Do Large Language Models (Really) Need Statistical Foundations?

Weijie Su

University of Pennsylvania

May 24, 2025

Abstract

Large language models (LLMs) represent a new paradigm for processing unstructured data, with applications across an unprecedented range of domains. In this paper, we address, through two arguments, whether the development and application of LLMs would genuinely benefit from foundational contributions from the statistics discipline. First, we argue affirmatively, beginning with the observation that LLMs are inherently statistical models due to their profound data dependency and stochastic generation processes, where statistical insights are naturally essential for handling variability and uncertainty. Second, we argue that the persistent black-box nature of LLMs—stemming from their immense scale, architectural complexity, and development practices often prioritizing empirical performance over theoretical interpretability—renders closed-form or purely mechanistic analyses generally intractable, thereby necessitating statistical approaches due to their flexibility and often demonstrated effectiveness. To substantiate these arguments, the paper outlines several research areas—including alignment, watermarking, uncertainty quantification, evaluation, and data mixture optimization—where statistical methodologies are critically needed and are already beginning to make valuable contributions. We conclude with a discussion suggesting that statistical research concerning LLMs will likely form a diverse "mosaic" of specialized topics rather than deriving from a single unifying theory, and highlighting the importance of timely engagement by our statistics community in LLM research.

1 Introduction

Consider a thought experiment where an octopus under the seabed connects to a submarine cable and then eavesdrops on human conversations without any prior knowledge of human language. All it has access to are the utterances of one speaker and the corresponding responses of the other. One may ask: can the octopus ever learn to understand human language, the content of these conversations, and even possess some level of intelligence, purely based on human conversations by passively listening to arbitrarily large amounts of these paired observations?

Email: suw@wharton.upenn.edu.

This thought experiment illustrates how large language models (LLMs) are essentially developed (Kottke, 2023). LLMs are massive neural networks, predominantly based on the Transformer architecture (Vaswani et al., 2017), trained on immense text corpora—including human-written content, code, and various other forms of text—to predict the next word (formally called a "token") given the preceding sequence of tokens (Radford et al., 2018; Brown et al., 2020). By design, an LLM is an autoregressive model that attempts to learn language purely by learning statistical patterns in human-generated text, rather than explicit linguistic rules or semantic grounding.

Interestingly, when Ilya Sutskever and his colleagues at OpenAI proposed next-token prediction as a training strategy in 2018 (Radford et al., 2018), very few researchers believed that such a simple training paradigm would yield capabilities resembling "understanding" of language (Bender et al., 2021; Matthew and Toner, 2024).[1] In stark contrast, when browser-based ChatGPT launched in late 2022, it created a global sensation as the public marveled at its ability to generate human-like text and handle wide-ranging tasks (Bubeck et al., 2023).[2] Not only did ChatGPT impress users with its near Turing test level conversational abilities (Jones and Bergen, 2025), but it also demonstrated elementary reasoning skills—enough to carry out basic statistical analyses and data visualization (Tu et al., 2024; Lin and Zhu, 2025). These capabilities have continued to advance at a rapid pace.

Given that LLMs involve many statistical aspects, a burgeoning body of statistical research on LLMs has emerged (Ji et al., 2025). Nonetheless, the question we pose here is whether LLMs would genuinely benefit from statistical foundations that our community develops. Furthermore, we ask if such statistical contributions would lead to improved development, deployment, and application, particularly to guide and enhance their real-world use? In other words, we move beyond the discussion of how LLMs can be used to enhance statistical analysis and statistical education to focus on whether statistical methodology and insights can improve LLMs themselves.[3]

A straightforward argument supporting the need for statistics in LLMs comes from Richard Sutton's influential essay, *The Bitter Lesson* (Sutton, 2019). The recent Turing Award laureate observed that enduring progress in artificial intelligence (AI) over decades has primarily stemmed from leveraging increased computational power and data scale, rather than incorporating hard-coded human knowledge or intricate model design. Statistical approaches are well suited to leveraging massive data via computation, as they are designed to let the data speak for themselves. Indeed, Sutton articulated in his essay that "statistics and computation came to dominate the field" of AI.

Given the repeated validation of the bitter lesson (Yousefi and Collins, 2024), there are compelling reasons to believe that statistics will undoubtedly benefit LLMs. However, we should recognize that the term "statistics" in Sutton's context is broad. It encompasses both purely predictive algorithms—such as neural networks, boosting, and random forests—and what might be termed in-

[1] In this paper, we exclude non-GPT models such as BERT from our discussion, which use a masked-language-modeling objective (Devlin et al., 2019).

[2] Although the NLP community had already been impressed by GPT-3 in 2020, ChatGPT's web interface made these advances far more widely accessible.

[3] For clarity, this paper does not address how LLMs can be used to enhance statistical analysis or statistical education. Interested readers are referred to Tu et al. (2024).

ferential statistics. This latter category roughly corresponds to what Bradley Efron termed "estimation and attribution" methods (Efron, 2020) or what Leo Breiman called "data-modeling" methods in his *Two Cultures* paper (Breiman, 2001). These approaches focus on interpretable parameters and uncertainty quantification to make inferences from noisy observations. In this paper, we ask whether LLMs, as they become more powerful and widely deployed, would benefit from the interpretable, inferential, and uncertainty-aware techniques that statistics in this narrower sense can offer.

We argue affirmatively that LLMs require more statistical contributions for their continued advancement. Although LLMs by design are purely predictive algorithms, they differ significantly from prior methods in this category, including even pre-Transformer neural networks (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014). This distinction emerges in two key ways: First, LLMs operate on an unprecedented diversity of data types, including almost all possible forms of text data. While earlier models could process text (Manning and Schutze, 1999), it is the first time a single model can integrate and process text almost as seamlessly as numbers, enabling a unified approach to handling tasks as diverse as code generation, language translation, and data analysis, marking a substantial departure from previous predictive algorithms primarily focused on structured or image data. This strong data dependency positions LLMs as compressors of vast human data, as suggested by the influential metaphor "ChatGPT is a blurry JPEG of the web" (Chiang, 2023). Second, the generative and stochastic nature of LLMs, which arises from next-token prediction, makes the models themselves random, with outputs necessarily involving variability and uncertainty (Huang et al., 2025). Together, these characteristics lead to numerous statistical challenges and opportunities for principled treatment of variability, uncertainty, calibration, and inference, especially when LLMs are used in high-stakes decision-making.

Moreover, we contend that for many LLM-related problems, statistics may be not just useful but sometimes perhaps the only viable approach. To make this case, we highlight the black-box status of LLMs, meaning it is difficult to understand how they arrive at their decisions. While the degree of black-boxness can vary, LLMs are arguably among the most complex digital systems ever constructed. We will provide evidence that this black-box nature is not merely a temporary state but likely to persist. Consequently, deriving LLM behavior from first principles via mathematical modeling, as is often possible in physics, appears highly challenging, if not infeasible. As such, when direct mathematical modeling is impractical, statistical modeling often offers a flexible and effective approach to shedding light on complex systems. Indeed, statistical approaches allow us to posit and test relationships between observable variables (like inputs and outputs) and potentially unobservable latent factors, even before their internal mechanisms are fully understood (Candès and Sabatti, 2020). Examples include understanding how different data mixtures influence model capabilities or quantifying the uncertainty of the outputs of LLMs using conformal prediction (Liu et al., 2025b; Mohri and Hashimoto, 2024; Cherian et al., 2024).

Having argued for the fundamental necessity of statistics for LLMs, we outline several statistical research avenues on LLMs, which are illustrative rather than exhaustive, reflecting the rapid evolu-

tion of the field. These range from the formulation of principled methods for aligning models with human values, to leveraging the models' probabilistic outputs for tasks like content verification, to rigorously evaluating model capabilities and quantifying output uncertainty, and finally to optimizing the central role of data in shaping LLM performance. Given that LLMs operate in an end-to-end manner—directly processing text inputs to produce text outputs that can inform decisions or trigger actions—they permeate an ever-wider range of applications, continuously generating new statistical challenges and opportunities. Echoing Efron's sentiment about algorithmic developments originating outside statistics, "future progress, especially in scientific applicability, will depend heavily on us" (Efron, 2020).

2 LLMs as Statistical Models

While designed as a predictive method where the underlying data-generating mechanism is treated as unknown, LLMs exhibit characteristics that distinguish them significantly from many other predictive algorithms (Breiman, 2001; Efron, 2020). Predictive methods like decision trees or support vector machines often rely on feature engineering or, equivalently, carefully chosen kernels, whereas modern LLMs operate in a nearly end-to-end data-centric manner, trained directly on vast quantities of raw data with minimal human intervention beyond data selection and cleaning.[4]

Owing to this strong data dependency, the capabilities of LLMs are largely determined by the properties and scale of the training data (Thrush et al., 2024). This fact is quantitatively captured by scaling laws, which demonstrate a predictable relationship between model capabilities and the volume of training data during the pre-training phase (Kaplan et al., 2020). In essence, scaling laws suggest that LLMs are what they are trained on, elevating data to the most critical component of their development. As evidence, while open-source LLM developers often release model weights and sometimes some technical details of their training processes, the composition and preparation of their training datasets are almost never disclosed (Grattafiori et al., 2024; Liu et al., 2024).

The pivotal role of data continues after the pre-training phase. To achieve proficiency in complex tasks, LLMs typically undergo specialized post-training (Ouyang et al., 2022). This necessitates the use of vast corpora of high-quality and meticulously annotated examples to impart nuanced understanding and desired behaviors to LLMs. This has led to the rise of a substantial data-labeling and annotation industry for training, refining, and aligning AI models.

One might argue that this data-centricity is shared to some extent with smaller deep learning models, not solely LLMs. Yet LLMs stand apart due to two characteristics:

Anything as numeric. LLMs operate directly on unstructured information—plain language, code, numbers, or even symbolic mathematics—by converting diverse data types into high-dimensional numeric vectors often thought to lie in a "semantic" space (Mikolov et al., 2013). This enables flex-

[4]Indeed, data curation is nontrivial in practice, but from a methodological perspective, no detailed human knowledge of linguistic constructs is explicitly hard-coded into LLMs at training time.

ible execution of transformations within this semantic space, which LLMs ultimately map back to text. This allows LLMs to process anything representable as text almost as readily as regression models handle numbers. In effect, LLMs instantiate a general-purpose engine operating on numeric representations of virtually all forms of text data. Indeed, Geoffrey Hinton metaphorically described GPT-4 as emerging like a butterfly from billions of nuggets of understanding accumulated throughout human history (Hinton, 2023).

Stochastic nature of generation. The predominant training paradigm for modern LLMs is next-token prediction (Radford et al., 2018).[5] Crucially, this generative process is inherently stochastic, depending not only on the prior context but also on the data used for training LLMs. This randomness is not merely an artifact but arguably a necessity for modeling human language, which itself—whether in stories, manuals, essays, creative writing, code, or mathematical proofs—is generative and rarely follows a deterministic path. This contrasts sharply with many earlier deep learning applications, particularly in classification, where outputs often correspond to fixed, deterministic ground truths.

These two characteristics—the universal processing of data through numerical embedding and the inherent stochasticity in generation—arguably position LLMs in a way that resonates more closely with data-modeling or inferential methods than typical predictive algorithms. Recognizing this allows us to appreciate the potential and necessity of statistical insights for LLMs beyond their initial purely predictive design. Consequently, we can effectively treat LLMs as *statistical models*.

Statistics becomes particularly relevant when a system involves dependence on data and the inference of patterns from that data. Indeed, the data-hungry nature of LLMs, coupled with their ability to process "anything as numeric," leads to complex data-dependent patterns. The sheer scale and heterogeneity of LLM training data present many opportunities and significant challenges from a statistical perspective. For instance, understanding how different data sources contribute to specific model capabilities (e.g., models trained on larger proportions of code tend to empirically exhibit stronger programming abilities) is crucial for optimizing data mixtures to achieve desired performance, in both the pre-training and post-training phases (Xie et al., 2023). Furthermore, as suggested by researchers like Ilya Sutskever, relying solely on existing human-generated data may already be insufficient (Sutskever, 2024), and therefore the generation of high-quality synthetic data becomes vital. This is a task where statistical principles for experimental design and data augmentation could prove invaluable.

Furthermore, the variability and uncertainty stemming from the stochastic nature of LLMs demand statistical analysis. Because LLM outputs are inherently variable, understanding and quantifying the uncertainty associated with their responses is critical. This is especially true in scientific applications where reproducibility is required, and in high-stakes decision-making, such as medical diagnostics, where the model's confidence level can drastically influence subsequent actions. Statis-

[5]While other objectives exist (e.g., masked language modeling in BERT) and new approaches are emerging (e.g., diffusion models for text), the core task remains generative.

tics offers a rich toolkit for analyzing variability, uncertainty, and miscalibration, and subsequently reducing them.

Conversely, the stochastic generation process not only presents challenges but also enables novel statistical techniques. Watermarking, for instance, leverages the model's probabilistic token generation to embed statistically detectable signals based on pseudorandomness, allowing for provable distinction between AI-generated and human-written text, potentially without compromising quality (Kirchenbauer et al., 2023). Another example is LLM alignment via reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), which uses the Bradley–Terry model (Bradley and Terry, 1952) to represent LLMs' preference distributions. Such techniques would be impossible if LLM generation were deterministic.

The need for statistical foundations is further amplified by the unprecedented breadth of LLM applications. LLMs are being integrated into coding assistants, writing tools, autonomous agents, scientific discovery platforms, medical information retrieval, and countless other domains. This rapid and widespread deployment into diverse, often novel environments constantly surfaces new challenges related to privacy, copyright, attribution, fairness, ethical considerations, and the need for mechanisms like machine unlearning (Cao and Yang, 2015) to remove specific sensitive knowledge. While addressing these complex issues for the trustworthy use of LLMs undoubtedly requires interdisciplinary collaboration, statistical insights and methodologies are particularly crucial and effective in developing solutions, with the additional advantage of typically being computationally light.

3 LLMs as Black-box Models

The data-centric and stochastic characteristics of LLMs already present a compelling case for incorporating statistical methodologies into their development and application. Nevertheless, one might wonder whether non-statistical approaches—for instance, purely mechanistic, first-principles methods—could also achieve similar goals. In this section, we deepen the argument by positing that for many crucial challenges surrounding LLMs, statistics may not merely be useful but often the most viable, or perhaps the only practical, path forward. This necessity arises primarily from the profound *black-box* nature of LLMs.

Whether a field leans heavily on statistical methodology often depends on the extent to which its underlying mechanisms are understood. Fields like classical physics, where fundamental principles are generally well-established, can often rely on deductive mathematical modeling to predict system behavior. While empirical data remains crucial for validation, data analysis often serves to confirm theories or estimate parameters within known models. Conversely, in fields like biology, particularly neuroscience, a vast number of unknowns and high-dimensional interactions render the internal workings inaccessible or too complex to model from first principles. Consequently, when confronting these kinds of black-box situations, researchers typically rely heavily on statistical inference to discern patterns, test hypotheses, and build useful predictive or explanatory models directly

from observational data, even without a complete mechanistic understanding. It is arguably this difference in mechanism transparency, among other factors, that contributes to the reality that statistical methodologies are more prevalent in biomedical research than in physics.

We argue that LLMs, in their current state and likely trajectory, fall firmly into the category of complex systems where the black-box nature necessitates a reliance on statistical approaches. Their status as black boxes is likely a persistent feature arising from the following two factors:

Inherent complexity and huge scale. LLMs are among the most complex computational systems ever built. The Transformer architecture, upon which almost all proprietary LLMs are based, involves intricate compositions of a variety of components such as multi-head attention, layer normalization, gating mechanisms, and nonlinear transformations, interacting across billions to nearly trillions of parameters (Vaswani et al., 2017). Indeed, the sheer size of LLMs appears necessary. On the theoretical front, Bubeck and Sellke (2021) showed that neural networks satisfying certain regularity conditions necessitate a vast number of parameters. Empirically, scaling laws further confirm that performance consistently improves with model size (Kaplan et al., 2020). This immense scale makes a detailed, analytical understanding from bottom-up principles practically intractable.

Non-uniqueness of architectures and optimizers. The black-box nature is compounded by the fact that there is not one single "correct" architecture for LLMs achieving high performance. While the Transformer has been dominant, simplified Transformer and mixture-of-experts Transformer variants are shown empirically to work well (He and Hofmann, 2024; Dai et al., 2024). Moreover, non-attention-based architectures like state-space models (Mamba) (Gu and Dao, 2024), recurrent structures (RWKV) (Peng et al., 2023), and even potentially large-scale LSTMs with sufficient memory show promise or competitive results (Schmidt, 2023). Similarly, various optimization algorithms—including Adam (Kingma and Ba, 2015), SGD, AdamW (Loshchilov and Hutter, 2019), Shampoo (Gupta et al., 2018), and the more recent Muon optimizer (Jordan et al., 2024)—have been effective in training these massive models. This lack of a uniquely optimal design for LLMs is reflective of the empirical trial-and-error approach driving both architectural and algorithmic innovations. Indeed, this empirical flexibility fosters co-adaptation between architectures and optimizers. Interestingly, a popular hypothesis within the AI community posits that Adam's effectiveness may partially result from neural architectures being inadvertently "overfitted" to its optimization characteristics (Orabona, 2020).

The confluence of immense complexity, necessarily large scale, and non-unique design makes it highly improbable that we can understand the behavior of LLMs from neatly closed-form laws in the way physicists often model physical phenomena. LLMs are thus de facto black boxes and likely to remain so for the foreseeable future (see Section 5 for elaboration). As Stephen Wolfram has argued, complex systems may sometimes be computationally irreducible, meaning their behavior cannot be predicted by simple, interpretable rules (Wolfram and Gad-el Hak, 2003). Consequently, attempts to build comprehensive mathematical theories to unveil the inner workings of LLMs are

likely to face fundamental challenges, if not prove impossible in practice.[6]

When dealing with a complex, stochastic black-box system where the true underlying process is unknown or intractable, we must resort to studying the system through its inputs and outputs, along with any available intermediate measurements and potentially latent factors. This necessitates the use of approximate and data-driven models to capture observed behaviors. Such models, built from data to approximate an unknown or intractable underlying process, are inherently statistical. The resulting statistical models are, in the spirit of George Box, necessarily wrong in the sense of being incomplete, yet potentially useful for prediction, understanding correlations, and guiding development (Box, 1976). Indeed, this perspective resonates with Alexei Efros' call to treat AI research more like experimental biology, focusing on using statistical methodology to make progress based on empirical observation, hypothesis testing, and data-driven modeling (Efros, 2023).

4 Statistical Topics on LLMs

In this section, we illustrate a number of research directions where we believe statistical principles can directly enhance both the development and application of LLMs. These directions either exploit the generative and data-driven essence of LLMs (Section 2) or take the viewpoint of LLMs as complex black-box systems (Section 3). This list is not exhaustive, reflecting the dynamic nature of the field, but serves to highlight the breadth of opportunities, and we anticipate that additional statistical challenges will arise as LLM capabilities and deployment settings continue to evolve. Moreover, many of these research areas demand modest computational resources, often requiring only API access to existing models, thus making them accessible to many researchers.

LLM alignment. Alignment is the process of steering AI models toward human preferences, intended goals, and ethical principles. Because human preferences and ethics can often be represented quantitatively via statistical models, statistical principles naturally arise for developing principled and trustworthy alignment methods.

- *Alignment from human feedback:* The technique of RLHF involves training a reward model based on human comparisons of LLM outputs (Ouyang et al., 2022). Formally, the preference distribution between two possible responses y, y' to a prompt x is modeled using the Bradley–Terry model:

$$\mathbb{P}(y \text{ is preferred over } y'|x) = \frac{\mathrm{e}^{r(x,y)}}{\mathrm{e}^{r(x,y)} + \mathrm{e}^{r(x,y')}},$$

  where the reward $r(x, y)$ is trained from pairwise comparison data from human labelers using maximum likelihood estimation. The LLM is then fine-tuned using the reward model to maximize the expected reward, subject to a constraint that penalizes excessive deviation from

[6]For completeness, we recognize efforts in mechanistic interpretability that aim to shed light on internal computations. These investigations, however, are computationally intensive and often yield localized or partial insights rather than a comprehensive, predictive approach to demystifying the entire system (Elhage et al., 2021).

the reference model. Owing to the stochasticity inherent in the Bradley–Terry model with noisy human feedback, this process is statistical in nature. This offers numerous opportunities for statisticians to analyze reference model misspecification, sample efficiency of preference data collection, and the generalization of learned preferences (Zhu et al., 2023; Chakraborty et al., 2024; Swamy et al., 2025; Ye et al., 2025). For example, recent studies demonstrate that the current approaches to RLHF inadequately represent the full spectrum of human preferences and can introduce statistical bias during fine-tuning (Liu et al., 2025a; Xiao et al., 2024).

- *Privacy and machine unlearning:* LLMs trained on vast datasets may inadvertently memorize and expose sensitive or copyrighted information. Differential privacy provides a rigorous mathematical framework that offers statistical guarantees against information leakage by introducing controlled noise during training (Dwork et al., 2006). A critical statistical consideration, therefore, is to optimize the trade-off between privacy protection and model utility, particularly for proprietary LLMs. Recent research has focused on enhancing this trade-off by implementing differential privacy during the fine-tuning phase of pre-trained models that were initially trained on public data (Li et al., 2022). However, further research is needed to meet the level of trade-off required by proprietary LLMs. A related area is machine unlearning, which aims to efficiently eliminate the influence of specific data points (e.g., due to privacy requests or copyright concerns) without retraining (Cao and Yang, 2015). This approach presents significant statistical challenges in precisely defining and verifying effective "forgetting" while preserving model capabilities (Yao et al., 2024; Zhang et al., 2024b).

- *Fairness:* LLMs can inherit and amplify societal biases present in their training data, leading to unfair or discriminatory outputs across different demographic groups (e.g., based on gender, race, religion) (Santurkar et al., 2023; Kotek et al., 2023). While addressing fairness is a complex socio-technical problem, statistics provides indispensable tools for defining, measuring, and mitigating bias. This includes developing quantitative fairness metrics, auditing models for systematic biases using statistical tests, and incorporating fairness considerations into various stages of the LLM pipeline—from data curation and pre-training objectives to alignment processes or even directly during generation of outputs (Zhang et al., 2024a; Chakraborty et al., 2024).

Exploiting the generative interface. The autoregressive nature of LLMs, specifically, next-token prediction, allows one to treat LLMs as a black-box machine that outputs a multinomial distribution, which is used to sample the next token. This probabilistic viewpoint allows us to develop statistical methods by leveraging the sampling properties of multinomial distributions, without needing to delve into the complexity of how the Transformer architecture computes the distributions.

- *Watermarking:* To distinguish LLM-generated text from human-written content, watermark-

ing techniques embed statistically detectable signals into the generation process based on cryptographic pseudorandomness (Kirchenbauer et al., 2023; Aaronson, 2023). Formally, the next token w_{t+1} is decoded as $\mathcal{S}(\boldsymbol{P}_t, \zeta_t)$, where the decoder $\mathcal{S}$ is deterministic or can incorporate sampling randomness, $\boldsymbol{P}_t$ is the multinomial distribution for drawing the $(t+1)^{\text{st}}$ token, and ζ_t is a pseudorandom variable that can be computed from the preceding context and a private key. The detection problem can be naturally framed as statistical hypothesis testing, based on the observation that when text is not watermarked (under the null hypothesis), w_{t+1} is independent of $\mathcal{S}(\boldsymbol{P}_t, \zeta_t)$, while when the text is watermarked (under the alternative hypothesis), $w_{t+1} = \mathcal{S}(\boldsymbol{P}_t, \zeta_t)$. The latter case necessarily induces dependence between the tokens and pseudorandom variables, even without knowledge of the multinomial distributions. This framework opens avenues for applying statistical decision theory to design watermarking schemes and detection rules that achieve optimal or near-optimal detection performance while preserving text quality (Li et al., 2025). A significant practical challenge involves ensuring robustness against adversarial modifications, such as paraphrasing or translation, necessitating the development of schemes with provable statistical guarantees under such attacks (Pang et al., 2024; Li et al., 2024a). Additionally, there is substantial interest in leveraging watermarking techniques for other purposes, such as detecting data misappropriation (Cai et al., 2025).

- *Speculative Sampling:* This technique accelerates text generation of LLMs by employing a smaller, faster "draft" model to propose candidate tokens, which are subsequently accepted or rejected by the larger "target" model based on a comparison of their respective output distributions (Leviathan et al., 2023; Chen et al., 2023). To add detail on speculative sampling, let $\boldsymbol{P}_t$ and $\boldsymbol{Q}_t$ denote the multinomial distributions of the larger and smaller LLMs, respectively, for predicting the t^{th} token—assuming the index of the last token in the prefix is 0. Denote by $x_1, \ldots, x_T$ the tokens sequentially proposed by the smaller LLM conditioned on the preceding tokens. Then, beginning from $t = 1$, the larger LLM accepts token x_t with probability $\min\{1, P_t(x_t)/Q_t(x_t)\}$, and terminates the process when the first rejection occurs. The final step of speculative sampling appends one additional token for this epoch by sampling from $\max\{\mathbf{0}, \boldsymbol{P}_t - \boldsymbol{Q}_t\}$ after normalization, where t denotes the position at which the smaller LLM's proposed token is first rejected. As is clear, speculative sampling is a form of rejection sampling that achieves the maximum coupling between the prediction probabilities of the two LLMs.

  The rationale behind this technique is that, for the larger LLM, evaluating the probability of a proposed token is computationally more efficient than decoding a token, which typically requires processing all possible tokens in the (extensively large) vocabulary. Consequently, the efficiency gain depends critically on the acceptance rate, which is inherently a statistical quantity depending on how the target LLM's probability distributions align with those of the smaller one. Statistical analysis is crucial for optimizing the trade-off between the com-

putational cost of the draft model and the expected speedup from the acceptance rate. For instance, this statistical technique has been implemented in DeepSeek V3 for efficient multi-token prediction (Li et al., 2024b; Liu et al., 2024) and can be integrated with watermarking (Hu and Huang, 2024). More opportunities exist for incorporating statistical insights into adaptive variations of this technique and integrating it with other methods.

- *Tokenization:* Tokenization breaks down text into discrete tokens to form the categories from which LLMs sample. Tokenization fundamentally impacts the statistical properties of the data fed into the model and the distributions it outputs. However, most current tokenizers (e.g., byte-pair encoding (Gage, 1994)) are based on heuristic compression algorithms and lack statistical guarantees. There is a need for statistically principled tokenization methods that optimize for criteria like information rate or minimal sequence length across diverse text types. Furthermore, statistical analysis is required to understand how tokenization efficiency and potential induced biases vary across different languages, domains (e.g., code, scientific literature), and demographic groups (Phan et al., 2024; Yang et al., 2024).

Assessment of LLM behavior. Understanding and quantifying the reliability, limitations, and capabilities of LLMs presents significant challenges, exacerbated by their stochastic and black-box nature. This inherently necessitates statistical modeling for assessing LLM behavior with confidence statements.

- *Uncertainty quantification and calibration:* LLM outputs exhibit uncertainty stemming from both inherent randomness in the generation process and knowledge limitations (Yadkori et al., 2024). For trustworthy applications, particularly in high-stakes scenarios, quantifying the uncertainty of LLM outputs becomes crucial. Among various approaches addressing uncertainty, conformal prediction has emerged as a statistically rigorous and flexible methodology for providing prediction sets with distribution-free coverage guarantees—a characteristic particularly suited to the black-box nature of LLMs (Mohri and Hashimoto, 2024; Cherian et al., 2024). Furthermore, the utility of uncertainty estimates depends on calibration—the degree to which a model's expressed confidence aligns with its actual accuracy. Given that aligned LLMs are often miscalibrated (Achiam et al., 2023), developing methods that simultaneously quantify uncertainty and restore calibration is an important research direction (Huang et al., 2024; Xiao et al., 2025; Wang et al., 2025; Liu et al., 2025c).

- *Evaluation:* Assessing LLM capabilities across diverse tasks using benchmarks such as MMLU (Hendrycks et al., 2021), TruthfulQA (Lin et al., 2022), and GSM8K (Cobbe et al., 2021) is essential not only for tracking progress but also for guiding AI development (Silver and Sutton, 2025; Yao, 2025; Gao et al., 2025). However, the probabilistic and complex nature of LLMs introduces substantial statistical challenges in evaluation. Statistically grounded methods are required to quantify the variance and reliability of evaluation scores (Miller, 2024; Polo et al., 2024), with statistical models such as item response theory being employed for this purpose

(Madaan et al., 2024). Nevertheless, this area faces an "evaluation crisis," where reported benchmark scores frequently inflate perceived capabilities, partly due to evaluation gaming—the process of optimizing models specifically for benchmark performance (Khomenko, 2025). This phenomenon bears resemblance to statistical issues such as p-hacking. Consequently, we need rigorous statistical principles for robust measurement and protection against overfitting to evaluation datasets.

The central role of data. The capabilities of LLMs are fundamentally determined by the data used in pre-training and fine-tuning (Yue et al., 2025). This gives rise to numerous statistical challenges related to understanding and optimizing the relationship between data characteristics and model performance.

- *Data mixture and attribution:* An essential challenge is determining the optimal composition of diverse data sources (e.g., web text, books, code, scientific papers) to train an LLM that achieves specific desired capabilities, often under resource constraints (Xie et al., 2023). While heuristic understanding exists—for instance, that a higher proportion of code in the training data generally leads to stronger coding abilities—the complex, high-dimensional relationship between data mixture and emergent abilities is largely unknown (Thrush et al., 2024). Statistical modeling, particularly regression-based approaches, offers a simple yet effective approach to investigating these dependencies (Liu et al., 2025b). Closely related to data mixture is the problem of data attribution, which seeks to identify the specific training samples that most influence a particular model output or behavior. Data attribution is crucial for addressing legal concerns like copyright infringement and enhancing transparency, but poses significant challenges due to the black-box nature of LLMs. Statistical techniques such as influence functions (Koh and Liang, 2017) and kernel approximation (Park et al., 2023) have been used to address these challenges for relatively small-scale models, yet significant research effort is needed to adapt and scale these methods effectively for large LLMs.

- *Synthetic data and model collapse:* As the scale of LLMs continues to grow, relying solely on existing human-generated data ("the fossil fuel of AI," as put by Ilya Sutskever in his keynote speech at NeurIPS 2024) may become insufficient or cost-prohibitive. Consequently, synthetic data is becoming increasingly vital for its scalability and cost-effectiveness (Eldan and Li, 2023; Adler et al., 2024; Yang et al., 2025; Tian and Shen, 2025). Statistics offers valuable tools for guiding the synthetic data generation process, including methods for assessing data quality, controlling distributional properties to match desired targets, and designing efficient data augmentation strategies (Angelopoulos et al., 2023). However, a major risk arises from recursively training models on their own synthetic outputs, which can lead to a degradation of model quality, loss of diversity, and divergence from the true data distribution—a phenomenon termed model collapse (Shumailov et al., 2024). Understanding the underlying mechanisms driving model collapse and developing statistically sound methodologies to mitigate it, po-

tentially by adaptively mixing real and synthetic data or by imposing specific distributional constraints, represents a promising research direction where statistical insights are needed (Gerstgrasser et al., 2024; Dey and Donoho, 2024).

- *Scaling laws:* Scaling laws are empirical observations that quantitatively relate an LLM's performance to factors such as the size of the training dataset, the number of model parameters, and the computational resources allocated for training (Kaplan et al., 2020). Among the various forms of scaling laws, Hoffmann et al. (2022) demonstrated that

$$L = E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}}$$

effectively captures how the pre-training loss L depends on the number N of model parameters and the number D of training tokens, where E denotes the entropy of natural text and A, α, B, and β are constants. These laws offer significant practical value by enabling researchers to predict potential performance gains from increased scale, thereby guiding resource allocation strategies for training progressively larger models without requiring exhaustive experimentation (Achiam et al., 2023). Scaling laws are fundamentally statistical, as they model empirically observed relationships between the scale of training resources and model performance. From a theoretical perspective, however, scaling laws present intriguing questions for statisticians. The observation that model performance continues to improve with increasing model size N, seemingly without saturation even when N becomes extremely large, challenges classical statistical learning theories that predict model performance saturation once complexity exceeds the intrinsic dimensionality of the data. Investigating the statistical underpinnings of these empirical laws, potentially through the lens of nonparametric estimation or approximation theory in high dimensions, may yield deeper insights into the learning dynamics of LLMs and potentially lead to novel statistical methodologies.

Other research directions. The rapidly evolving landscape of LLMs continues to generate novel statistical challenges that extend beyond the aforementioned categories, many of which remain incompletely formulated from a statistical perspective and thus present opportunities for contributions from our community. In the development of small LLMs, which are needed for deployment in edge devices, empirical evidence has revealed that knowledge distillation from larger LLMs often outperforms training from scratch (Guo et al., 2025), which calls for developing statistically efficient distillation methods. Accordingly, there is a need for owners of proprietary LLMs to develop sampling strategies that limit the effectiveness of distillation by business competitors (Savani et al., 2025). The recent emergence of reasoning models that employ latent intermediate steps via chain-of-thought (Wei et al., 2022) suggests that latent variable modeling might be valuable for understanding the mechanisms underlying the effectiveness of latent reasoning outputs (Muennighoff et al., 2025). Furthermore, the very recent advent of diffusion-based LLMs presents an opportunity for statistical analysis to elucidate the fundamental comparisons between the autoregressive and

diffusion-based strategies (Nie et al., 2025; Dou et al., 2024). Moreover, as LLMs are frequently deployed as evolving API-based services, the development of statistically grounded techniques to detect unannounced updates and consequent behavioral shifts is critical for ensuring the reliability of downstream applications (Dima et al., 2025). Finally, the modification of multinomial distributions for next-token prediction through a Bayesian perspective represents another frontier where statistical insights can directly inform LLM development (Zhuang et al., 2025).

5 Discussion

The past decade and a half, starting with the advent of AlexNet, has witnessed a remarkable advancement of purely predictive algorithms, with LLMs emerging as perhaps the most striking example. While LLMs share many defining aspects of purely predictive algorithms, their versatility in handling a variety of data types—especially unstructured text—and flexibility in applications across unprecedentedly diverse domains clearly distinguish LLMs from earlier purely predictive models, including pre-Transformer neural networks used in classification. Indeed, it may be more accurate to consider LLMs as enabling a new data-processing paradigm that converts and unifies diverse text-based inputs into a numeric form that can then be transformed and generated back into text, creating new forms of data amenable to analysis. Analogous to how genome-wide association studies once catalyzed the development of high-dimensional statistics, there are good reasons to believe that the continued progress of LLMs will open up an entire class of problems for which statistical methodologies will thrive.

Even taking a narrower scope, we argue further that classical inferential statistical principles—particularly those aligned with the "data-modeling" culture (Breiman, 2001) and the "estimation and attribution" perspective (Efron, 2020)—are becoming increasingly relevant for LLMs. The inherently stochastic nature of LLM generation makes statistical approaches suitable for quantifying and understanding uncertainty and variability. Moreover, while their architecture is in principle known, the stronger case for a statistical viewpoint emerges because the nearly trillions of parameters that LLMs operate on lack straightforward interpretation. Indeed, when faced with a system exhibiting black-box complexity, approximating its behavior through data-driven, testable, and refutable statistical models is perhaps the only tractable and effective approach, as statistical methods have a long history of proving effective even when underlying mechanisms are not fully understood. This black-box complexity requires that we formulate solutions for some use cases of LLMs—especially those requiring stability, robustness, and interpretability—using smaller, interpretable probabilistic models or inferring relationships between interpretable factors. Indeed, current (and most likely future) approaches—whether mitigating alignment biases, verifying content origin, quantifying the reliability of generated content, or assessing the influence of specific data subsets—often involve statistical reasoning, hypothesis testing, or parameter estimation within a relevant probabilistic framework.

One may hope that the internal workings of LLMs will eventually become transparent and

amenable to purely mechanistic analysis (Elhage et al., 2021), which would perhaps eliminate the need for statistical approaches to the problems we present in Section 4. Yet strong evidence suggests otherwise.

Hypothesis of perpetual black-box state-of-the-art models. A crucial argument supporting our claim is that state-of-the-art models are constantly evolving, driven by empirical gains achieved through increasing scale, architectural modifications often optimized for hardware efficiency, and new training heuristics. Consequently, the theoretical understanding of these models lags significantly behind practitioner-led advancements, a trend observable since the introduction of AlexNet (Su, 2024), leading to a widening gap between what can be rigorously understood and the capabilities of the latest LLMs.

The persistent black-box status of LLMs suggests that a single, unifying "grand statistical foundation for LLMs" is unlikely to emerge. Consequently, statistical research in this area will likely proceed in a bottom-up fashion, driven by the need to solve specific problems and address particular applications. This problem-driven approach is expected to yield a mosaic of specialized statistical frameworks and techniques tailored to distinct challenges. While this implies diversity in methodology, our personal experience working on several LLM-related problems shows that fundamental statistical thinking and inferential reasoning remain consistently crucial and effective across different contexts. This lack of a single unifying foundation presents a wealth of research opportunities for statisticians with varied skillsets.

Although this paper primarily focuses on inferential or classical statistics, this emphasis does not imply that statistical research on LLMs should be exclusively inferential. Rather, progress will likely require a blend of both inferential and predictive statistical approaches. Indeed, the boundary between these two statistical cultures is often blurry, especially in practice. Furthermore, advancing statistical methodology in LLM research requires recognizing the interplay between statistics and data science, particularly the significant engineering component inherent in the latter (Donoho, 2017). Embracing data science practices, such as robust data and code sharing, will also be vital for accelerating progress across this diverse research landscape and maximizing the collective impact of statistical contributions (Donoho, 2024).

While we have argued for the utility and necessity of statistical contributions to LLM development and application, we have not addressed the matter of timing. It remains uncertain whether generative AI, particularly LLMs, will lead to artificial general intelligence. However, it appears highly probable that these technologies will constitute a lasting and significant component of the future AI landscape. This presents a significant and timely opportunity for the statistics community (Lin et al., 2025). However, the risk of missing the chance to shape these AI technologies might arise if our community delays active engagement, since researchers in other fields, such as computer science—where younger generations often receive substantial statistical training—may develop these solutions on their own. While these contributions would still leverage statistical ideas, they might adopt a different flavor or lack the rigor that principled statistical approaches could

provide. Waiting for the field of LLMs to "stabilize" or for problems to become "well-defined" risks allowing non-statistical or less-statistically grounded methodologies to occupy domains where principled statistical approaches would be more appropriate. However, the potential "equilibrium" is not necessarily *unique*. Principled statistical approaches, if they come late, might not necessarily replace less-statistically grounded approaches that arrive earlier, especially in a field like LLMs involving significant engineering, scientific, and business considerations. Therefore, it is crucial for statisticians to be proactive to ensure that the development and application of LLMs benefit fully from the depth and rigor of statistical science.

Acknowledgments

The author would like to thank Xiang Li for comments on an earlier version of the manuscript. This work was supported in part by NSF DMS-2310679, a Meta Faculty Research Award, and Wharton AI for Business.

References

S. Aaronson. Watermarking of large language models. `https://simons.berkeley.edu/talks/scott-aaronson-ut-austin-openai-2023-08-17`, August 2023. URL `https://simons.berkeley.edu/talks/scott-aaronson-ut-austin-openai-2023-08-17`.

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

B. Adler, N. Agarwal, A. Aithal, D. H. Anh, P. Bhattacharya, A. Brundyn, J. Casper, B. Catanzaro, S. Clay, J. Cohen, et al. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*, 2024.

A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, and T. Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 610–623, 2021.

G. E. Box. Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799, 1976.

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

L. Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

S. Bubeck and M. Sellke. A universal law of robustness via isoperimetry. *Advances in Neural Information Processing Systems*, 34:28811–28822, 2021.

S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.

Y. Cai, L. Li, and L. Zhang. A statistical hypothesis testing framework for data misappropriation detection in large language models. *arXiv preprint arXiv:2501.02441*, 2025.

E. Candès and C. Sabatti. Discussion of the paper "prediction, estimation, and attribution" by Bradley Efron. *Journal of the American Statistical Association*, 115(530):656–658, 2020.

Y. Cao and J. Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480. IEEE, 2015.

S. Chakraborty, J. Qiu, H. Yuan, A. Koppel, F. Huang, D. Manocha, A. S. Bedi, and M. Wang. MaxMin-RLHF: Alignment with diverse human preferences. In *International Conference on Machine Learning*, pages 6116–6135. PMLR, 2024.

C. Chen, S. Borgeaud, G. Irving, J.-B. Lespiau, L. Sifre, and J. Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023.

J. J. Cherian, I. Gibbs, and E. J. Candès. Large language model validity via enhanced conformal prediction methods. In *Advances in Neural Information Processing Systems*, volume 37, pages 114812–114842, 2024.

T. Chiang. ChatGPT is a blurry jpeg of the web, February 2023. URL `https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web`. Accessed: 2025-05-05.

K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

D. Dai, C. Deng, C. Zhao, R. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu, Z. Xie, Y. Li, P. Huang, F. Luo, C. Ruan, Z. Sui, and W. Liang. Deepseekmoe: Towards ultimate expert

specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1280–1297, 2024.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 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*, pages 4171–4186, 2019.

A. Dey and D. Donoho. Universality of the $\pi^2/6$ pathway in avoiding model collapse. *arXiv preprint arXiv:2410.22812*, 2024.

A. Dima, J. Foulds, S. Pan, and P. Feldman. You've changed: Detecting modification of black-box large language models. *arXiv preprint arXiv:2504.12335*, 2025.

D. Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4): 745–766, 2017.

D. Donoho. Data science at the singularity. *Harvard Data Science Review*, 6(1), 2024.

Z. Dou, S. Kotekal, Z. Xu, and H. H. Zhou. From optimal score matching to optimal sampling. *arXiv preprint arXiv:2409.07032*, 2024.

C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

B. Efron. Prediction, estimation, and attribution. *Journal of the American Statistical Association*, 115(530):636–655, 2020.

A. Efros. Ai is (mostly) an experimental science. work like a biologist rather than a mathematician. X (formerly Twitter), 7 2023. URL `https://x.com/turingbook/status/1679718448249864194`. Retrieved from X (formerly Twitter).

R. Eldan and Y. Li. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*, 2023.

N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

P. Gage. A new algorithm for data compression. *The C Users Journal*, 12(2):23–38, 1994.

T. Gao, J. Jin, Z. T. Ke, and G. Moryoussef. A comparison of DeepSeek and other LLMs. *arXiv preprint arXiv:2502.03688*, 2025.

M. Gerstgrasser, R. Schaeffer, A. Dey, R. Rafailov, H. Sleight, J. Hughes, T. Korbak, R. Agrawal, D. Pai, A. Gromov, D. A. Roberts, D. Yang, D. L. Donoho, and S. Koyejo. Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. In *First Conference on Language Modeling*, 2024.

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

V. Gupta, T. Koren, and Y. Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pages 1842–1850. PMLR, 2018.

B. He and T. Hofmann. Simplifying transformer blocks. In *ICLR 2024*, 2024.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

G. Hinton. GPT-4 as humanity's butterfly. `https://x.com/geoffreyhinton/status/1635739459764322330`, March 2023. Twitter post.

J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 30016–30030, 2022.

Z. Hu and H. Huang. Inevitable trade-off between watermark strength and speculative sampling efficiency for language models. *arXiv preprint arXiv:2410.20418*, 2024.

L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

X. Huang, S. Li, M. Yu, M. Sesia, H. Hassani, I. Lee, O. Bastani, and E. Dobriban. Uncertainty in language models: Assessment through rank-calibration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 284–312, 2024.

W. Ji, W. Yuan, E. Getzen, K. Cho, M. I. Jordan, S. Mei, J. E. Weston, W. J. Su, J. Xu, and L. Zhang. An overview of large language models for statisticians. *arXiv preprint arXiv:2502.17814*, 2025.

C. R. Jones and B. K. Bergen. Large language models pass the Turing test. *arXiv preprint arXiv:2503.23674*, 2025.

K. Jordan, Y. Jin, V. Boza, J. You, F. Cesista, L. Newhouse, and J. Bernstein. Muon: An optimizer for hidden layers in neural networks. `https://kellerjordan.github.io/posts/muon/`, 2024. Accessed: 2025-05-24.

J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

L. Khomenko. Too many AIs, 2025. URL `https://dev.to/leeaao/too-many-ais-24nb`.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein. A watermark for large language models. In *International Conference on Machine Learning*, pages 17061–17084. PMLR, 2023.

P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR, 2017.

H. Kotek, R. Dockum, and D. Sun. Gender bias and stereotypes in large language models. In *Proceedings of the ACM Collective Intelligence Conference*, pages 12–24, 2023.

J. Kottke. The octopus test for large language model AIs, 2023. URL `https://kottke.org/23/03/the-octopus-test-for-large-language-model-ais`. Accessed: 2025-05-05.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105, 2012.

Y. Leviathan, M. Kalman, and Y. Matias. Fast inference from transformers via speculative decoding. In *International Conference on Machine Learning*, pages 19274–19286. PMLR, 2023.

X. Li, F. Tramer, P. Liang, and T. Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations*, 2022.

X. Li, F. Ruan, H. Wang, Q. Long, and W. J. Su. Robust detection of watermarks for large language models under human edits. *arXiv preprint arXiv:2411.13868*, 2024a.

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

Y. Li, F. Wei, C. Zhang, and H. Zhang. EAGLE: Speculative sampling requires rethinking feature uncertainty. In *Proceedings of the 41st International Conference on Machine Learning*, pages 28935–28948, 2024b.

C. W. Lin and W. Zhu. Divergent llm adoption and heterogeneous convergence paths in research writing. *arXiv preprint arXiv:2504.13629*, 2025.

S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, 2022.

X. Lin, T. Cai, D. Donoho, H. Fu, T. Ke, J. Jin, X.-L. Meng, A. Qu, C. Shi, P. Song, Q. Sun, W. Wang, H. Wu, B. Yu, H. Zhang, T. Zheng, H. Zhou, J. Zhou, H. Zhu, and J. Zhu. Statistics and AI: A fireside conversation. *Harvard Data Science Review*, 7(2), 2025.

A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, et al. DeepSeek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

K. Liu, Q. Long, Z. Shi, W. J. Su, and J. Xiao. Statistical impossibility and possibility of aligning LLMs with human preferences: From Condorcet paradox to Nash equilibrium. *arXiv preprint arXiv:2503.10990*, 2025a.

Q. Liu, X. Zheng, N. Muennighoff, G. Zeng, L. Dou, T. Pang, J. Jiang, and M. Lin. Regmix: Data mixture as regression for language model pre-training. In *International Conference on Learning Representations*, 2025b.

X. Liu, T. Chen, L. Da, C. Chen, Z. Lin, and H. Wei. Uncertainty quantification and confidence calibration in large language models: A survey. *arXiv preprint arXiv:2503.15850*, 2025c.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

L. Madaan, A. K. Singh, R. Schaeffer, A. Poulton, S. Koyejo, P. Stenetorp, S. Narang, and D. Hupkes. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*, 2024.

C. Manning and H. Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

B. Matthew and H. Toner. The surprising power of next word prediction: Large language models explained, Part 1. Georgetown University, March 2024. URL `https://cset.georgetown.edu/article/`

the-surprising-power-of-next-word-prediction-large-language-models-explained-part-1/. Accessed May 5, 2025. Explains the mechanism and notes limitations are discussed in subsequent parts.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 2013.

E. Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024.

C. Mohri and T. Hashimoto. Language models with conformal factuality guarantees. In *International Conference on Machine Learning*, pages 36029–36047. PMLR, 2024.

N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candès, and T. Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

S. Nie, F. Zhu, Z. You, X. Zhang, J. Ou, J. Hu, J. Zhou, Y. Lin, J.-R. Wen, and C. Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

F. Orabona. Neural networks (maybe) evolved to make Adam the best optimizer, 2020. URL https://parameterfree.com/2020/12/06/neural-network-maybe-evolved-to-make-adam-the-best-optimizer/.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

Q. Pang, S. Hu, W. Zheng, and V. Smith. Attacking LLM watermarks by exploiting their strengths. In *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*, 2024.

S. M. Park, K. Georgiev, A. Ilyas, G. Leclerc, and A. Madry. TRAK: Attributing model behavior at scale. In *International Conference on Machine Learning*, pages 27074–27113. PMLR, 2023.

B. Peng, E. Alcaide, Q. G. Anthony, A. Albalak, S. Arcadinho, S. Biderman, H. Cao, X. Cheng, M. N. Chung, L. Derczynski, et al. RWKV: Reinventing RNNs for the Transformer era. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

B. Phan, M. Havasi, M. Muckley, and K. Ullrich. Understanding and mitigating tokenization bias in language models. *arXiv preprint arXiv:2406.16829*, 2024.

F. M. Polo, L. Weber, L. Choshen, Y. Sun, G. Xu, and M. Yurochkin. tinyBenchmarks: Evaluating llms with fewer examples. In *International Conference on Machine Learning*, 2024.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. 2018.

S. Santurkar, E. Durmus, F. Ladhak, C. Lee, P. Liang, and T. Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, pages 29971–30004. PMLR, 2023.

Y. Savani, A. Trockman, Z. Feng, A. Schwarzschild, A. Robey, M. Finzi, and J. Z. Kolter. Antidistillation sampling. *arXiv preprint arXiv:2504.13146*, 2025.

L. Schmidt. Are Transformers necessary? A data-centric view on generalization. Workshop on Large Language Models and Transformers at the Simons Institute for the Theory of Computing, University of California, Berkeley, 2023. URL `https://simons.berkeley.edu/talks/ludwig-schmidt-university-washington-2023-08-18`.

I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

D. Silver and R. S. Sutton. Welcome to the era of experience, 2025. Preprint of a chapter that will appear in the book Designing an Intelligence, published by MIT Press.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

W. J. Su. Envisioning future deep learning theories: Some basic concepts and characteristics. *Science China Information Sciences*, 67(10):203101, 2024.

I. Sutskever. Sequence to sequence learning with neural networks: what a decade. NeurIPS 2024 Keynote, 2024. URL `https://www.youtube.com/watch?v=1yvBqasHLZs`.

R. S. Sutton. The bitter lesson. `http://www.incompleteideas.net/IncIdeas/BitterLesson.html`, 2019. Accessed: 2024-05-24.

G. Swamy, S. Choudhury, W. Sun, Z. S. Wu, and J. A. Bagnell. All roads lead to likelihood: The value of reinforcement learning in fine-tuning. *arXiv preprint arXiv:2503.01067*, 2025.

T. Thrush, C. Potts, and T. Hashimoto. Improving pretraining data using perplexity correlations. *arXiv preprint arXiv:2409.05816*, 2024.

X. Tian and X. Shen. Conditional data synthesis augmentation. *arXiv preprint arXiv:2504.07426*, 2025.

X. Tu, J. Zou, W. Su, and L. Zhang. What should data science education do with large language models? *Harvard Data Science Review*, 6(1), 2024.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Z. Wang, Q. Wang, Y. Zhang, T. Chen, X. Zhu, X. Shi, and K. Xu. SConU: Selective conformal uncertainty in large language models. *arXiv preprint arXiv:2504.14154*, 2025.

J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

S. Wolfram and M. Gad-el Hak. A new kind of science. *Applied Mechanics Reviews*, 56(2):B18–B19, 2003.

J. Xiao, Z. Li, X. Xie, E. Getzen, C. Fang, Q. Long, and W. J. Su. On the algorithmic bias of aligning large language models with RLHF: Preference collapse and matching regularization. *arXiv preprint arXiv:2405.16455*, 2024.

J. Xiao, B. Hou, Z. Wang, R. Jin, Q. Long, W. J. Su, and L. Shen. Restoring calibration for aligned large language models: A calibration-aware fine-tuning approach. In *International Conference on Machine Learning*, 2025.

S. M. Xie, H. Pham, X. Dong, N. Du, H. Liu, Y. Lu, P. S. Liang, Q. V. Le, T. Ma, and A. W. Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36:69798–69818, 2023.

Y. A. Yadkori, I. Kuzborskij, A. György, and C. Szepesvári. To believe or not to believe your llm. *arXiv preprint arXiv:2406.02543*, 2024.

J. Yang, Z. Wang, Y. Lin, and Z. Zhao. Problematic tokens: Tokenizer bias in large language models. In *2024 IEEE International Conference on Big Data (BigData)*, pages 6387–6393. IEEE, 2024.

Z. Yang, N. Band, S. Li, E. Candès, and T. Hashimoto. Synthetic continued pretraining. In *International Conference on Learning Representations*, 2025.

S. Yao. The second half, 2025. URL `https://ysymyth.github.io/The-Second-Half/`. Online essay about AI development stages.

Y. Yao, X. Xu, and Y. Liu. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475, 2024.

K. Ye, H. Zhou, J. Zhu, F. Quinzan, and C. Shi. Robust reinforcement learning from human feedback for large language models fine-tuning. *arXiv preprint arXiv:2504.03784*, 2025.

M. Yousefi and J. Collins. Learning the bitter lesson: Empirical evidence from 20 years of CVPR proceedings. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 175–187, 2024.

Y. Yue, Z. Chen, R. Lu, A. Zhao, Z. Wang, S. Song, and G. Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*, 2025.

L. Zhang, A. Roth, and L. Zhang. Fair risk control: A generalized framework for calibrating multi-group fairness risks. In *International Conference on Machine Learning*, pages 59783–59805. PMLR, 2024a.

R. Zhang, L. Lin, Y. Bai, and S. Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024b.

B. Zhu, M. Jordan, and J. Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pages 43037–43067. PMLR, 2023.

Y. Zhuang, L. Liu, C. Singh, J. Shang, and J. Gao. Text generation beyond discrete token sampling. *arXiv preprint arXiv:2505.14827*, 2025.