in expected loss over the deployment task mix and quantify the uncertainty in that estimate. When human review is limited, the design question is which cases to review to make the comparison as precise as possible.

Suppose errors have equal cost. Systems A and B err on 5% and 8% of routine tasks and on 30% and 15% of difficult tasks, respectively. A benchmark with 90% routine tasks favors A (7.5% versus 8.7%). If only 40% of deployment tasks are routine and the error rates within each type carry over, B has the lower rate (12.2% versus 20.0%). The ranking reverses as the share of routine tasks crosses five-sixths (Figure 2). Even when 90% of tasks are routine, system B has the lower expected loss if an error on a difficult task costs twice as much as an error on a routine task.

The reweighting idea from Section 3.4 can also help anticipate performance in deployment. Benchmarks such as Holistic Evaluation of Language Models compare systems across scenarios and criteria [55]. To reweight a benchmark comparison, assume the mean loss difference within each task type carries over to deployment. Estimating each system's expected loss separately requires its own within-task mean loss to carry over. The benchmark must cover every deployment task type and record the outcomes needed to score losses using the costs of errors in deployment. Score the outcomes using that loss, average within each task type, and weight those means by deployment frequencies. Collect new evidence for task types the benchmark misses or where the relevant mean loss or loss

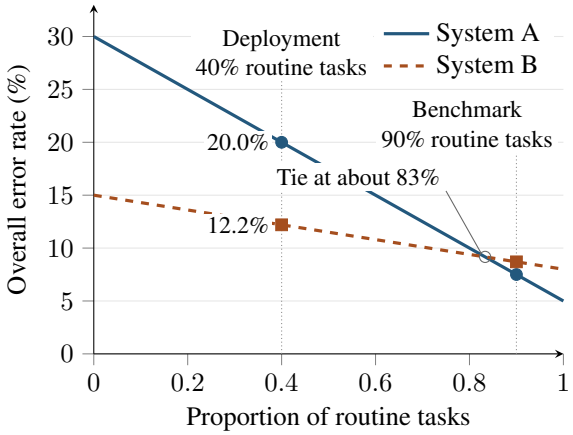


FIG 2. *With equal error costs, a change in task mix can reverse the ranking of two systems even when their error rates within each task type stay fixed. The rates tie when five-sixths of tasks are routine. System B has the lower rate below that point, and system A above it.*

difference may change. Document how responses were sampled and rated [40, 57]. If user ratings may reflect fluency as well as correctness, assess correctness separately [45].

A deployment study can estimate task frequencies and measure losses on those tasks. Testing both systems on the same cases can improve precision when their losses are positively correlated. Confidence intervals should account for both uncertainty in the estimated task frequencies and losses and dependence from paired cases, shared prompts, or shared raters. Use fresh cases for an independent evaluation if the benchmark was used to select or tune the system [33].

When human scoring is expensive, a fixed AI judge can score many responses. In a smaller, independent sample, both the judge and reviewers score the responses, with reviewers using a common rubric. Prediction-powered inference (Section 2) uses the differences on that sample to correct the automated average. In a study of one multilingual summarization system using the Seahorse dataset, researchers estimated how often its summaries were fully supported by their source documents. They used the AI judge's confidence scores to group the responses and allocate samples from existing human ratings. With 300 sampled ratings, stratified prediction-powered inference yielded confidence intervals for the mean score about as narrow on average as those based on 600 human ratings alone [38, Figure 2].

To estimate the mean under the human rubric, sample cases at random for review within each task type and weight the corrected means by deployment frequencies. With the automated sampling design fixed, known deployment frequencies, and equal review costs, allocate reviews roughly in proportion to each task type's frequency times the standard deviation of the difference between human and automated scores in that type. A pilot can esti-

mate those standard deviations [38]. For a paired comparison, calculate the correction on each reviewed case as the human-rated difference between the systems minus the automated difference. Within each task type, add the mean correction to the average automated difference, then weight these corrected differences by deployment frequencies. With equal costs per paired review, the same allocation rule uses the standard deviation of this paired correction.

Rare harms can require a dedicated study. If a fixed version of a document assistant discloses no restricted material in 1,000 independent cases sampled from a specified setting, the one-sided 95% upper confidence bound on the disclosure probability in that setting is about 0.3%. To assess overall deployment risk, evaluators need to know how often each setting arises and how likely disclosure is in each setting.

When people act on AI outputs, evaluate the whole workflow. In a meta-analysis of 106 experimental studies, human–AI teams performed worse, on average across comparisons, than the better of the human-only and AI-only conditions, with more favorable results for content creation than for decision-making [66]. A clinical risk prediction, for example, can prompt treatment that changes the patient outcomes later used to evaluate the model [63]. Randomly assigning comparable users to different workflows can help estimate their effects on decision quality, harms, and workload [62].

AI can change the evidence available for evaluation. One system may score another's outputs, and a deployed system can change the outcomes measured later. Statistical methods help evaluators use limited human reviews to correct automated assessments and design studies of how the whole workflow performs. Decision theory ties that evidence to a choice about whether and how to deploy a system.

4. EDUCATION AND STATISTICAL LEADERSHIP

Generative AI changes how students learn and what their work reveals about their understanding. Students across disciplines need practice using AI in their own fields and judging whether its conclusions follow from the evidence. An AI tutor can offer hints as students work or produce a solution they may struggle to explain. In a randomized study in an undergraduate physics course, students using a purpose-built AI tutor scored higher on a post-test than students taught the same material in an active-learning class [48]. In a randomized high-school mathematics study, students given a general-purpose chat interface did better on practice problems with AI but scored 17% lower on a later exam without AI than students who practiced without it. Students who used the tutor designed to give hints did not score significantly differently from those who practiced without AI [6]. What students can do with AI and what they can do independently are both worth assessing.

Statistical reasoning helps students decide what an AI-assisted analysis can establish. Consider a student who uses AI to summarize a survey. The student needs to ask whose views the sample represents, whether the questions measure what they are intended to measure, and how much uncertainty surrounds the reported differences. An introductory statistics course can develop this judgment by having students trace an AI-assisted conclusion back to the data, explain what would change their minds, and recognize when they need further evidence or expert advice. That practice builds on the emphasis on statistical investigation in the American Statistical Association's college guidelines [2, 39].

Statistics majors should also learn to design studies and evaluate procedures when AI helps determine what evidence is collected. Suppose a model uses the initial responses from that survey to decide who receives follow-up questions. Students should be able to determine what the follow-up responses reveal, what assumptions would justify conclusions about the wider population, and which additional measurements would be most informative. This work draws on probability, inference, and computation, as well as collaboration with scientists who understand how the measurements are made. Statistical decision theory gives students a way to compare follow-up designs by the accuracy they can achieve with the resources available. AI changes what students need to learn about where evidence comes from. In turn, statistical training prepares graduates to help design AI systems and evaluate their effects in practice.

5. DISCUSSION

Statistics helps determine what evidence an AI system can provide and what decisions that evidence can support. The five problems in Section 3 show how the design and use of AI systems, in turn, raise new statistical questions. In watermarking, statistical testing guides the choice of signal to embed in generated text. The need to verify edited text has raised questions about when that signal can be detected and when its location can be established.

Statistical decision theory provides a principled, general framework for comparing procedures and designing new ones. The best performance attainable with the same evidence and constraints gives a benchmark for judging how much improvement is possible. If that benchmark falls short of the accuracy a decision requires, the framework helps compare different ways to gather, generate, or share evidence.

The same reasoning applies to scientific discovery. AI predictions can help researchers choose what to measure next. How a model selects candidates affects what the

measurements can establish, and independent evidence can test the resulting claims. The expected improvement in a scientific decision should justify the cost of gathering that evidence.

Statistics has a distinctive role in the age of AI. Researchers across disciplines use AI to pursue their own questions, from exploring conjectures to planning experiments. Statistical research helps shape the tools they use by establishing what systems can learn under practical constraints and what evidence can support their use in a new setting. Advances on these questions can improve systems used far beyond the application that first motivated the work.

This exchange gives statistics a promising future, with new questions for theory and new opportunities to put its results to work. Working with scientists and developers, statisticians can help build AI systems that expand what researchers can investigate and design studies that test the resulting claims with independent evidence.

STATEMENT OF AI USE

The author used OpenAI tools to assist with revising portions of the text, proofreading, and refining the figures. The author takes full responsibility for the accuracy, integrity, and final form of the manuscript.

REFERENCES

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318. ACM. DOI: 10.1145/2976749.2978318.

[2] American Statistical Association (2026). *Guidelines for Assessment and Instruction in Statistics and Data Science Education (GAISE) College Report*. https://asacollegegaise.org/. Accessed November 12, 2026.

[3] Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. (2023). Prediction-powered inference. *Science*, 382(6671), 669–674. DOI: 10.1126/science.adi6000.

[4] Anthropic (n.d.). How Claude marks AI-generated content. Product documentation. Accessed November 12, 2026.

[5] Auddy, A., Cai, T. T., and Chakraborty, A. (2026). Minimax and adaptive transfer learning for nonparametric classification under distributed differential privacy constraints. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(3), 903–929. DOI: 10.1093/jrsssb/qkaf070.

[6] Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., and Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences*, 122(26), e2422633122. DOI: 10.1073/pnas.2422633122.

[7] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, 79(1–2), 151–175. DOI: 10.1007/s10994-009-5152-4.

[8] Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. 2nd ed. Springer, New York. DOI: 10.1007/978-1-4757-4286-2.

[9] Bhattacharyya, S. and Bonnerjee, S. (2026). textGrain watermarking is harder to detect: can sum-tests still do it optimally? Preprint, arXiv:2610.17655. https://arxiv.org/abs/2610.17655.

[10] Blackwell, D. (1953). Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2), 265–272. DOI: 10.1214/aoms/1177729032.

[11] Blackwell, D. and Girshick, M. A. (1954). *Theory of Games and Statistical Decisions*. John Wiley & Sons, New York.

[12] Blanchet, J. H., Cai, T. T., Li, X., Liu, H., Long, Q., and Su, W. J. (2026). Optimal watermark localization in mixed-source large language model texts. Preprint, arXiv:2608.14906, version 2, revised October 15, 2026. https://arxiv.org/abs/2608.14906v2.

[13] Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3), 199–231. DOI: 10.1214/ss/1009213726.

[14] Cai, T. T. (2025). Statistics at the crossroads: Challenges and opportunities in the age of AI. *IMS Bulletin*, 54(8), 4–6. https://imstat.org/wp-content/uploads/2025/11/Bulletin54_8.pdf.

[15] Cai, T. T., Chakraborty, A., and Vuursteen, L. (2026). Optimal federated learning for nonparametric regression with heterogeneous distributed differential privacy constraints. *Journal of the American Statistical Association*, 1–87. Advance online publication, September 8, 2026. DOI: 10.1080/01621459.2026.2731119.

[16] Cai, T. T., Jeng, X. J., and Jin, J. (2011). Optimal detection of heterogeneous and heteroscedastic mixtures. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(5), 629–662. DOI: 10.1111/j.1467-9868.2011.00778.x.

[17] Cai, T. T., Li, X., Long, Q., Su, W. J., and Wen, G. G. (2025). Optimal detection for language watermarks with pseudorandom collision. Preprint, arXiv:2510.22007. https://arxiv.org/abs/2510.22007.

[18] Cai, T. T. and Pu, H. (2024). Transfer learning for nonparametric regression: Non-asymptotic minimax analysis and adaptive procedure. Preprint, arXiv:2401.12272. https://arxiv.org/abs/2401.12272.

[19] Cai, T. T., Wang, Y., and Zhang, L. (2021). The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5), 2825–2850. DOI: 10.1214/21-AOS2058.

[20] Cai, T. T., Wang, Y., and Zhang, L. (2023). Score attack: A lower bound technique for optimal differentially private learning. Preprint, arXiv:2303.07152, version 3, revised August 15, 2026. https://arxiv.org/abs/2303.07152v3.

[21] Cai, T. T., Wang, Y., and Zhang, L. (2026). The cost of privacy: Rates of convergence for parameter estimation with differential privacy. Corrected version, arXiv:1902.04495v6, September 18, 2026. https://arxiv.org/abs/1902.04495v6.

[22] Cai, T. T. and Wei, H. (2021). Transfer learning for nonparametric classification: Minimax rate and adaptive classifier. *The Annals of Statistics*, 49(1), 100–128. DOI: 10.1214/20-AOS1949.

[23] Cai, T. T. and Wei, H. (2022). Distributed adaptive Gaussian mean estimation with unknown variance: Interactive protocol helps adaptation. *The Annals of Statistics*, 50(4), 1992–2020. DOI: 10.1214/21-AOS2167.

[24] Cai, T. T. and Wei, H. (2022). Distributed nonparametric function estimation: Optimal rate of convergence and cost of adaptation. *The Annals of Statistics*, 50(2), 698–725. DOI: 10.1214/21-AOS2124.

[25] Cai, T. T. and Wei, H. (2024). Distributed Gaussian mean estimation under communication constraints: Optimal rates and communication-efficient algorithms. *Journal of Machine Learning Research*, 25(37), 1–63. https://www.jmlr.org/papers/v25/21-0316.html.

[26] Cai, T. T. and Wu, Y. (2014). Optimal detection of sparse mixtures against a given null distribution. *IEEE Transactions on Information Theory*, 60(4), 2217–2232. DOI: 10.1109/TIT.2014.2304295.

[27] Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., et al. (2024). Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70), 1–53. https://www.jmlr.org/papers/v25/23-0870.html.

[28] Dahabreh, I. J., Robertson, S. E., Tchetgen, E. J., Stuart, E. A., and Hernán, M. A. (2019). Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2), 685–694. DOI: 10.1111/biom.13009.

[29] Dathathri, S., See, A., Ghaisas, S., Huang, P.-S., McAdam, R., Welbl, J., Bachani, V., Kaskasoli, A., Stanforth, R., et al. (2024). Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035), 818–823. DOI: 10.1038/s41586-024-08025-4.

[30] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 1, pages 4171–4186. DOI: 10.18653/v1/N19-1423.

[31] Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4), 745–766. DOI: 10.1080/10618600.2017.1384734.

[32] Duchi, J. C., Jordan, M. I., and Wainwright, M. J. (2018). Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521), 182–201. DOI: 10.1080/01621459.2017.1389735.

[33] Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., and Roth, A. (2015). The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248), 636–638. DOI: 10.1126/science.aaa9375.

[34] Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, Lecture Notes in Computer Science, vol. 3876, pages 265–284. Springer. DOI: 10.1007/11681878_14.

[35] Dwork, C. and Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4), 211–407. DOI: 10.1561/0400000042.

[36] European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). Articles 4(2) and 6.

[37] European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act), as amended. Consolidated text as of July 27, 2026, Article 50(2).

[38] Fisch, A., Maynez, J., Hofer, R. A., Dhingra, B., Globerson, A., and Cohen, W. W. (2024). Stratified prediction-powered inference for effective hybrid evaluation of language models. *Advances in Neural Information Processing Systems*, 37, 111489–111514. DOI: 10.52202/079017-3541.

[39] GAISE College Report ASA Revision Committee (2016). *Guidelines for Assessment and Instruction in Statistics Education College Report 2016*. American Statistical Association. https://www.amstat.org/asa/files/pdfs/GAISE/GaiseCollege_Full.pdf.

[40] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. DOI: 10.1145/3458723.

[41] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. DOI: 10.18653/v1/2020.acl-main.740.

[42] Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., and Ramage, D. (2018). Federated learning for mobile keyboard prediction. Preprint, arXiv:1811.03604, version 2, revised February 28, 2019. https://arxiv.org/abs/1811.03604v2.

[43] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9.

[44] Huang, B., Zhu, H., Zhu, B., Ramchandran, K., Jordan, M. I., Lee, J. D., and Jiao, J. (2024). Towards optimal statistical watermarking. Preprint, arXiv:2312.07930, version 3. https://arxiv.org/abs/2312.07930v3.

[45] Jacobs, A. Z. and Wallach, H. (2021). Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385. ACM. DOI: 10.1145/3442188.3445901.

[46] Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301), 790–794. DOI: 10.1126/science.aaf7894.

[47] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. DOI: 10.1038/s41586-021-03819-2.

[48] Kestin, G., Miller, K., Klales, A., Milbourne, T., and Ponti, G. (2025). AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15, 17458. DOI: 10.1038/s41598-025-97652-6.

[49] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., and Goldstein, T. (2023). A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, PMLR 202, 17061–17084. https://proceedings.mlr.press/v202/kirchenbauer23a.html.

[50] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234–1240. DOI: 10.1093/bioinformatics/btz682.

[51] Li, S., Cai, T. T., and Li, H. (2022). Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1), 149–173. DOI: 10.1111/rssb.12479.

[52] Li, X., Ruan, F., Wang, H., Long, Q., and Su, W. J. (2025). A statistical framework of watermarks for large language models: Pivot, detection efficiency and optimal rules. *The Annals of Statistics*, 53(1), 322–351. DOI: 10.1214/24-AOS2468.

[53] Li, X., Ruan, F., Wang, H., Long, Q., and Su, W. J. (2026). Robust detection of watermarks for large language models under human edits. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(2), 491–515. DOI: 10.1093/jrsssb/qkaf056.

[54] Li, X., Wen, G., Chen, X., Long, Q., Jain, A., Joly, F., Lam, M., Song, Q., and Su, W. (2026). textGrain: Entropy-calibrated watermarking for language model text. OpenAI technical report, October 5, 2026.

[55] Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., et al. (2023). Holistic evaluation of language models. *Transactions on Machine Learning Research*, August. https://openreview.net/forum?id=iO4LZibEqW.

[56] McMahan, B., Moore, E., Ramage, D., Hampson, S., and Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, PMLR 54, 1273–1282. https://proceedings.mlr.press/v54/mcmahan17a.html.

[57] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229. ACM. DOI: 10.1145/3287560.3287596.

[58] Mosalaganti, S., Obarska-Kosinska, A., Siggel, M., Taniguchi, R., Turoňová, B., Zimmerli, C. E., Buczak, K., Schmidt, F. H., Margiotta, E., et al. (2022). AI-based structure prediction empowers integrative structural analysis of human nuclear pores. *Science*, 376(6598), eabm9506. DOI: 10.1126/science.abm9506.

[59] OpenAI (2026). Our approach to EU text provenance rules. October 5, 2026.

[60] OpenAI (n.d.). Provenance signals in content processed by OpenAI tools. Product documentation. Accessed November 19, 2026.

[61] Panda, A., Tang, X., Nasr, M., Choquette-Choo, C. A., and Mittal, P. (2025). Privacy auditing of large language models. In *Proceedings of the 13th International Conference on Learning Representations*. Conference paper.

[62] Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, Cambridge. DOI: 10.1017/CBO9780511803161.

[63] Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. (2020). Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, PMLR 119, 7599–7609. https://proceedings.mlr.press/v119/perdomo20a.html.

[64] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67. https://www.jmlr.org/papers/v21/20-074.html.

[65] Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, 33(1), 1–67. DOI: 10.1214/aoms/1177704711.

[66] Vaccaro, M., Almaatouq, A., and Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12), 2293–2303. DOI: 10.1038/s41562-024-02024-1.

[67] Wald, A. (1950). *Statistical Decision Functions*. John Wiley & Sons, New York.

[68] Wasserman, L. and Zhou, S. (2010). A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489), 375–389. DOI: 10.1198/jasa.2009.tm08651.

[69] Willett, K. W., Lintott, C. J., Bamford, S. P., Masters, K. L., Simmons, B. D., Casteels, K. R. V., Edmondson, E. M., Fortson, L. F., Kaviraj, S., et al. (2013). Galaxy Zoo 2: Detailed morphological classifications for 304 122 galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 435(4), 2835–2860. DOI: 10.1093/mnras/stt1458.