Abstract: Large language models (LLMs) are increasingly used in archival financial economics research, particularly for large-scale text analysis. LLMs have demonstrated strong performance across a diverse range of tasks, including financial text classification, summarization, and predictions. However, because LLM outputs are inherently stochastic, their application in archival research raises significant concerns about robustness and replicability. I demonstrate that commonly used single-run LLM measures can generate economically misleading inference such as coefficient instability and spurious significance. I identify two primary use cases of LLMs in archival research and highlight the distinct sources of variation associated with each. For each use case, I propose a statistical algorithm designed to mitigate stochasticity and ensure replicable empirical results. The algorithms are task-agnostic and provide sufficient statistics for reproducible research.