

WEEKLY DIGEST: HUMANITIES & CULTURAL STUDIES

Biases in AI Evaluation and Historical Reexamination

Recent studies have revealed biases in Large Language Models (LLMs) used for evaluation, leading to unfair assessments. Meanwhile, a historian has reexamined a Chinese tea chest label, challenging its previously believed connection to the Boston Tea Party. These developments highlight the importance of critically evaluating AI models and reexamining historical artifacts. By doing so, researchers can gain a more accurate understanding of the past and improve AI systems.

KEY FINDINGS

- ✓ Researchers have identified biases in Large Language Models (LLMs) used as evaluators, including position bias and verbosity bias.
- ✓ These biases arise from the models' training data and can lead to unfair assessments.
- ✓ A historian has reexamined a Chinese tea chest label and discovered that it was not a relic from the Boston Tea Party as previously believed.

IMPLICATIONS

- The discovery of biases in LLMs highlights the need for more transparent and explainable AI evaluation systems.
- The reexamination of historical artifacts can challenge prevailing narratives and provide new insights into the past.
- Improving AI models and historical research can have significant implications for fields such as education, law, and cultural heritage preservation.

Key Metrics: None explicitly mentioned, but the identification of position bias and verbosity bias in LLMs implies a quantitative analysis of model performance.

1. Why You Shouldn't Always Trust LLMs as Judges: Understanding Bias in Automated Evaluation

Category: Research

Departments: Computer Science, Political Science & Public Administration, Philosophy, Psychology

Topics: Large Language Models, Bias Mitigation, Evaluation Metrics, Fairness in AI

Impact Score: 23.0

Rubric Scores

Research Relevance: 4

Educational Value: 4

Innovation / Novelty: 3

Practical Impact: 3

Interdisciplinary Potential: 4

Ethical / Policy Implications: 5

Researchers have identified biases in Large Language Models (LLMs) used as evaluators, which can lead to unfair assessments. The biases, including position bias and verbosity bias, arise from the models' training on human-written text and tendency to rely on priors rather than evidence. Specifically, LLMs are prone to favoring answers based on their position or length, rather than quality, particularly when evaluating comparable answers. These findings suggest that LLMs should not be trusted as sole judges in evaluation tasks.

<https://www.analyticsvidhya.com/blog/2026/08/why-you-shouldnt-trust-llm-judges/>

2. The mystery of the Chinese tea chest label

Category: Research

Departments: History

Topics: AI Ethics & Safety, Historical Document Analysis, Linguistic Research Methods

Impact Score: 11.0

Rubric Scores

Research Relevance: 2

Educational Value: 2

Innovation / Novelty: 2

Practical Impact: 1

Interdisciplinary Potential: 3

Ethical / Policy Implications: 1

Historian Tristan Brown has reexamined a Chinese tea chest label previously believed to be a relic from the Boston Tea Party. Through archival and linguistic research, Brown discovered that the label was actually created nearly a century after the event, in the 1860s or 1870s, and features the name of an American trading firm, Smith, Archer & Co. The finding reveals a new story connecting Chinese migration, Pacific commerce, and the construction of historical memory in America. Brown's research was published in the American Historical Review and explores how ordinary objects acquire historical authority.

<https://news.mit.edu/2026/mystery-of-the-chinese-tea-chest-label-0811>