# StatEval: A Comprehensive Benchmark for Large Language Models in Statistics

Yuchen Lu[*,1], Run Yang[*,1], Yichen Zhang[*,1], Shuguang Yu[*,1], Runpeng Dai[2], Wenxin E[1], Siran Gao[1], Yirui Huang[1], Chenjing Xi[1], Haibo Hu[1], Yueming Fu[1], Qinglan Yu[1], Xiaobing Wei[1], Jiani Gu[1], Rui Sun[1], Jiaxuan Jia[1], Fan Zhou[†,1]

[1]Shanghai University of Finance and Economics
[2]University of North Carolina at Chapel Hill
[*]Equal contribution [†]Corresponding author: zhoufan@mail.shufe.edu.cn

**Abstract**

Large language models (LLMs) have demonstrated remarkable advances in mathematical and logical reasoning, yet statistics, as a distinct and integrative discipline, remains underexplored in benchmarking efforts. To address this gap, we introduce **StatEval**, the first comprehensive benchmark dedicated to statistics, spanning both breadth and depth across difficulty levels. StatEval consists of 13,817 foundational problems covering undergraduate and graduate curricula, together with 2374 research-level proof tasks extracted from leading journals. To construct the benchmark, we design a scalable multi-agent pipeline with human-in-the-loop validation that automates large-scale problem extraction, rewriting, and quality control, while ensuring academic rigor. We further propose a robust evaluation framework tailored to both computational and proof-based tasks, enabling fine-grained assessment of reasoning ability. Experimental results reveal that while closed-source models such as GPT5-mini achieve below 57% on research-level problems, with open-source models performing significantly lower. These findings highlight the unique challenges of statistical reasoning and the limitations of current LLMs. We expect StatEval to serve as a rigorous benchmark for advancing statistical intelligence in large language models. All data and code are available on our web platform: https://stateval.github.io/.

*Keywords:* Statistical reasoning; Automated extraction, Proof verification; Multi-agent evaluation

# 1 Introduction

Large language models (LLMs) have advanced rapidly in recent years (Brown et al., 2020; Touvron et al., 2023), demonstrating remarkable progress in complex reasoning (Guo et al., 2025), fluent text generation, and even automated proof discovery (Yu et al., 2025). These advances have spurred growing adoption of LLMs across education, data science, and research, where they are increasingly used for tutoring, problem explanation, data analysis, and hypothesis formulation (Wu et al., 2021; Polu and Sutskever, 2020; Khan et al., 2023; Gao et al., 2023). However, despite their broad deployment in quantitative domains, the field of *statistics*, which forms the foundation of modern data-driven science, has received little attention in LLM evaluation.

Statistics differs fundamentally from other quantitative disciplines. Rather than focusing on symbolic manipulation or fixed-form computation, it emphasizes reasoning under uncertainty, connecting probability theory, inference, regression, Bayesian analysis, multivariate methods, and asymptotic theory into a unified framework. Yet existing large-scale LLM evaluations rarely cover these competencies: statistical problems account for less than 3% of recent reasoning benchmarks (Paster et al., 2025), and when included, they are typically treated as isolated probability puzzles without structured categorization or coverage of inferential reasoning (Gao et al., 2024). This gap makes it impossible to rigorously assess whether LLMs can function as capable statisticians or support data-driven scientific discovery.

To bridge this critical gap, we introduce **StatEval**, the first large-scale benchmark dedicated to evaluating large language models on *statistical reasoning*. With nearly **20,000 meticulously curated problems**, **StatEval** covers the entire spectrum of statistics, from basic undergraduate exercises to advanced research-level challenges, captures the full

breadth and depth of the discipline, as illustrated in Figure 1. Its unprecedented scale and comprehensive coverage make it among the largest reasoning benchmarks to date. A concise yet systematic scoring framework is further established to ensure reliable assessment of model performance across diverse statistical tasks.

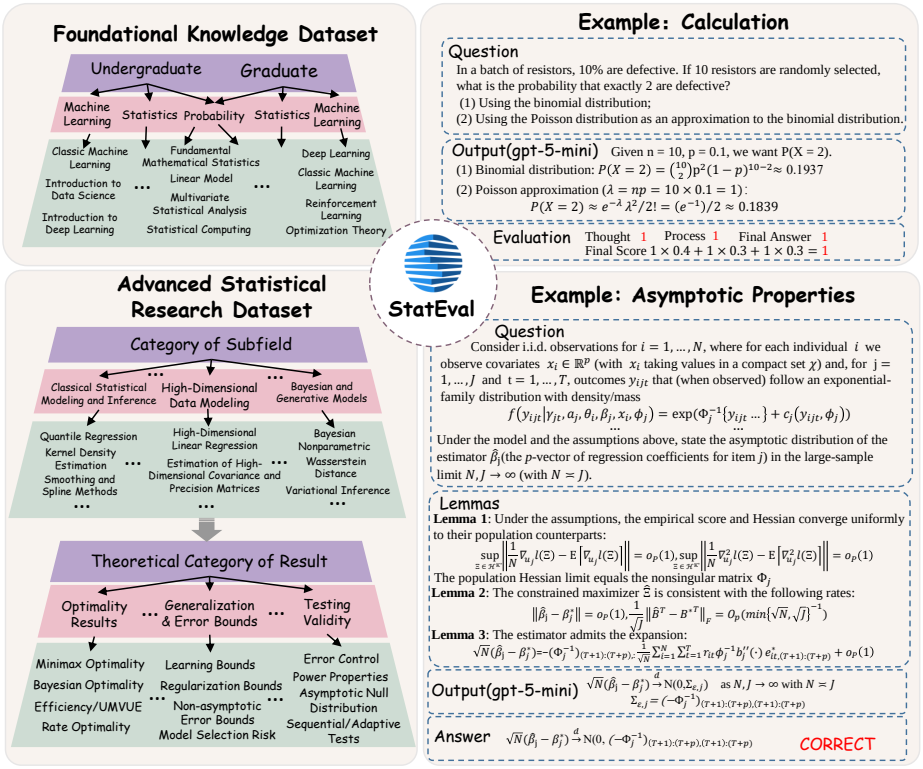



Figure 1: Overview of **StatEval**, illustrating the Foundational Knowledge Dataset, Advanced Statistical Research Dataset, and example evaluations on tasks such as statistical hypothesis testing and asymptotic properties of estimators.

StatEval is first organized along a difficulty axis, which defines two datasets. (i) The foundational knowledge dataset contains 13817 problems drawn from over **50** textbooks and course materials spanning undergraduate to doctoral levels, including both multiple-choice and short question–answer formats that assess LLMs' mastery of core statistical knowl-

edge. (ii) The Statistical research dataset consists of 2,374 verifiable, proof-based questions sourced from **18** top-tier peer-reviewed journals, designed to assess LLMs' reasoning ability on research-oriented tasks requiring rigorous derivations and proof-style solutions. Within each dataset, we further introduce a disciplinary axis, organized as a two-level taxonomy that covers more than 30 subdomains, including probability, inference, regression, Bayesian methods, multivariate statistics, asymptotic theory, experimental design, and machine learning. This dual-axis design enables fine-grained analysis of model performance across both foundational and advanced areas of statistics. All problems are presented in text-only format, ensuring that the evaluation directly probes reasoning ability rather than reliance on computational tools.

StatEval is built through a multi-agent LLM pipeline with human-in-the-loop verification, designed to balance scalability and rigor. The pipeline comprises four key agents: file conversion, context segmentation, problem generation, and quality control. Initially, a corpus of over 10,000 textbooks and research papers is converted into clean, LaTeX-compatible text using multi-modal models such as MinerU (Wang et al., 2024). Subsequently, the context segmentation agent applies an LLM-driven regular-expression framework to identify structural elements `Theorem`, `Lemma`, and `Example` and extract them with their relevant context while filtering unrelated material to reduce noise. Thereafter, the problem-generation agent reformulates extracted theorems into self-contained question–answer pairs under a rubric that enforces appropriate difficulty, single-answer focus, and quantitative verifiability. Finally, a GPT-5-based quality control agent verifies rubric adherence and contextual completeness before human experts review correctness and difficulty. Feedback from these reviews, especially representative failures, is periodically incorporated as few-shot exemplars to refine segmentation and generation in later iterations. This pipeline automates the transformation of scholarly materials into standardized, verifiable evaluation data, pro-

viding a scalable and continuously improving framework for benchmark construction in scientific domains.

Preliminary results demonstrate that StatEval presents a significant challenge. LLMs show steadily improving performance on fundamental statistical knowledge, particularly in domains such as machine learning and applied regression, yet their abilities remain uneven. Accuracy drops substantially on more basic but essential areas such as probability theory and linear models, which may be due to an overrepresentation of popular subjects in the training corpora, leaving foundational topics underexposed. At the Statistical research dataset, model performance is even more limited. Accuracy is significantly lower than on widely used datasets such as AIME25, underscoring the distinctive nature of statistics problems that require formal proofs and theoretical derivations. Even the most advanced proprietary systems, such as GPT5-mini and Gem2.5-flash, achieve only 57.62% and 51.14% accuracy, respectively, while the best open-source model reaches just 51.10%. These findings highlight both the intrinsic difficulty of statistical reasoning and the importance of StatEval in exposing capability imbalances that remain hidden in narrower benchmarks.

Our contributions are as follows:

1. We present StatEval, the first benchmark dedicated to statistics, providing comprehensive coverage across sub-disciplines and difficulty levels.

2. We develop a scalable multi-agent plus human-in-the-loop pipeline that automates the extraction of problems from scientific corpora while preserving academic rigor.

3. We conduct systematic experiments on state-of-the-art LLMs, uncovering significant gaps in their statistical reasoning and pointing to directions for future improvement.

The structure of this paper is organized as follows. Chapter 2 introduces the dataset and its distributional characteristics. Chapter 3 describes the data production pipeline.

Chapter 4 presents our scoring strategy. Chapter 5 reports the experimental results and analysis.

The Appendix offers supplementary information, including figures, experimental results, prompts, and further discussions. The dataset and evaluation code are publicly available; readers may access the full resources via our GitHub repository at https://github.com/StatEval/StatEval. In addition, detailed results of model evaluations can be found on our web platform at https://StatEval.github.io/.

# 2 Breakdown of StatEval

In this section, we provide a systematic description of StatEval, a benchmark that spans both fundamental and frontier aspects of statistics. StatEval is structured along the difficulty axis into two parts: a foundational knowledge dataset and a statistical research dataset. For each dataset, we detail its data sources, the distribution of problems, and the integration of the disciplinary axis, which supports fine-grained evaluation across a wide spectrum of statistical subdomains.

## 2.1 Foundational knowledge dataset

The foundational knowledge dataset evaluates large language models on their mastery of basic statistical knowledge and their ability to solve classic problems through foundational reasoning, spanning both undergraduate foundations and graduate-level training. To provide a comprehensive overview, we analyze the dataset from three complementary perspectives: source, question formats, and disciplinary structure.

**Data Sources.** The dataset consists of 13,817 problems, with 6,336 at the undergraduate and 7,481 at the graduate level. Problems are drawn from three primary sources: (i) 45 classical textbooks in statistics and related fields, which supply the majority of problems and ensure full curriculum coverage; (ii) over one thousand carefully verified, exam-style questions from graduate entrance examinations and curated exercise collections, introduced for the first time as an open benchmark; and (iii) recommended problems from publicly available courses at leading international universities, supplemented by online resources, enhancing topical diversity.

**Question Formats.** Complementing the source analysis, we categorize problems by format. The dataset includes 1,517 multiple-choice questions and 12,300 open-ended Question and Answer (QA) questions, Multiple-choice items, following conventions established by benchmarks such as MMLU and SciQA, primarily test factual recall and concept recognition. Open-ended questions, in contrast, require explicit derivations, detailed reasoning, or formal proofs, offering a rigorous evaluation of both reasoning ability and structured problem-solving skills.

**Disciplinary Structure.** To further organize the dataset content, problems are hierarchically classified. At the primary tier, problems are grouped into three domains: Probability, Statistics, and Machine Learning. At the second tier, each domain is divided into course-level subjects with distinct undergraduate and graduate coverage. For instance, undergraduate probability includes Elementary Probability, Stochastic Processes, and Elementary Time Series, whereas graduate-level probability extends to Advanced Probability (including stochastic processes), Advanced Time Series and Information Theory. The detailed distribution of problems across levels, domains, and subfields is summarized in

Table 4, with Figure 2 providing a visualization of category proportions alongside representative source textbooks. This layered design ensures that StatEval captures both foundational training and advanced theoretical rigor, providing a principled framework for fine-grained evaluation of LLMs across educational stages and statistical subdomains.

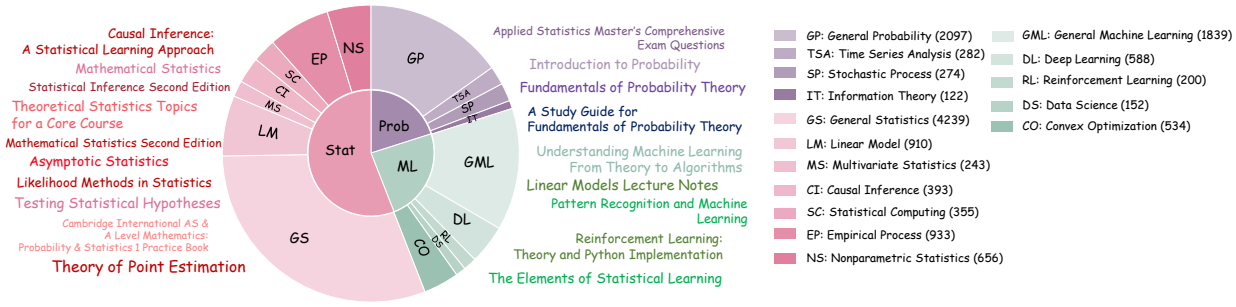



Figure 2: Disciplinary classification of foundational-level datasets

## 2.2 Statistical research dataset

The statistical research dataset is designed to benchmark the ability of LLMs to perform structured, multi-step reasoning on research-level statistical problems. Each task in the dataset is constructed to embody the core components of modern statistical reasoning, including precise assumptions, intermediate lemmas, and measure-theoretic or algorithmic arguments, all leading to rigorously justified conclusions.

**Data Sources.** This dataset contains 2,374 proof-based research tasks derived from 2,719 research articles published between 2020 and 2025 across 18 leading journals (Table 3). The corpus covers top-tier statistics venues (e.g., Annals of Statistics, Biometrika, JASA) and closely related fields in econometrics, probability, and machine learning.

**Question Formats.** Problem formulation is standardized by extracting peer-reviewed theorems, propositions, and lemmas from the selected articles and recasting them as research-level proof tasks. Each task is centered around a *single quantitative objective* , a precisely defined target such as finding an exact constant, a closed-form expression, a distributional form, a convergence rate, or an explicit bound with constants. This approach preserves the full complexity of original research problems while enabling objective, criteria-based solution verification. Integrating research-level difficulty with concrete endpoints supports reliable LLM output evaluation, consistent with best practices established in leading benchmarks (e.g., MATH-500, AIME). Further details on document collection, filtering, and quality control are provided in Section 3.

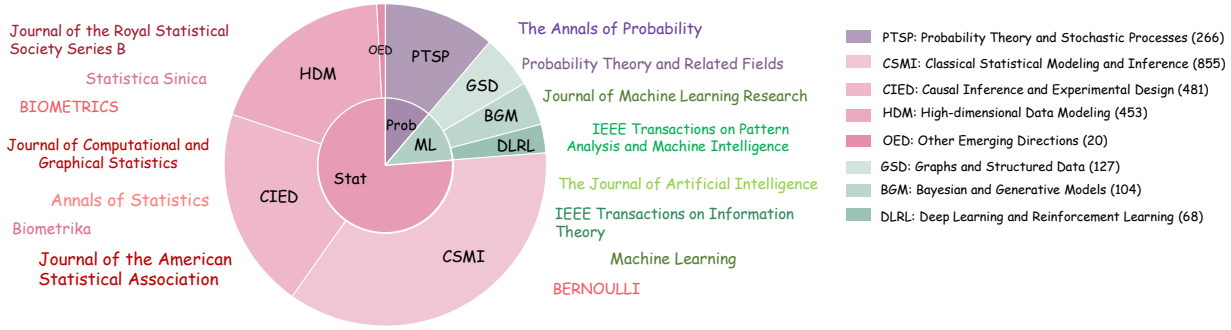



Figure 3: Disciplinary classification of statistical research datasets

**Disciplinary Structure.** In terms of organization, the disciplinary axis of the research dataset extends the taxonomy introduced in the foundational knowledge dataset (Section 2.1). While the fundamental knowledge is organized around curricular subjects in statistics education, the research dataset expands these into research-level categories in order to capture the complexity of modern statistical inquiry and enable finer-grained evaluation of reasoning ability.

At the primary tier, the original three domains: Probability, Statistics, and Machine

Learning are refined into eight specialized subdomains: Classical Statistical Modeling and Inference, Causal Inference and Experimental Design, High-dimensional Data Modeling, Probability Theory and Stochastic Processes, Graphs and Structured Data, Bayesian and Generative Models, Deep Learning and Reinforcement Learning, and Other Emerging Directions.

To further enhance granularity, a new second-tier classification is introduced, based on theoretical property type. This framework distinguishes among categories of theoretical results, including Asymptotic Properties, Identifiability and Consistency, Distributional Properties, Generalization and Error Bounds, Optimality Results, Testing Validity, Convergence and Stability, Structural Guarantees, among others, as summarized in Table 5.

This structure preserves continuity with the foundational dataset while introducing new axes that capture the breadth of frontier research, allowing StatEval to evaluate LLMs not only on fundamental subject knowledge but also on their ability to address advanced reasoning and derivation in cutting-edge statistical problems. Further details regarding the distribution across research areas and theoretical property categories are provided in Figure 3 and Appendix B .

# 3 Data Processing Pipeline

The data processing pipeline serves two complementary purposes: (i) to support large-scale, reliable extraction of Statistical research datasets from heterogeneous sources, and (ii) to conduct systematic quality inspection for the Foundational Knowledge dataset. The pipeline integrates LLMs with human-in-loop verification to ensure both scalability and precision, and incorporates few-shot feedbacks from human reviews to iteratively enhance performance. It consists of four core agents, each responsible for a distinct stage of the

end-to-end workflow (Figure 4).

1. **File-Conversion Agent:** This agent converts raw documents of diverse formats, PDFs, scanned files, and LaTeX sources into clean, structured text. We employ multimodal large language models (MLLMs), such as *MinerU*, for optical character recognition (OCR) and text reconstruction, converting outputs into LaTeXform. This standardization preserves mathematical expressions and notations while enabling seamless downstream processing and semantic parsing.

2. **Context-Segmentation Agent:** This agent extracts theorems together with their relevant context to support the subsequent problem-generation stage. It operates through an LLM-driven regular-expression framework, where the model dynamically generates and applies customized regular expressions to identify structural elements such as `Theorem`, `Lemma`, and `Example`. To ensure that each extracted fragment is self-contained, the agent also retrieves preceding definitions, assumptions, and other semantically related sections. We employ Gemini-Flash-Lite with a 1M-token context window to efficiently process long documents, capturing extended contextual dependencies while maintaining high throughput. This approach yields accurate and contextually complete theorem segments that provide reliable inputs for the following Problem-Generation Agent.

3. **Problem-Generation Agent:** This agent transforms extracted theorems and their surrounding context into question–answer (QA) pairs under a rubric-based generation framework. A concrete example of such generated QA pairs is provided in Appendix D for reference. As this stage requires precise mathematical reasoning rather than structural parsing, we employ the reasoning-optimized GPT-5 model to ensure the

faithful reconstruction of problems into self-contained and verifiable forms. Each question is generated according to the following rubric:

(a) **Appropriate difficulty:** Questions should align with advanced coursework or research-level challenges, neither trivial nor overly open-ended.

(b) **Self-containment:** Each problem must include sufficient background to be solved independently from the source text.

(c) **No leakage:** Questions must not reveal intermediate steps, proof structures, or final results.

(d) **Single-answer constraint:** Each problem should have exactly one well-defined solution.

(e) **Quantitative verifiability:** The answer must be a single quantitative object—an explicit number, closed-form expression, distribution, rate, or bound with constants, enabling objective evaluation.

This rubric ensures that all generated QA pairs preserve theoretical rigor while remaining suitable for automated assessment.

4. **Quality-Control Agent:** Each QA pair undergoes validation by an independent LLM-based quality-control agent, implemented with GPT-5, which re-evaluates adherence to the rubric checklist and checks for internal consistency between questions and answers. This stage serves as an automated filter, ensuring that only theoretically sound and structurally complete problems are passed to human review.

5. **Human Check & Feedback:** Samples that pass all automatic checks proceed to human verification by domain experts, who confirm semantic correctness, difficulty appropriateness, and dataset classification. Feedback from this manual re-

view—particularly examples of agent failures or subtle misclassifications—is periodically collected and incorporated as few-shot exemplars to improve the segmentation and generation agents in subsequent iterations. This feedback loop enhances robustness and precision over time without interrupting large-scale automated processing.

This pipeline enables fully automated conversion of scholarly materials into standardized evaluation data. By combining LLM-based reasoning, dynamic regular-expression segmentation, and human-in-the-loop verification, StatEval achieves both scalability and reliability.

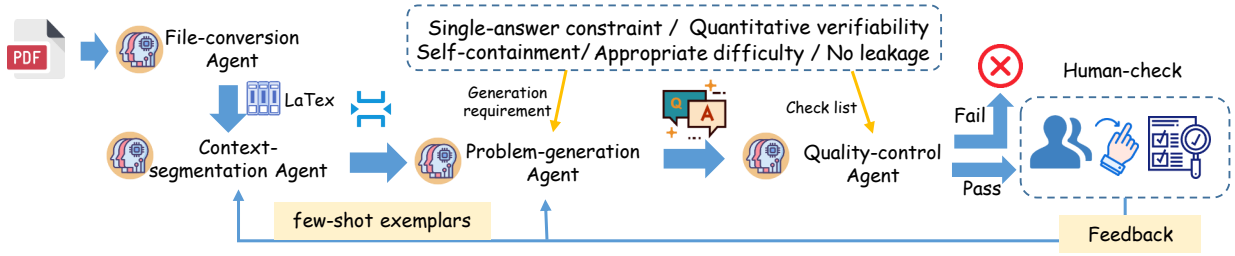



Figure 4: Overview of the StatEval data processing pipeline. Each agent corresponds to a major functional stage in the automated extraction and verification process.

# 4 Scoring Framework

To ensure rigorous and interpretable evaluation, StatEval defines different scoring schemes for Multiple-Choice Questions and Open-Ended QA Questions respectively.

## 4.1 Scoring for Multiple-Choice Questions

Multiple-choice questions are graded by exact answer matching. A model receives a score of 1 if its selected option matches the correct answer and 0 otherwise. No partial credit is

given, ensuring a clear and objective standard consistent with common academic testing practice.

## 4.2 Scoring for Open-Ended QA Questions

Open-ended questions, including those from both the Foundational Knowledge dataset and the Statistical Research dataset, are evaluated through a process-based scoring pipeline designed to assess both final correctness and the quality of intermediate reasoning. The pipeline consists of four sequential components:

1. **Reasoning Step Extraction:** The model's response is parsed to identify key reasoning steps, including assumptions, logical transitions, and intermediate derivations. This stage reconstructs the complete reasoning chain to capture how the final result is obtained.

2. **Outcome Extraction:** Each reasoning step is further analyzed to extract its quantitative or symbolic outcome (e.g., computed values, derived expressions, or identified distributions). This ensures that both the logical structure and the resulting intermediate outcomes are available for later verification.

3. **LLM Judging:** A dedicated LLM evaluator compares the extracted reasoning steps and outcomes with the reference solution. It verifies correctness, checks whether each step is necessary and sufficient, and detects any logical inconsistency or missing justification.

4. **Scoring:** Based on the LLM Judge's evaluation, each step is assigned binary scores along three dimensions: (i) *Reasoning Accuracy*, (ii) *Step Completeness*, and (iii)

*Final Answer Correctness.* The aggregated score for one evaluation pass is computed as

$$S_{\text{final}}^{(i)} = \alpha S_r^{(i)} + \beta S_s^{(i)} + (1 - \alpha - \beta) S_a^{(i)},$$

where $S_r^{(i)}, S_s^{(i)}, S_a^{(i)} \in \{0, 1\}$ and $\alpha = 0.4, \beta = 0.3$. To get conservative results, scoring is repeated three times with different random seeds, and the final score is defined as

$$S_{\text{final}} = \min\{S_{\text{final}}^{(1)}, S_{\text{final}}^{(2)}, S_{\text{final}}^{(3)}\}.$$

This four-step design separates reasoning reconstruction from correctness judgment, enabling fine-grained and interpretable evaluation. The framework outputs two complementary indicators: (1) a *final score* reflecting overall correctness, and (2) a *process score* reflecting reasoning quality and stepwise consistency. An illustration of this pipeline is shown in Figure 5.

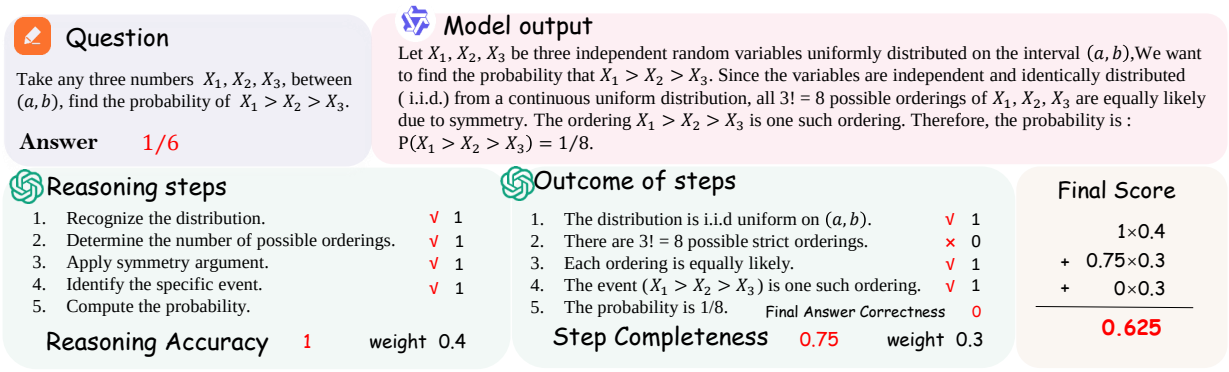



Figure 5: Examples and scoring procedures in StatEval

# 5 Experiments

In this section, we sequentially present the evaluation details and results for each dataset. To accommodate computational and API cost constraints, we construct a compact evalua-

tion subset, termed **StatEval-mini**, by sampling from both the Foundational Knowledge and Statistical Research datasets.

From the foundational dataset, which contains 13,817 problems, we select 1,300 representative questions stratified by subject, difficulty, question type, and problem length, ensuring comprehensive coverage across undergraduate and graduate levels. Similarly, from the statistical research dataset, we select 2,000 questions following the same stratification principles. In this manner, **StatEval-mini** serves as a representative, balanced, and computationally efficient subset of the full StatEval benchmark, enabling cost-effective yet reliable evaluation.

**Evaluation Protocol** All evaluations follow the scoring framework introduced in Section 4. For each model, the final leaderboard score is computed as the total points obtained divided by the total possible points across all questions. Each question carries a maximum score of 1.0, but partial credit (e.g., 0.5 or 0.8) may be awarded for open-ended problems according to the process-based scoring pipeline.

We adopt OpenAI's latest model, **GPT-5**, as the judging model responsible for automatic scoring under this protocol. The evaluated models include a diverse set of open-source and closed-source systems covering a wide range of parameter scales and architectures. The open-source series comprises **LLaMA-3.1**, **GPT-OSS** (20B and 120B), **DeepSeek-V3.1**, and **Qwen** models (30B and 235B). The closed-source series includes **GPT-5** variants and **Gemini-2.5** models. Together, these models provide a comprehensive landscape for evaluating statistical reasoning capabilities.

## 5.1 Evaluation Results

**Results for the foundational knowledge dataset.** Table 1 summarizes the performance of various large language models at both undergraduate and graduate levels. Overall, closed-source models consistently outperform open-source models across all subject areas and in overall mean scores, with GPT-5 achieving the highest overall mean of 82.85, demonstrating the strongest comprehensive capability to date.

Table 1: Foundational knowledge dataset results by academic level and domain

| Model | Graduate | | | | Undergraduate | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|
| | Prob. | Stat. | ML | Mean | Prob. | Stat. | ML | Mean | Mean |
| **Open-source Models** | | | | | | | | | |
| DeepSeek-V3.1 | 51.09 | 43.27 | 44.49 | 45.13 | 80.27 | 64.57 | 47.78 | 62.88 | 53.98 |
| Qwen2.5-72B | 68.86 | 64.58 | 62.51 | 64.62 | 78.49 | 75.98 | 72.57 | 75.68 | 70.42 |
| Qwen3-30B | 71.46 | 71.07 | 61.49 | 67.46 | 73.76 | 74.27 | 77.37 | 75.09 | 71.49 |
| Qwen3-235B | 73.43 | 76.92 | 68.46 | 73.13 | 84.29 | 77.55 | 80.56 | 80.57 | 76.96 |
| LLaMA-3.1-8B | 47.98 | 40.71 | 34.74 | 39.98 | 48.94 | 44.18 | 43.12 | 45.30 | 42.79 |
| GPT-OSS-120B | 74.42 | 75.43 | 73.31 | 74.46 | 88.76 | 84.91 | 83.32 | 85.57 | 80.27 |
| GPT-OSS-20B | 68.19 | 72.19 | 67.48 | 69.69 | 85.53 | 81.76 | 77.10 | 81.40 | 75.77 |
| **Closed-source Models** | | | | | | | | | |
| GPT-5 | 78.94 | 82.31 | 71.28 | 78.72 | 88.23 | 87.46 | 85.90 | 87.24 | 82.85 |
| GPT-5 mini | 78.66 | 78.14 | 71.61 | 76.50 | 86.84 | 81.63 | 86.17 | 84.69 | 80.37 |
| Gemini-2.5 Pro | 72.15 | 79.12 | 68.78 | 75.43 | 88.53 | 86.84 | 85.24 | 86.90 | 80.88 |
| Gemini-2.5 Flash | 71.27 | 75.25 | 69.27 | 73.11 | 86.73 | 80.23 | 78.53 | 81.67 | 77.23 |

Among open-source models, Qwen3-235B performs notably well, reaching an overall mean of 76.96 and narrowing the gap with some closed-source models. Other open-source models, including LLaMA-3.1-8B and DeepSeek-V3.1, exhibit lower performance across both undergraduate and graduate tasks, highlighting that model scale and training optimization remain important factors affecting performance on foundational statistical tasks.

**Results for the statistical research dataset.** Table 2 presents model performance across subfields and result categories, reflecting their reasoning capabilities in frontier statistical tasks. A clear distinction emerges between closed-source and open-source models, with closed-source systems, particularly the GPT-5 family, consistently outperforming open-source alternatives. The GPT5-mini series and its effort-tuned variants achieve the highest overall scores, exceeding 60% on average, while Gemini models demonstrate moderate but balanced performance across statistical subfields. Open-source Qwen models, though trailing in overall averages, show competitive potential in specific areas, especially in probability.

Field-level results indicate the strongest reasoning performance in *Probability* and *Statistics*, where top models surpass 70%, whereas *Machine Learning* tasks remain more challenging, with even the best GPT models scoring lower. Analysis by result category reveals that GPT-5 models excel in *Identifiability & Consistency* and *Testing Validity*, highlighting their strength in rigorous statistical reasoning and hypothesis evaluation. Gemini models show relative advantages in *Distributional Properties* and *Structural Guarantees*, while Qwen models improve gradually in *Probability* and *Distributional Properties* as model scale increases.

Overall, these results illustrate a hierarchy of reasoning ability in frontier statistical tasks: GPT5-mini and its tuned variants define the strongest benchmarks, Gemini models

occupy an intermediate position, and Qwen models represent a developing open-source alternative with emerging strengths. This emphasizes that both model scale and targeted optimization are crucial for advancing statistical reasoning capabilities in research-level contexts.

Table 2: Statistical research dataset results by domain and property type

| Model | Subfields | | | Result Categories | | | | | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ML | Prob | Stat | Asymp | Conv | Dist | Gen | Ident | Opt | Struct | Test | Mean |
| *Closed-source models* | | | | | | | | | | | | |
| Gem2.5-flash | 44.10 | 58.65 | 53.26 | 53.38 | 49.24 | 56.04 | 21.23 | 67.35 | 53.45 | 60.47 | 49.30 | 51.14 |
| Gem2.5-flashlite | 36.03 | 51.50 | 43.11 | 40.82 | 40.15 | 46.31 | 17.12 | 56.36 | 43.53 | 50.00 | 46.51 | 41.58 |
| GPT5-mini | 48.56 | 66.54 | 59.46 | 62.20 | 54.55 | 63.42 | 25.00 | 71.48 | 52.16 | 62.79 | 63.26 | 57.62 |
| GPT5-nano | 42.66 | 53.76 | 47.61 | 48.91 | 46.21 | 48.99 | 20.89 | 61.51 | 42.24 | 54.65 | 55.81 | 47.05 |
| *Open-source models* | | | | | | | | | | | | |
| GPT-oss-120B | 41.28 | 56.39 | 47.18 | 53.38 | 42.42 | 50.67 | 18.49 | 60.82 | 45.69 | 65.12 | 61.86 | 49.49 |
| GPT-oss-20B | 34.26 | 48.87 | 34.26 | 44.44 | 37.12 | 43.29 | 18.15 | 55.33 | 39.66 | 52.33 | 48.84 | 42.21 |
| Qwen3-235B | 43.29 | 62.03 | 52.08 | 53.86 | 50.76 | 59.06 | 20.55 | 67.70 | 48.28 | 55.81 | 49.77 | 51.10 |
| Qwen3-30B | 41.20 | 59.77 | 44.93 | 49.28 | 49.24 | 53.36 | 18.15 | 61.17 | 41.38 | 60.47 | 53.49 | 47.43 |

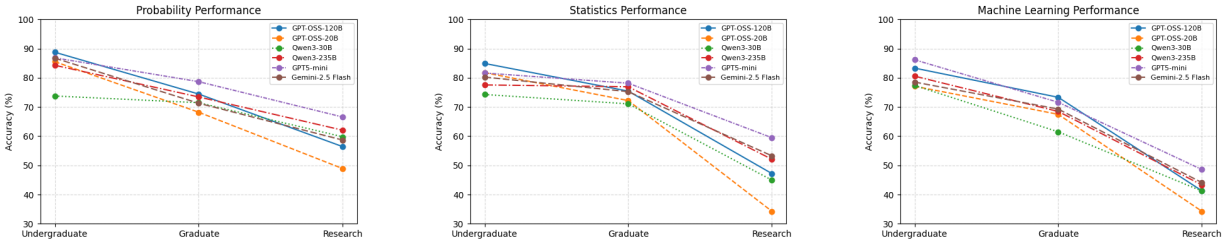
**Note:** ML = Machine Learning; Prob = Probability; Stat = Statistics; Asymp = Asymptotic Properties; Conv = Convergence & Stability; Dist = Distributional Properties; Gen = Generalization & Error Bounds; Ident = Identifiability & Consistency; Opt = Optimality Results; Struct = Structural Guarantees; Test = Testing Validity.

**Performance by Subject Area** . Across both foundational and research-level tasks, model performance exhibits consistent patterns among the three core statistical subjects: Probability, Statistics, and Machine Learning. For undergraduate foundational tasks, top-performing closed-source models (e.g., GPT-5 mini-H) achieve 72.9% in Probability, 63.1% in Statistics, and 51.4% in Machine Learning. At the graduate level, these models show slight improvements in Probability (73–74%) and Statistics (64–65%), while performance in Machine Learning remains lower (52–58%).

In research-level tasks, the pattern persists: closed-source models maintain relatively strong and balanced performance in Probability (66–73%) and Statistics (59–65%), whereas

Machine Learning remains the most challenging domain, with lower scores particularly for open-source models (e.g., Qwen3-30B achieves 44.9% in Statistics but only 41.2% in Machine Learning). Across both datasets, closed-source models consistently outperform open-source models by 8–15 percentage points.

These results indicate that while foundational and frontier statistical reasoning is captured robustly, Machine Learning tasks, especially at research level, pose greater difficulty for current LLMs. Figure 6 visualizes these trends, highlighting the relative robustness of Probability and Statistics compared to the decline observed in Machine Learning performance.




(a) Probability (b) Statistics (c) Machine Learning

Figure 6: Performance of models across the three core subfields: Probability, Statistics, and Machine Learning. Each bar represents the mean score of the respective model in the corresponding subfield, highlighting differences between closed- and open-source models and the effect of effort-tuned variants.

**Research Capability by Result Category.** In research-level tasks derived from the statistical research dataset, model performance exhibits notable variation across reasoning categories. GPT-5 series models, including their high- and medium-effort variants, demonstrate outstanding performance in Identifiability & Consistency (74–77%) and Testing Validity (64–72%), reflecting strong capabilities in rigorous statistical reasoning and

theoretical verification.

The Gemini series achieves comparatively higher scores in Distributional Properties (up to 59%) and Structural Guarantees (up to 60%), though performance in more complex reasoning categories, such as Optimality Results and Generalization & Error Bounds, is moderately lower (42–50%). Open-source models, including Qwen3-235B and Qwen3-30B, show solid performance in probability-related reasoning (50–62%) but lag in optimization- and generation-related reasoning (39–48% and 16–21%, respectively), highlighting the potential for targeted fine-tuning to enhance theoretical derivation skills.

Figure 7 illustrates these contrasts, emphasizing the advantage of closed-source and effort-tuned models in research-level statistical reasoning across diverse categories.

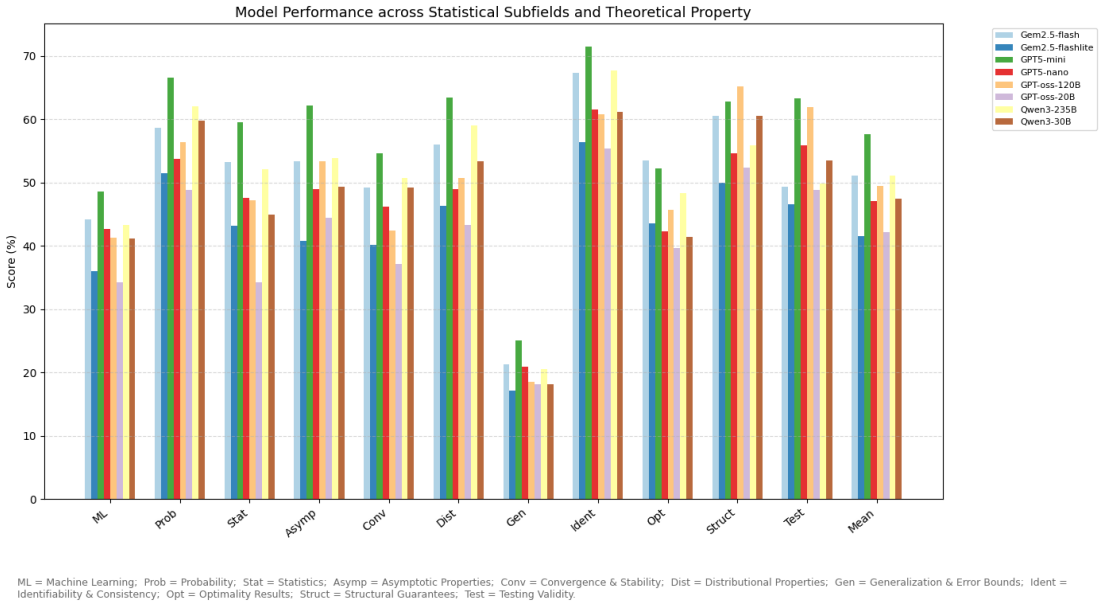



Figure 7: Model performance across all statistical reasoning result categories. Closed-source models, open-source models are shown for comparison.

# 6 Related work

With the rapid advancement of large language models (LLMs), systematic reasoning has become a key strength for solving complex problems in mathematics and science. Early progress was driven by prompting techniques such as "chain of thought" (Wei et al., 2022) and "tree of thought" (Yao et al., 2023), which encouraged stepwise reasoning. More recently, direct training approaches, including reward modeling (Uesato et al., 2022) and frameworks like Reinforcement Learning with Verifiable Reward (RLVR) (Guo et al., 2025; Dai et al., 2025; Zheng et al., 2025), have further improved multi-step reasoning, resulting in notable gains in mathematical reasoning (Guo et al., 2025), code generation (Wang et al., 2025), multimodal reasoning (Liu et al., 2025), and information extraction (Dai et al., 2025). A variety of benchmarks have been developed to measure these advances. However, LLMs' abilities in rigorous statistical problem solving remain underexplored, primarily due to the lack of dedicated benchmarks capturing the complexity of statistical inference and theoretical reasoning.

To contextualize current progress in mathematical and applied statistical evaluation benchmarks, we categorize existing datasets as follows:

1. **Mathematical reasoning benchmarks** primarily assess logical and quantitative skills, but rarely emphasize statistics. The multidisciplinary **MMLU** dataset (Hendrycks et al., 2020) includes only high-school level statistics, focusing on basic concepts. **MATH** (Hendrycks et al., 2021) contains a limited set of probability and counting problems, mainly from competition-style exercises. **MathBench** (Liu et al., 2024) incorporates some undergraduate-level statistics, yet topics remain relatively simple. Competition-focused datasets such as **Omni-MATH** (Gao et al., 2024) and **OlympiadBench** (He et al., 2024) emphasize probability problems, while **UG-**

**MathBench** (Xu et al., 2025) draws on university grading exercises but lacks highly challenging or comprehensive statistical tasks. Classical benchmarks like **GSM8K** (Cobbe et al., 2021) and **MATHVERSE** (Zhang et al., 2024) largely exclude statistics.

2. **Research-level benchmarks** aim to capture reasoning in cutting-edge academic work. **TheoremQA** (Chen et al., 2023) collects university-level theorem application problems across mathematics, physics, finance, and computer science to evaluate systematic theorem application. **RealMath** (Zhang et al., 2025) converts theorem proofs from arXiv papers into structured QA pairs for evaluation, though it mostly omits statistics. In the statistics domain, **PaperBench** (Starace et al., 2025) assembles evaluation questions from ICML papers in collaboration with authors, but its workflow is highly manual and coverage of theoretical statistics is limited. Other research-oriented benchmarks, such as **SciAssess** (Cai et al., 2024) and **SciFIBench** (Roberts et al., 2024), emphasize multimodal understanding (e.g., figures, document hierarchies) rather than formal theorem proving.

3. **Data analysis benchmarks** focus primarily on practical tasks such as coding, workflow execution, and quantitative reasoning, rather than theoretical inference. **StatQA** (Zhu et al., 2024) evaluates column identification, method selection, and hypothesis testing; **QR-Data** (Liu et al., 2024) examines quantitative reasoning on tabular datasets. Other benchmarks, including **MLAgentBench** (Huang et al., 2023), **DSBench** (Jing et al., 2024), **DSCodeBench** (Ouyang et al., 2025), and **StatLLM** (Song et al., 2025), assess programming and application of statistical software, but do not target systematic reasoning or formal proof ability.

Despite the availability of these benchmarks, theoretical statistical inference remains largely unaddressed, and existing datasets focus on either applications or basic exercises.

**StatEval** fills this gap as the first comprehensive benchmark for formal statistical reasoning, spanning both foundational and research-level problems across the discipline.

**Automated Evaluation Paradigms.** To reduce manual evaluation costs, the **"LLM-as-a-judge"** paradigm (Ashktorab et al., 2025) is widely adopted, but its common binary (correct/incorrect) scoring is unsuitable for complex statistical tasks, reducing evaluation accuracy and stability (Shi et al., 2024). Strategies such as pairwise comparisons improve accuracy, but increase computational overhead (Jiang et al., 2023; Xu et al., 2025); Product-of-Experts (PoE) reduces pairwise comparisons but still lacks fine-grained evaluation and relies on LLMs' black-box judgment (Liusie et al., 2024). Meanwhile, **formal proof assistants** like Lean 4(Moura and Ullrich, 2021), which enable rigorous verification in mathematics (e.g., FormalMath (Yu et al., 2025), PutnamBench (Tsoukalas et al., 2024)), face fundamental challenges in statistics, as statistical proofs involve random variables, asymptotic arguments, and multiple solution paths that are difficult to formalize.

In summary, despite progress in reasoning benchmarks and evaluation paradigms for LLMs, rigorous assessment of advanced statistical theory remains largely unaddressed. To fill this gap, we introduce **StatEval**, a benchmark and evaluation framework that systematically measures both complex problem-solving and proof-based reasoning in statistics, enabling robust and fine-grained analysis of LLM capabilities.

# 7 Conclusion

While LLM evaluations have largely focused on logic and mathematics, statistical reasoning remains underexplored. To fill this gap, we introduce **StatEval**, the first comprehensive benchmark covering both foundational (13,000+ undergraduate-level) and research-level (2,000+ literature-based) statistical problems across varying difficulty and interdisciplinary

applications. Our systematic evaluation reveals that even the strongest closed-source models struggle with research-level tasks, particularly in advanced machine learning theory. These results underscore the need and potential for enhancing LLMs' statistical reasoning, providing a benchmark and reference for future development of research-oriented statistical AI tools.

# References


Ashktorab, Z., E. M. Daly, E. Miehling, W. Geyer, M. S. Cooper, T. Pedapati, M. Desmond, Q. Pan, and H. J. Do (2025). Evalassist: A human-centered tool for llm-as-a-judge. *arXiv preprint arXiv:2507.02186*.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems 33*, 1877–1901.

Cai, H., X. Cai, J. Chang, S. Li, L. Yao, C. Wang, Z. Gao, H. Wang, Y. Li, M. Lin, et al. (2024). Sciassess: Benchmarking llm proficiency in scientific literature analysis. *arXiv preprint arXiv:2403.01976*.

Chen, W., M. Yin, M. Ku, P. Lu, Y. Wan, X. Ma, J. Xu, X. Wang, and T. Xia (2023). Theoremqa: A theorem-driven question answering dataset. *arXiv preprint arXiv:2305.12524*.

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

Dai, R., L. Song, H. Liu, Z. Liang, D. Yu, H. Mi, Z. Tu, R. Liu, T. Zheng, H. Zhu, et al.

(2025). Cde: Curiosity-driven exploration for efficient reinforcement learning in large language models. *arXiv preprint arXiv:2509.09675*.

Dai, R., T. Zheng, R. Yang, K. Yu, and H. Zhu (2025). R1-re: Cross-domain relation extraction with rlvr. *arXiv preprint arXiv:2507.04642*.

Gao, B., F. Song, Z. Yang, Z. Cai, Y. Miao, Q. Dong, L. Li, C. Ma, L. Chen, R. Xu, et al. (2024). Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*.

Gao, C., J. Gu, and T. Ma (2023). Scientific discovery in the age of artificial intelligence. *Nature Reviews Physics 5*, 646–662.

Guo, D., D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

He, C., R. Luo, Y. Bai, S. Hu, Z. L. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang, et al. (2024). Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*.

Hendrycks, D., C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt (2020). Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Hendrycks, D., C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt (2021). Measuring mathematical problem solving with the math dataset. In *Advances in Neural Information Processing Systems*, Volume 34, pp. 3512–3524.

Huang, Q., J. Vora, P. Liang, and J. Leskovec (2023). Benchmarking large language models as ai research agents. In *NeurIPS 2023 Foundation Models for Decision Making Workshop*.

Jiang, D., X. Ren, and B. Y. Lin (2023). Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*.

Jing, L., Z. Huang, X. Wang, W. Yao, W. Yu, K. Ma, H. Zhang, X. Du, and D. Yu (2024). Dsbench: How far are data science agents from becoming data science experts? *arXiv preprint arXiv:2409.07703*.

Khan, S. et al. (2023). Large language models in education: Opportunities and challenges. In *Proceedings of the Learning at Scale Conference*.

Liu, H., Z. Zheng, Y. Qiao, H. Duan, Z. Fei, F. Zhou, W. Zhang, S. Zhang, D. Lin, and K. Chen (2024). Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. *arXiv preprint arXiv:2405.12209*.

Liu, R., D. Yu, T. Zheng, R. Dai, Z. Li, W. Yu, Z. Liang, L. Song, H. Mi, P. Tokekar, and D. Yu (2025). Vogue: Guiding exploration with visual uncertainty improves multimodal reasoning. *arXiv preprint arXiv:2510.01444*.

Liu, X., Z. Wu, X. Wu, P. Lu, K.-W. Chang, and Y. Feng (2024). Are llms capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. *arXiv preprint arXiv:2402.17644*.

Liusie, A., V. Raina, Y. Fathullah, and M. Gales (2024). Efficient llm comparative assessment: a product of experts framework for pairwise comparisons. *arXiv preprint arXiv:2405.05894*.

Moura, L. d. and S. Ullrich (2021). The lean 4 theorem prover and programming language. In *International Conference on Automated Deduction*, pp. 625–635. Springer.

Ouyang, S., D. Huang, J. Guo, Z. Sun, Q. Zhu, and J. M. Zhang (2025). Dscodebench: A realistic benchmark for data science code generation. *arXiv preprint arXiv:2505.15621*.

Paster, K., M. Dos Santos, Z. Azerbayev, and J. Ba (2025). Openwebmath: An open dataset of high-quality mathematical web text (2023). *URL https://arxiv. org/abs/2310 6786*.

Polu, S. and I. Sutskever (2020). Generative language modeling for automated theorem proving. *arXiv preprint arXiv:2009.03393*.

Roberts, J., K. Han, N. Houlsby, and S. Albanie (2024). Scifibench: Benchmarking large multimodal models for scientific figure interpretation. *Advances in Neural Information Processing Systems 37*, 18695–18728.

Shi, L., C. Ma, W. Liang, X. Diao, W. Ma, and S. Vosoughi (2024). Judging the judges: A systematic study of position bias in llm-as-a-judge. *arXiv preprint arXiv:2406.07791*.

Song, X., L. Lee, K. Xie, X. Liu, X. Deng, and Y. Hong (2025). Statllm: A dataset for evaluating the performance of large language models in statistical analysis. *arXiv preprint arXiv:2502.17657*.

Starace, G., O. Jaffe, D. Sherburn, J. Aung, J. S. Chan, L. Maksin, R. Dias, E. Mays, B. Kinsella, W. Thompson, et al. (2025). Paperbench: Evaluating ai's ability to replicate ai research. *arXiv preprint arXiv:2504.01848*.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tsoukalas, G., J. Lee, J. Jennings, J. Xin, M. Ding, M. Jennings, A. Thakur, and S. Chaudhuri (2024). Putnambench: Evaluating neural theorem-provers on the putnam mathematical competition. *Advances in Neural Information Processing Systems 37*, 11545–11569.

Uesato, J., N. Kushman, R. Kumar, F. Song, N. Siegel, L. Wang, A. Creswell, G. Irving, and I. Higgins (2022). Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Wang, B., C. Xu, X. Zhao, L. Ouyang, F. Wu, Z. Zhao, R. Xu, K. Liu, Y. Qu, F. Shang, et al. (2024). Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.

Wang, H., L. Li, C. Qu, F. Zhu, W. Xu, W. Chu, and F. Lin (2025). To code or not to code? adaptive tool integration for math language models via expectation-maximization. *arXiv preprint arXiv:2502.00691*.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems 35*, 24824–24837.

Wu, Y., S. Polu, and I. Sutskever (2021). Autoformalization with large language models. *arXiv preprint arXiv:2106.01344*.

Xu, X., J. Zhang, T. Chen, Z. Chao, J. Hu, and C. Yang (2025). Ugmathbench: A diverse and dynamic benchmark for undergraduate-level mathematical reasoning with large language models. *arXiv preprint arXiv:2501.13766*.

Xu, Y., L. Ruis, T. Rocktäschel, and R. Kirk (2025). Investigating non-transitivity in llm-as-a-judge. *arXiv preprint arXiv:2502.14074*.

Yao, S., D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan (2023). Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems 36*, 11809–11822.

Yu, Q., Z. Zhang, R. Zhu, Y. Yuan, X. Zuo, Y. Yue, W. Dai, T. Fan, G. Liu, L. Liu, et al. (2025). Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Yu, Z., R. Peng, K. Ding, Y. Li, Z. Peng, M. Liu, Y. Zhang, Z. Yuan, H. Xin, W. Huang, et al. (2025). Formalmath: Benchmarking formal mathematical reasoning of large language models. *arXiv preprint arXiv:2505.02735*.

Zhang, J., C. Petrui, K. Nikolić, and F. Tramèr (2025). Realmath: A continuous benchmark for evaluating language models on research-level mathematics. *arXiv preprint arXiv:2505.12575*.

Zhang, R., D. Jiang, Y. Zhang, H. Lin, Z. Guo, P. Qiu, A. Zhou, P. Lu, K.-W. Chang, Y. Qiao, et al. (2024). Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer.

Zheng, T., H. Zhang, W. Yu, X. Wang, X. Yang, R. Dai, R. Liu, H. Bao, C. Huang, H. Huang, et al. (2025). Parallel-r1: Towards parallel thinking via reinforcement learning. *arXiv preprint arXiv:2509.07980*.

Zhu, Y., S. Du, B. Li, Y. Luo, and N. Tang (2024). Are large language models good statisticians? *Advances in Neural Information Processing Systems 37*, 62697–62731.

# A Journal sources

In this section, we provide a detailed description of the specific journal sources used in the Statistical Research dataset. The dataset comprises 2,374 proof-based research tasks derived from 2,719 research articles published between 2020 and 2025 across 18 leading journals. This dataset emphasizes high-quality theoretical contributions, including both core statistical theory and interdisciplinary developments in probability, econometrics, and machine learning. For a complete list of the journals included, please refer to Table 3.

Table 3: Research Data Sources by Discipline

| **Discipline** | **Journal Name** |
|---|---|
| **Statistics** | Annals of Statistics, Biometrika, Journal of the American Statistical Association |
| | Journal of the Royal Statistical Society Series B, Bernoulli, Biometrics |
| | Statistica Sinica, Journal of Computational and Graphical Statistics |
| **Econometrics** | Journal of Econometrics, Journal of Business & Economic Statistics, Econometrica |
| **Machine Learning** | Journal of Machine Learning Research, Machine Learning |
| | IEEE Transactions on Pattern Analysis and Machine Intelligence |
| | IEEE Transactions on Information Theory, Journal of Artificial Intelligence |
| **Probability Theory** | The Annals of Probability, Probability Theory and Related Fields |

Our selection includes top-tier peer-reviewed journals in statistics, such as *Annals of Statistics*, *Biometrika*, *Journal of the American Statistical Association*, and *Journal of the Royal Statistical Society Series B*, alongside journals from related fields, including econometrics, probability theory, and theoretical machine learning. This combination ensures that the benchmark captures a diverse set of formal theorem-proof reasoning tasks spanning multiple research domains.

To facilitate data acquisition and maintain copyright compliance, corresponding *arXiv* versions were retrieved for all articles where available. This approach ensures open access

to the content and allows conversion of theoretical results into structured tasks without relying on supplementary arXiv-only papers.

The journal selection was guided by several considerations. First, inclusion of top-tier, peer-reviewed journals guarantees that foundational and methodological contributions are well represented. Second, journals from adjacent fields capture the interdisciplinary nature of modern statistics, reflecting the integration of probability, econometrics, and machine learning theory. Third, application-focused journals, such as *Annals of Applied Statistics*, were deliberately excluded, as the benchmark emphasizes formal theoretical reasoning rather than applied case studies. After processing, the final dataset includes 2,374 extractable theorem-proof problems suitable for benchmarking.

# B Research Category

Frontier statistical literature can be classified from two perspectives: first, by research topics or subject areas, covering directions from classical statistical modeling to deep learning and causal inference; second, by the type of theoretical results, which systematically summarize the properties and performance of statistical methods. We provide a detailed description of these two perspectives below.

## B.1 Subject Area Classification

**Classical Statistical Modeling and Inference.** This area primarily focuses on traditional statistical methods and their theoretical properties, including linear and generalized linear models, quantile regression, kernel density estimation, smoothing and spline methods, Fréchet regression, distribution embeddings, and shape/geometric statistics. It also covers heteroscedasticity and robust estimation, nonparametric change-point detection, multiple

testing and false discovery rate control, selective and sequential testing, bootstrap and resampling methods, high-dimensional optimization, and approximate algorithms. These methods form the foundational tools of statistical science, and understanding them is essential for both methodological development and applied analysis.

**High-dimensional Data Modeling.** This area studies statistical modeling under high-dimensional or complex data structures, including high-dimensional regression (Lasso, Ridge, sparse regression), covariance and precision matrix estimation, high-dimensional principal component analysis, independence and correlation testing, tensor regression, multi-response/multivariate analysis, functional data analysis and functional time series, as well as semi-parametric survival models and individualized risk estimation. The high-dimensional setting arises in many modern applications, and methods in this category are critical to ensure reliable inference when the number of variables is large relative to the sample size.

**Bayesian and Generative Models.** This area includes Markov Chain Monte Carlo methods (Gibbs sampling, HMC), variational inference, Bayesian nonparametrics (Dirichlet Process, Gaussian Process), hierarchical models and Bayesian factor analysis, approximate Bayesian computation, Wasserstein distance and optimal transport, as well as deep generative models and score-based flow models. Bayesian methods provide a coherent probabilistic framework for incorporating prior information and handling uncertainty, making this category indispensable for both theory and practice.

**Probability Theory and Stochastic Processes.** Focuses on random structures and their analysis, including random graphs and networks, stochastic walks and mixing, extreme value theory, stochastic PDEs, random matrices and combinatorial probability, geometric probability, high-dimensional and classical time series, change-point detection, covariance and autocorrelation estimation, and Markov and hidden Markov models. Probability theory

forms the mathematical backbone of statistics, and a deep understanding of stochastic processes is critical for deriving rigorous inference results and modeling complex random phenomena.

**Graphs and Structured Data.** Covers graph/network learning, community detection, higher-order graph structures, graphical models, tensor decomposition and multiway representation, graph regression and high-dimensional graph modeling, network sampling and inference, as well as keyword and citation network analysis. Structured data arise ubiquitously in applications such as social networks, biological networks, and bibliometrics, making this area essential for modern statistical analysis.

**Deep Learning and Reinforcement Learning.** Includes policy gradient and actor-critic methods, distributed reinforcement learning, RLHF and bilevel RL, kernel-based RL, risk optimization, sample complexity analysis, LLM/statistical analysis, deep neural network regression, causal estimation, transfer learning and robustness, and neural network-based splines. These methods are at the forefront of data-driven modeling, bridging statistics and AI, and they provide powerful tools for prediction, optimization, and causal reasoning in complex environments.

**Causal Inference and Experimental Design.** Covers individualized treatment rules, propensity score matching, robust estimation, latent/mediation causal effects, experimental and observational design, offline RL causal policy estimation, functional/continuous treatment effects, hierarchical effects, quantile treatment effects, negative controls, and causal structure learning. Causal inference is central for understanding mechanisms and making reliable decisions, making this category indispensable for both applied and methodological research.

**Other Emerging Directions.** Includes stochastic/geometric object statistics, model validation and structural deviation analysis, power-law and tail analysis, robust inference,

shape/evolutionary statistics, predictive risk quantification, and physics-informed machine learning and simulation methods. These emerging areas capture new challenges in modern data science and provide avenues for innovative methods that cannot be addressed by classical approaches alone.

## B.2 Theoretical Property Classification

From the perspective of theoretical property, frontier statistical literature often focuses on performance guarantees, feasibility, and uncertainty quantification. The main subcategories include:

**Optimality Results.** Studies the optimal properties of methods or estimators, such as minimax optimality, Bayesian optimality, efficiency/UMVUE, and rate optimality. Understanding optimality is crucial to benchmark methods and guide the development of statistically efficient procedures. This category also encompasses computational feasibility, including algorithmic complexity, polynomial-time solvability, hardness/lower bounds, and statistical-computational trade-offs. Considering computational aspects ensures that theoretically optimal methods are practically implementable on large-scale data.

**Asymptotic Properties.** Covers asymptotic distribution, asymptotic efficiency, higher-order asymptotic expansions, high-dimensional asymptotics, and related properties. These results provide insights into the long-sample behavior of estimators and test statistics, helping to justify approximations, design confidence intervals, and inform inference procedures in complex or high-dimensional settings.

**Testing Validity.** Focuses on controlling type I error, power properties, asymptotic null distributions, and sequential/adaptive testing. Valid hypothesis testing ensures scientific conclusions are statistically justified, making this category essential for evaluating claims in experimental and observational studies. Results here guide the construction of

tests that are both reliable and robust under various modeling assumptions.

**Convergence and Stability.** Concerns algorithmic convergence, stochastic process convergence, stability analysis, and non-asymptotic convergence results. Stability and convergence analysis guarantees that iterative procedures, learning algorithms, and stochastic optimization methods are well-behaved, reproducible, and numerically robust, which is crucial for both theory and practical applications.

**Generalization and Error Bounds.** Includes learning bounds, regularization bounds, model selection risk, non-asymptotic error bounds, and information-theoretic guarantees such as minimax lower bounds, sample complexity limits, and recoverability thresholds. These results are fundamental to understanding how well methods perform on unseen data, under model misspecification, or when data is limited, thereby providing guidance for model selection, algorithm design, and assessment of statistical reliability.

**Identifiability and Consistency.** Studies parameter identifiability, causal identifiability, and consistency of estimators. Without identifiability, estimates may be non-unique or meaningless, and without consistency, estimates may not converge to the true parameters as sample size increases. This category ensures that inference procedures are theoretically sound and interpretable.

**Distributional Properties.** Covers central limit theorems, extreme value/tail properties, concentration inequalities, random matrix theory, dependence structures, and risk/uncertainty quantification including confidence intervals, posterior variability, and robust risk bounds. These properties provide a rigorous understanding of variability and dependence, underpin uncertainty assessment, and inform decisions in settings where data may be noisy, dependent, or high-dimensional.

**Structural Guarantees.** Focuses on recoverability of graphical models, manifold/-geometric guarantees, causal/SEM identifiability, sparsity/low-rank recovery, and related

structural properties. Structural guarantees ensure that statistical models can capture essential patterns in the data, that recovery is feasible under realistic assumptions, and that the resulting models are interpretable and reproducible in practice.

# C Prompt

**Prompt for Foundational knowledge dataset**

You are a helpful assistant designed to help with statistical problems. When given a statistical question, you should answer the question according to the user's prompt. For calculation questions in statistics, follow this format: Here is a calculation problem in statistics. Please write out the calculation process as well as the final answer.

——@Problem: question

——@Calculation process:

——@Final answer:

**Prompt for Extracting Key Steps**

You are now a statistics teacher, and students have provided some answers to a calculation problem in statistics. Your task is to extract the key calculation steps and the final calculation result (the last step) from the students' answers and return the extracted content in markdown format. The goal is to extract key information, not expand the answer, so you need to return to the key steps, including brief textual explanations and formula calculations, without adding your own understanding or comments. There is no need to explain why the information is extracted this way. Now, you need to extract from the following answer:

——@Student's answer:{response}

——@Extracted answer:

**Prompt for Foundational knowledge dataset Evaluation (Part 1)**

You are a helpful grading assistant designed to help with statistical problems. When given a response and a ground truth solution, you should score the response according to the user's grading criteria.

You are now a grading teacher for a statistics exam. I will provide you with a statistics problem, including the question, reference answer, and the key steps output by the model. Based on the model's solution steps, please give the average score for the thought process, the step-by-step calculation, and the final answer score. The specific requirements are as follows:

——Thought process average score: Based on the reference answer and the model's solution steps, evaluate whether the thought process behind each step of the model's solution is correct. The thought process can differ from the reference answer as long as it is correct. It is not necessary to calculate the result correctly, as long as the calculation idea is correct. For each correct step, give 1 point, and 0 points for incorrect ones. Then calculate the average score for the thought process.

——Step average score: Based on the reference answer and the model's solution steps, evaluate whether the calculation steps are correct. Unlike the thought process score, the step score requires the correct calculation result. For each correct step, give 1 point, and 0 points for incorrect ones. Then calculate the average score for the steps.

——Final answer score: Match the model's final answer with the ground truth answer. For a prove question, see whether the question has been correctly proved. If it matches, give 1 point; otherwise, give 0 points.

**Prompt for Foundational knowledge dataset Evaluation (Part 2)**

You don't need to provide a detailed explanation for your scores, just give the scores directly.

——@Question: {question}

——@Ground-truth answer: {answer}

——@Model output: {extracted steps}

——@Thought process average score:

——@Step average score:

——@Final answer score:

**Prompt for Statistical research dataset Task**

You are a helpful assistant designed to help with statistical problems. When given a statistical question, you should answer the question according to the user's prompt. This is a theorem proof problem from a statistics journal, involving high-difficulty theorem reasoning. Based on the supplementary materials and the theorem statement provided, please provide a detailed proof process.

——@Relevant Supplementary Materials: {supplementary materials}

——@Problem: {question}

——@Please provide your proof process and results:

**Prompt for Statistical research dataset Task Evaluation**

You are a rigorous grader specializing in statistical proof assessments. Your core task is to judge whether a predicted answer aligns with the gold standard answer for theoretical questions in statistical proofs, following these strict and precise criteria:

**1. Criteria for Non-Constant Components** Non-constant components refer to expressions dependent on dimensions, sample sizes, or variables (e.g., $O(\sqrt{n})$, $\|\widehat{\pi}_1 - \pi_1^*\|_2$). For such components:

- The order of the *dominant term* (the term determining the overall order/magnitude) must be *exactly consistent* with the gold standard to count as correct.
- If the result contains multiple terms, correctness is determined solely by the dominant term: only when the dominant term's order is strictly identical, the answer is correct (non-dominant terms with lower order/magnitude may be included or omitted without penalty).
- Minor differences in non-essential constant coefficients (that do not alter the core order of the dominant term) do not affect correctness; however, any discrepancy in the dominant term's order renders the answer incorrect.

**2. Criteria for Constant Components** Constant components refer to fixed numerical values, constant terms in final results, or fixed coefficients definitive to the conclusion (e.g., $1-\alpha$, constant 2 in $2\sigma$). For such components: - The predicted answer must be *exactly identical* to the gold standard.

**3. Criteria for Formatting Differences** Purely formatting differences (e.g., spacing, parentheses placement, LaTeX notation variations) that do not alter the mathematical value or the order of the dominant term in non-constant components will not be penalized—count these as correct.

**Final Requirement** After evaluating against the above criteria, provide a clear verdict of "Correct" or "Incorrect".

# D Example of statistic research question

**Context (Part 1)**

Let $(X, Z, Y)$ denote a full-data vector and consider iid calibration observations

$$O_i = (C_i, G_{C_i}(X_i, Z_i, Y_i)), \quad i = 1, \ldots, n,$$

where the coarsening indicator $C$ takes values in $\{0, 1, \infty\}$ with

$$G_0(X, Z, Y) = X, \quad G_1(X, Z, Y) = (X, Z), \quad G_\infty(X, Z, Y) = (X, Z, Y).$$

Assume a missing-at-random (MAR) mechanism such that for $k \in \{0, 1, \infty\}$,

$$P(C = k \mid X, Z, Y) = \omega(k, G_k(X, Z, Y)).$$

Define the stage-2 outcome regression

$$m_2^*(\theta, x, z) = P(Y \leq \theta \mid X = x, Z = z, C = \infty),$$

the stage-1 outcome regression

$$m_1^*(\theta, x) = \int m_2^*(\theta, x, z)\, dF(z \mid C \geq 1, x),$$

the stage-2 propensity

$$\pi_2^*(x, z) = P(C = \infty \mid X = x, Z = z, C \geq 1),$$

and the stage-1 propensity

$$\pi_1^*(x) = \frac{P(C = 0 \mid X = x)}{P(C \geq 1 \mid X = x)}.$$

**Context (Part 2)**

Let estimators $\widehat{m}_2(\theta, x, z), \widehat{m}_1(\theta, x), \widehat{\pi}_2(x, z), \widehat{\pi}_1(x)$ be obtained from an independent training split. For $i = 1, \ldots, n$ define

$$G_i(\theta; \widehat{m}_2, \widehat{m}_1, \widehat{\pi}_2, \widehat{\pi}_1) = \mathrm{IF}(\theta, C_i, G_{C_i}(X_i, Z_i, Y_i); \widehat{m}_2, \widehat{m}_1, \widehat{\pi}_2, \widehat{\pi}_1),$$

where $\mathrm{IF}(\cdot)$ is given by

$$\begin{aligned}\mathrm{IF}(\theta, c, G_c(x, z, y); m_2, m_1, \pi_2, \pi_1) = {} & \frac{1\{c = \infty\}\, \pi_1(x)}{\pi_2(x, z)}\big(1\{y \le \theta\} - (1 - \alpha)\big) \\ & - 1\{c \ge 1\}\, \pi_1(x)\Big(\frac{1\{c = \infty\}}{\pi_2(x, z)} - 1\Big) \\ & \quad \cdot \big[m_2(\theta, x, z) - (1 - \alpha)\big] \\ & - \big(1\{c \ge 1\}\, \pi_1(x) - 1\{c = 0\}\big)\big[m_1(\theta, x) - (1 - \alpha)\big].\end{aligned}$$

**Theorem:** : Under the assumption that for each fixed $x$ and $z$, the maps $\theta \mapsto \widehat{m}_1(\theta, x)$ and $\theta \mapsto \widehat{m}_2(\theta, x, z)$ are right-continuous, the robust split conformal procedure guarantees the following finite-sample lower bound:

$$\begin{aligned}P(Y_{n+1} \le \widehat{r}_\alpha(X_{n+1}) \mid C_{n+1} = 0) \ge 1 - \alpha & - \frac{\|\widehat{\pi}_1 - \pi_1^*\|_2 \cdot \|\pi_2^*\|_4^{n_2/2}}{\|\widehat{\pi}_2\|_4 \cdot P(C = 0)} \\ & - \frac{\|\widehat{m}_1 - m_1^*\|_2 \cdot \sup_\theta \|m_1^*(\theta, X) - \widehat{m}_1(\theta, X)\|_2}{P(C = 0)} \\ & - \frac{\|\widehat{m}_2 - m_2^*\|_4 \cdot \sup_\theta \|m_2^*(\theta, X, Z) - \widehat{m}_2(\theta, X, Z)\|_4}{P(C = 0)}\end{aligned}$$

without assuming any additional continuity or moment conditions beyond the right-continuity stated.

**Question & Answer**

**Question:** Using the calibration set, define for any baseline $x$ of a test unit,

$$\widehat{r}_\alpha(x) = \inf\Big\{\theta : \frac{\sum_{i=1}^{n} G_i(\theta; \widehat{m}_2, \widehat{m}_1, \widehat{\pi}_2, \widehat{\pi}_1)}{n+1} + \frac{\widehat{m}_1(\theta, x) - (1-\alpha)}{n+1} \geq 0\Big\}.$$

State the finite-sample lower bound on the conditional coverage

$$P(Y_{n+1} \leq \widehat{r}_\alpha(X_{n+1}) \mid C_{n+1} = 0)$$

that the robust split conformal procedure guarantees in terms of $\alpha, \widehat{\pi}_1, \widehat{\pi}_2, \widehat{m}_1, \widehat{m}_2, \pi_2^*, m_2^*, m_1^*, \pi_1^*$, and $P(C = 0)$. Do not assume any additional continuity or moment conditions beyond the right-continuity stated.

**Answer:**

$$\begin{aligned} 1 - \alpha &- \frac{\|\widehat{\pi}_1 - \pi_1^*\|_2 \cdot \|\pi_2^*\|_4^{n_2/2}}{\|\widehat{\pi}_2\|_4 \cdot P(C=0)} \\ &- \frac{\|\widehat{m}_1 - m_1^*\|_2 \cdot \sup_\theta \|m_1^*(\theta, X) - \widehat{m}_1(\theta, X)\|_2}{P(C=0)} \\ &- \frac{\|\widehat{m}_2 - m_2^*\|_4 \cdot \sup_\theta \|m_2^*(\theta, X, Z) - \widehat{m}_2(\theta, X, Z)\|_4}{P(C=0)} \end{aligned}$$

(The bound accounts for estimation errors in propensities $(\pi_1^*, \pi_2^*)$ and outcome regressions $(m_1^*, m_2^*)$, scaled by $P(C = 0)$ and adjusted by the confidence level $\alpha$.)

# E Additional Tables

Table 4: Foundational Dataset composition: disciplinary structure

| Level | Domain | Subdomain | Number of Questions |
|---|---|---|---|
| UG | Probability | Elementary Probability | 1461 |
| | | Elementary Time Series | 176 |
| | | Stochastic Process | 274 |
| | Statistics | Elementary Statistics | 1531 |
| | | Linear Model | 910 |
| | | Multivariate Statistics | 243 |
| | | Causal Inference | 223 |
| | | Statistical Computing | 355 |
| | Machine Learning | General Machine Learning | 668 |
| | | Deep Learning | 343 |
| | | Data Science | 152 |
| | Total | | 6336 |
| G | Probability | Advanced Probability | 636 |
| | | Advanced Time Series Analysis | 106 |
| | | Information Theory | 122 |
| | Statistics | Advanced Statistics | 2708 |
| | | Empirical Process | 933 |
| | | Nonparametric Statistics | 656 |
| | | Causal Inference | 170 |
| | Machine Learning | General Machine Learning | 1171 |
| | | Deep Learning | 245 |
| | | Reinforcement Learning | 200 |
| | | Convex Optimization | 534 |
| | Total | | 7481 |
| **Overall Total** | | | **13817** |

Table 5: Research task distribution across areas and theoretical properties

| **Standard** | **Domain** | **Subdomain / Property Type** | **Number** |
|---|---|---|---|
| Research Area | Probability | Probability Theory and Stochastic Processes | 266 |
| | Statistics | Classical Statistical Modeling and Inference | 855 |
| | | Causal Inference and Experimental Design | 481 |
| | | High-dimensional Data Modeling | 453 |
| | | Other Emerging Directions | 20 |
| | Machine Learning | Graphs and Structured Data | 127 |
| | | Bayesian and Generative Models | 104 |
| | | Deep Learning and Reinforcement Learning | 68 |
| | Total | | 2374 |
| Theoretical Property | | Asymptotic Properties | 828 |
| | | Distributional Properties | 298 |
| | | Generalization and Error Bounds | 292 |
| | | Identifiability and Consistency | 291 |
| | | Optimality Results | 232 |
| | | Testing Validity | 215 |
| | | Convergence and Stability | 132 |
| | | Structural Guarantees | 86 |
| | Total | | 2374 |