Reliable Economics Research in the Age of Generative AI*

Valerie Zhang† Christopher Rigsby‡ Sean M. Song§

Abstract

Despite the swift advancements in LLM-enhanced abilities, general generative AI models still hallucinate and cannot reliably distinguish theoretically plausible from implausible foundations. To determine whether domain-specific generative AI can consistently improve academic research in empirical economics, finance, and accounting, we construct a retrieval-augmented generation (RAG) system based on an extensive corpus of accounting, finance, and economics papers and code. Compared with standalone commercial LLMs, the system nearly eliminates hallucinations, retrieves more relevant prior literature, and produces faster and more reliable replications. Our findings suggest that a domain-specific knowledge base enhances the reliability of AI agents in conducting empirical economics, finance, and accounting research, opening a path forward by automating transparent and theoretically grounded analyses.

*We thank all participants in the seminar at the University of California, Berkeley. This research acknowledges the financial support from the University of California, Berkeley.

†Corresponding Author; Haas School of Business, U.C. Berkeley. Email: valerie.zhang@haas.berkeley.edu

‡School of Business and Economics, Vrije Universiteit Amsterdam. Email: c.m.rigsby@vu.nl

§Haas School of Business, U.C. Berkeley. Email: sean_song@berkeley.edu

1 Introduction

What happens to empirical research in economics, finance, and accounting when the marginal cost of producing an academic paper converges to zero? Recent advances in generative artificial intelligence (AI) force the profession to take this question seriously. Large language models (LLMs) can automate substantial portions of the research process and write full drafts of projects at fractions of the usual time (e.g., Ciconte et al., 2025; Novy-Marx and Velikov, 2025). Although such tools could expand scholarly capacity, core concerns about research integrity and replicability remain (Eisfeldt and Schubert, 2025; Korinek, 2023; de Kok, 2025). Further, the peer-review infrastructure might collapse from an onslaught of such papers (Ciconte et al., 2025; Novy-Marx and Velikov, 2025). These risks are particularly consequential in any field where credibility depends on transparent links between prior literature, hypothesis development, and empirical execution—a description that fits empirical economics, finance, and accounting alike—so new methods to increase the credibility of AI-generated research are necessary. Retrieval-augmented generation systems (RAGs) offer a promising solution to one of the central limitations of LLMs: citation hallucinations and the inability to reliably distinguish theoretically plausible from implausible foundations.

We demonstrate that a domain-specific RAG architecture can eliminate citation hallucinations (reducing rates from as high as 36.7% to almost zero) when generating literature reviews on accounting research topics. We document these findings using Etalia, [1] a RAG system constructed from 8,075 papers published in leading accounting and finance journals. We examine two research questions. First, can a domain-specific RAG reduce hallucination rates when LLMs generate academic literature reviews? Second, does RAG improve the quality and accuracy of LLM-generated code for canonical empirical replications?

The questions matter for both researchers and journals. LLMs promise substantial effi-

[1] A partial, publicly accessible version of Etalia is available at `https://reliableaccountingresearch.ai`. The system is built and operated in accordance with publisher terms of use and applicable copyright law; it indexes vector embeddings and bibliographic metadata to support retrieval rather than redistributing full article text.

ciency gains in literature synthesis and coding tasks, yet hallucinated citations undermine trust and impose verification costs on readers. A literature review citing nonexistent studies wastes scholarly attention and damages author credibility. At the editorial level, automated evaluation of replication packages could improve pre-publication screening and strengthen reproducibility standards. Both applications require grounding model outputs in verified domain knowledge rather than relying solely on general training corpora (Eisfeldt and Schubert, 2025).

How do we ground model outputs in verified domain knowledge? With traditional large language models, users provide a prompt, and the model generates a response based solely on patterns learned during training to predict the next token. A RAG introduces an intermediary retrieval layer. Upon receiving a user prompt, the RAG first searches a structured knowledge base to identify the relevant context needed to answer the user's question. Next, the system passes both the user's prompt and the retrieved evidence to the language model, enabling responses grounded in verifiable scientific material rather than parametric memory alone.

We describe the construction of Etalia, a RAG built using accounting knowledge, in greater detail below. We construct Etalia by parsing PDFs from five leading journals, extracting structured sections using font-size heuristics, and embedding text into high-dimensional vectors with Sentence-BERT. The system integrates 376 replication packages spanning five programming languages, 93 WRDS sample codes, and 143 WRDS documentation manuals. Through an application programming interface (API), Etalia retrieves relevant context before prompting commercial LLMs, thereby constraining generation to field-specific evidence and methodological precedent.

The knowledge base underlying Etalia draws on 8,075 papers from five premier accounting and finance journals: the Review of Financial Studies (2,349 papers), The Accounting Review (2,136), the Journal of Accounting and Economics (1,624), the Journal of Accounting Research (1,060), and the Review of Accounting Studies (906). Approximately 4.7% of these

papers (376 total) include publicly available replication packages spanning five programming languages, sourced from the Review of Financial Studies and the Journal of Accounting Research. We supplement this corpus with 93 WRDS sample codes and 143 WRDS documentation manuals. For each paper, we directly extract full-text content from core sections—title, abstract, introduction, literature review, hypothesis development, research design, results, and conclusion—and convert statistical tables into structured records capturing variable names, sample sizes, coefficients, and t-statistics. A locally hosted large language model performs higher-order extraction tasks, including section segmentation, summary statistics parsing, hypothesis identification, and code documentation. We embed all textual and code content into a 384-dimensional vector space using Sentence-BERT, yielding approximately 794,175 searchable entries that support semantic retrieval at the paper, section, table, code, and variable levels.

We compare three baseline LLMs—GPT-5.2 (OpenAI), Claude 4.5 Sonnet (Anthropic), and Gemini 3 Pro (Google)—against the same models augmented with Etalia's retrieval system. For each test, baseline models receive only the user prompt, while augmented models receive the prompt together with context retrieved from Etalia's corpus, isolating the effect of retrieval augmentation while holding the underlying language model constant. Our first validation test examines citation accuracy: we prompt each of six model configurations to generate up to 10 citations per topic across 20 accounting research topics—10 classic topics representing established areas and 10 modern topics representing emerging fields—and validate every citation against the Crossref API using fuzzy title matching. Our second validation test evaluates code generation for empirical replications of two seminal papers, Sloan (1996) and Basu (1997), measuring whether models produce working code that connects to the WRDS database and replicates the main regression in each paper. We then compare estimated coefficients, sample sizes, and sample periods against published values.

Our analysis yields two principal findings. First, RAG eliminates citation hallucinations while preserving semantic relevance across both established and emerging accounting topics.

On classic topics, standalone LLMs hallucinate between 5% and 11% of citations; on modern topics, hallucination rates diverge sharply, with one model fabricating more than one in three references. All three Etalia-augmented models achieve 0% hallucination across both topic categories while maintaining comparable or higher relevance scores. Across 10 independent runs, augmented models sustain 0% hallucination with zero variance, whereas standalone models average between 6% and 35% depending on model and topic category. In emerging research areas in which training data are thin, augmented models also improve topical relevance, producing scores roughly 5 to 40 percentage points higher than their standalone counterparts.

Second, RAG improves code generation for empirical replications. For Sloan (1996), all six model configurations generate working code, but baseline models struggle with variable construction and sample identification: one baseline produces near-zero coefficients suggesting a fundamental coding error, and others use sample periods extending decades beyond the original 1962–1991 window. Etalia-augmented models more accurately match the original sample period and produce accrual and cash flow coefficients closer to the published estimates. For Basu (1997), the replication is more challenging—baseline models require up to three iterations to produce working code, and their coefficient estimates exhibit scaling errors or incomplete specifications. Etalia-augmented models produce estimates closer to Basu's published values; the best-performing augmented configuration recovers an intercept identical to the original estimate, closely matches the sample period of 1963–1990, and approximates the original sample size, whereas baseline models deviate substantially on one or more dimensions. Across both replications, RAG-provided context—particularly the original papers' variable definitions, sample restrictions, and regression specifications—guides models toward scientifically meaningful output rather than syntactically correct but empirically inaccurate code.

Collectively, these results indicate that domain-specific retrieval can transform LLMs from probabilistic text generators into research assistants capable of transparent and verifi-

able scholarly tasks.

Etalia also enables tracing of hypothesis development, considering both included and excluded references. Embedding each section of a paper allows researchers to rapidly evaluate whether hypotheses incorporate the relevant literature and whether omitted studies materially alter the proposed theoretical mechanism. This capability is increasingly important as generative AI lowers the cost of producing statistically significant but weakly motivated findings (Ciconte et al., 2025; Novy-Marx and Velikov, 2025). While stricter multiple-testing thresholds can mitigate false discoveries (Harvey et al., 2016; Heath et al., 2023), statistical safeguards alone cannot ensure that empirical tests emerge from coherent theoretical frameworks. Auditing hypothesis development provides a complementary mechanism by clarifying which intellectual traditions are incorporated and ignored when interpreting empirical evidence.

Our caveat is that we do not claim Etalia-augmented models can replace existing LLMs in assisting researchers with research tasks. Rather, we show that Etalia complements advances in LLMs. While LLMs provide the reasoning layer, Etalia serves as a knowledge base from which models can efficiently retrieve domain-specific information. In our setting, this domain-specific information consists of a collection of research questions, evidence, and methodologies accumulated over decades of accounting research. Etalia-augmented models leverage the capabilities of LLMs while grounding their outputs in reliable evidence. As LLM capabilities continue to improve, performance on many of the tasks we document may likewise improve. But retrieval-augmented systems such as Etalia will likely remain complementary to these newer models, as structured access to accounting knowledge enhances an LLM's reliability regardless of advances in underlying model capabilities.

Our study contributes to two strands of literature. First, we extend research on generative AI in accounting by demonstrating that domain-specific architectures can improve reliability relative to general-purpose models (Korinek, 2023; Eisfeldt and Schubert, 2025; Novy-Marx and Velikov, 2025). Second, we contribute to the replication literature by demonstrating that

domain-specific RAG can lower the costs that traditionally discourage systematic replication (Hail et al., 2020; Jensen et al., 2023). More broadly, our evidence suggests a path forward in which generative AI augments scholarly judgment by replicating empirical results and evaluating the theoretical foundations of these results.

2 Literature Review

2.1 Generative AI in Applied Economics Research

Recent research in economics, finance, and accounting uses generative AI, especially LLMs and their derivatives, as scalable tools for various stages of the research process (e.g., Korinek, 2023; Ludwig and Mullainathan, 2024; Eisfeldt and Schubert, 2025; Novy-Marx and Velikov, 2025). Generative AI can aid in the identification of research questions and hypotheses (Ludwig and Mullainathan, 2024), empirical design, analysis, and the creation of full research papers (Korinek, 2023; Ciconte et al., 2025; Novy-Marx and Velikov, 2025). To place structure on this tool, Ludwig et al. (2025) formalize an econometric framework so that estimation and prediction tasks can incorporate LLM outputs.

Researchers most commonly use LLMs for text generation, capture, and inference based on regulatory filings, firm disclosures, and data vendors (e.g., Breitung and Müller, 2025; de Kok, 2025). Many researchers focus on converting unstructured text into economically meaningful information for textual analysis. Domain-adapted LLMs (e.g., FinBERT) improve the measurement of sentiment, ESG content, and disclosure informativeness in financial texts, particularly when labeled training data are limited (Huang et al., 2023). General-purpose generative models can perform complex tasks but raise concerns related to prompt design, construct validity, and replicability (de Kok, 2025; Brodeur et al., 2025).

Current LLMs have several disadvantages. Data mining, spurious theoretical foundations, and mass production of low-quality content could damage knowledge discovery (Novy-Marx and Velikov, 2025). When LLMs select seemingly exogenous shocks for causal analysis and

write entire papers quickly, false significance appears almost surely (Ciconte et al., 2025). Further, these LLMs rely on general training and do not have a memory of field-specific literature, empirical methodology, and coding for proper context (Eisfeldt and Schubert, 2025). RAGs can address some of these problems (Eisfeldt and Schubert, 2025). Instead of predicting text by using only the training model, a RAG consults a database of specific documents, ranks the relevance of parts of the database, and provides more context when making an LLM request (Eisfeldt and Schubert, 2025). The RAG architecture allows researchers to quantify abstract topics, such as corporate culture (Li et al., 2026), and understand detailed technical text, such as zoning regulations (Bartik et al., 2025).

2.2 Replication in Applied Economics

Concerns about a replication crisis (or at least a weakening of credibility) have become widespread in accounting, finance, and economics (Maniadis et al., 2017; Hail et al., 2020; Jensen et al., 2023; Ankel-Peters et al., 2023; Breuer and Hofmann, 2025). Survey evidence in accounting, for example, indicates that a majority of accounting researchers think a lack of replication is common and consequential, even though few failed replication attempts are published (Hail et al., 2020). These beliefs are probably not speculative. Instead, they likely reflect accumulated private information from failed replication attempts, sensitivity analyses conducted during extensions, informal exchanges among researchers, and awareness of the fragility of researchers' own work (Salterio, 2014; Sharma, 2023). Such evidence rarely appears in top journals, so disciplines have an uneasy skepticism but limited systematic studies (Ankel-Peters et al., 2023). As a result, uncertainty about the credibility of key findings remains despite some successful replications (Luo et al., 2024). Further, in the absence of systematic replication and timely correction, flawed findings can persist and continue to shape subsequent research agendas (Joëts and Mignon, 2026).

This gap persists because researchers must choose between replication and original research. As Hail et al. (2020) argue, a replication is likely to lead to lower knowledge compared

to fresh studies, and the incentives for researchers, particularly for PhD students and junior faculty members, strongly favor avoiding conflict with established results. Replications typically face lower publication probabilities, weaker career rewards, and higher coordination costs (Salterio, 2014; Sharma, 2023). Conducting meaningful replications frequently requires reconstructing complex data pipelines, reverse-engineering undocumented coding decisions, and navigating ambiguities in variable construction and sample selection. Even when replications are published, they are often embedded within extensions rather than organized as comprehensive reassessments (Luo et al., 2024).

Empirical asset pricing, specifically the anomaly literature, rewards replication in top journals (McLean and Pontiff, 2016; Harvey et al., 2016; Chordia et al., 2020; Hou et al., 2020; Chen and Zimmermann, 2022; Bryzgalova et al., 2023; Chen and Velikov, 2023; Jensen et al., 2023; Cakici et al., 2024). Because researchers have frequently raised suspicions about data mining and multiple testing, research teams have answered the calls for systematic replication. The techniques and methods from the rich set of replications and the transparent data availability offer a template for other areas (Chen and Zimmermann, 2022; Jensen et al., 2023).

Researchers who want to replicate a paper must be meticulous, possibly more than the original authors, because failure to replicate does not necessarily imply that the original results cannot be replicated or reproduced (See, e.g., Frankel et al., 2002; Ashbaugh et al., 2003; Sharma, 2023). AI agents can attempt substantial replications across papers and different time periods with less manual intervention. Just as AI agents are lowering transaction costs in consumer markets (Shahidi et al., 2025), a RAG combined with plentiful research data can lower the time and resources required for replications (Novy-Marx and Velikov, 2025). Especially as researchers produce more research at increasing rates and journals struggle to cope with the expanded supply of papers (Novy-Marx and Velikov, 2025; Ciconte et al., 2025; Eisfeldt and Schubert, 2025), rapid replications will be more valuable.

3 Etalia's RAG Infrastructure

In this section, we describe the core infrastructure underlying Etalia's retrieval-augmented generation (RAG) framework. Traditional large language models operate in a closed-loop setting: users provide a prompt, and the model generates a response based solely on patterns learned during training to predict the next token.

Etalia introduces an intermediary retrieval layer. Upon receiving a user prompt, the system first searches a structured knowledge base to identify the relevant context needed to answer the user's question. Next, the system passes both the user's prompt and the retrieved evidence to the language model, enabling responses grounded in verifiable scientific material rather than parametric memory alone. We describe the construction and organization of the knowledge base in greater detail below.

3.1 Dataset Construction

Our corpus comprises 8,075 papers from five premier accounting and finance journals spanning publications from the 1970s through 2025. Table 1 Panel A summarizes the corpus composition. The corpus contains papers from the Review of Financial Studies (2,349 papers), The Accounting Review (2,136 papers), Journal of Accounting and Economics (1,624 papers), Journal of Accounting Research (1,060 papers), and Review of Accounting Studies (906 papers).

Table 1 Panel B documents additional replication code packages and sample code in our corpus. Approximately 4.7% of papers (376 total) include publicly available replication packages across five programming languages (Stata, SAS, R, Python, MATLAB). Replication packages are sourced from the Review of Financial Studies (272 packages) and the Journal of Accounting Research (104 packages). To complement this corpus, we incorporate 93 WRDS sample codes and 143 WRDS documentation manuals. The breadth and depth of this corpus are critical for both research questions: a comprehensive collection of published papers

provides the retrieval base needed to eliminate citation hallucinations, while the inclusion of replication code and WRDS documentation supplies the domain-specific programming context required to improve LLM-generated empirical code.

3.2 Text Extraction and Structural Parsing

We use automated PDF parsing to process each paper and construct the Etalia knowledge base. As Table 2 Panel A shows, we directly extract full-text content from core paper sections, including the title, abstract, introduction, literature review, hypothesis development, research design, results, and conclusion.

In addition to unstructured text, we convert each paper into structured research artifacts. Specifically, we extract summary statistics (variable names, sample sizes, means, standard deviations, minima, and maxima), regression tables (dependent variables, independent variables of interest, estimated coefficients, and t-statistics), and correlation matrices (variable pairs and pairwise correlations) to capture the quantitative backbone of empirical studies. This structured representation enables systematic comparison across papers and supports downstream analytical tasks.

We use locally hosted large language models to perform higher-order extraction tasks, including identifying section boundaries, parsing statistical tables, recovering hypothesized relationships, and documenting replication code.

Table 2 Panel B reports the specific type of information we use local LLMs to extract for each paper. First, the LLM performs section segmentation by identifying section boundaries from font-size changes and header patterns, enabling downstream retrieval at the section level rather than treating each paper as a monolithic document. Second, the LLM parses summary statistics tables, extracting variable names, sample sizes, means, and standard deviations into structured records that can be compared across studies. Third, the LLM identifies stated hypotheses (e.g., H1, H2, H3) and their predicted relationships between variables, linking each hypothesis to the empirical tests reported in subsequent sections.

Fourth, the LLM documents replication code by parsing inline comments and docstrings to catalog the purpose of each script and its role in the overall data pipeline. Together, these extraction tasks structure the knowledge base so that retrieval queries return not only relevant text but also the empirical details—variable definitions, statistical specifications, and coding patterns—that ground LLM outputs when generating literature reviews and replication code.

3.3 Code Analysis and Workflow Reconstruction

For papers with replication packages, we implement language-specific analyzers that parse code to extract computational workflows, data operations, and statistical methods. Each analyzer understands the syntax and semantics of its target language, detecting regression commands, variable definitions, data operations, and sample restrictions. The Stata analyzer identifies commands like reghdfe and xtreg; the SAS analyzer parses PROC SQL queries; the Python analyzer detects pandas operations and statistical libraries; the R analyzer recognizes lm, plm, and data manipulation operations; and the MATLAB analyzer identifies optimization routines and matrix operations.

We extract WRDS query patterns from 93 official sample codes, building a pattern library of 15 distinct data access patterns that enables automatic detection of data sources even when not explicitly mentioned. The system tracks 611 unique financial variables and documents 127 distinct variable construction patterns, representing multi-step transformations as dependency graphs. Workflow reconstruction analyzes file naming conventions, explicit dependencies, and script purposes to chain together the complete data flow from raw inputs through intermediate datasets to final analysis, reconstructing workflows with an average of 5.2 steps per paper.

3.4 Enhanced Content Understanding

We use large language models to move beyond structural extraction to semantic interpretation. Specifically, we use a local LLM on an Nvidia A100 GPU (Llama 3.3 70B-Instruct) to analyze each section to generate concise summaries, extract methodological details, identify key findings, and note table/figure references. For tables, the model interprets regression components including dependent variables, independent variables of interest, control variables, fixed effects specifications, and statistical significance patterns.

We also add additional features to understand regression specifications presented in tables. The local LLM extracts coefficient structures, model specifications, and sample information from statistical tables. Next, we use specific hypothesis extraction methods to identify predictions in the text of the paper and link them to empirical tests using language model analysis. We also document how the final sample is constructed, and which data filters are used in the study. Relatedly, we ask the LLM to summarize how authors use special treatment for outliers and missing data; how authors transform variables; and how authors merge different datasets to understand how the final sample is created.

3.5 Vector Embedding and Semantic Search

We embed all textual and code content into a 384-dimensional vector space using Sentence-BERT (all-MiniLM-L6-v2), enabling natural language queries based on semantic meaning rather than keyword matching. We create embeddings at five levels of granularity: paper-, section-, table-, code-, and variable-level.

The complete corpus contains approximately 794,175 searchable entries stored in NumPy compressed format with accompanying JSON metadata. We implement a hybrid retrieval architecture combining vector similarity search with structured metadata filters.

To implement our research design, we host Etalia on a remote server that exposes an API for programmatic querying. The system integrates semantic search over a structured academic corpus with large language model (LLM) generation, allowing responses to be

grounded in retrieved scholarly context. Users can constrain retrieval using metadata filters such as journal, publication year, author, and content type (e.g., code segments), enabling precise and reproducible queries.

4 Experimental Design

We compare three baseline LLMs—GPT-5.2 (OpenAI), Claude 4.5 Sonnet (Anthropic), and Gemini 3 Pro (Google)—against the same models augmented with Etalia's RAG system. For each test, baseline models receive only the prompt; augmented models receive the prompt plus retrieved context from Etalia's corpus. This design isolates the effect of retrieval augmentation while holding the underlying language model constant. All models use temperature settings of 0.7.

4.1 Literature Review Test

Our first validation test examines citation accuracy when LLMs generate literature reviews. We prompt each of six model configurations to provide up to 10 relevant academic citations per topic in structured JSON format, specifying author names, publication year, paper title, and journal name. The prompt constrains citations to five leading accounting journals: The Accounting Review, Journal of Accounting Research, Journal of Accounting and Economics, Review of Accounting Studies, and Contemporary Accounting Research. All models use a generation temperature of 0.7, reflecting realistic research interactions rather than near-deterministic settings.

We design 20 research topics across two categories: 10 classic topics representing established accounting research areas (e.g., earnings management and accruals quality, audit quality and auditor independence) and 10 modern topics representing emerging fields with sparser published literature (e.g., machine learning and AI in accounting, ESG disclosure, cryptocurrency and blockchain accounting). This split allows us to examine whether halluci-

nation patterns differ between well-documented research areas and newer domains in which training data are thinner.

For each topic, we compare three standalone LLMs—GPT-5.2, Claude, and Gemini—against the same models augmented with Etalia's retrieval system, which injects up to 20 semantically retrieved papers from the corpus as context before generation and instructs the LLM to cite only from the retrieved set.

We validate each generated citation against the Crossref API, an independent third-party registry of scholarly DOIs. A citation is classified as verified if Crossref returns a matching record with a fuzzy title similarity score of at least 75%, computed using the token set ratio algorithm from the rapidfuzz library. When the initial query filtered by publication year returns no match, we retry without the year constraint to account for minor discrepancies.

We also measure and report topical relevance among papers that are cited by each model using the average cosine similarity of the titles of the citation responses to the literature query. We use OpenAI text-embedding-3-small embedding models to calculate cosine similarity.

Table 3 reports four metrics for each model: the number of citations generated, the number verified in Crossref, the hallucination rate, and the mean semantic relevance score. To assess the stability of these results, we repeat the full pipeline 10 times, re-querying each model on all 20 topics at the same temperature and re-validating every citation against Crossref. We cache Crossref lookups to ensure deterministic validation; variation across runs arises entirely from stochastic LLM generation. Panel B of Table 3 reports the mean and standard error of the hallucination rate for each model configuration, split by topic category.

4.2 Replication Test

Our second validation test evaluates code generation quality for empirical replications of two seminal papers: Sloan (1996), documenting the accruals anomaly, and Basu (1997), documenting the asymmetric timeliness of earnings. We prompt each model to generate Python code that connects to the WRDS database and replicates the main regression in

each paper. For Sloan (1996), the code should query Compustat Annual Fundamentals, calculate balance sheet accruals, and estimate the regression of earnings on accruals and cash flows from operations (Table 3, Panel A). For Basu (1997), the code should estimate the reverse earnings-return regression $X_t/P_{t-1} = \beta_0 + \beta_1 D_t + \beta_2 R_t + \beta_3 D_t \times R_t + \varepsilon_t$ (Table 1, Panel A). We execute the generated code in a controlled environment with valid WRDS credentials and measure four metrics: whether the model generated final working code, the number of iterations required to produce working code, the average time per iteration, and total execution time. This test evaluates whether LLMs can produce functional research code that interacts with real financial databases.

Our third validation test compares the accuracy of LLM-generated regression estimates against published ground truth values. Using the code generated in the second test, we extract estimated coefficients and compare them against the values reported in Sloan (1996) Table 3, Panel A and Basu (1997) Table 1, Panel A. We also compare sample sizes and sample periods to assess whether models accurately reconstruct the original data. This test evaluates not just whether the code runs, but whether it produces scientifically meaningful results that align with established findings in the literature.

5 Discussion of Results

5.1 Results for Literature Review Test

Table 3 presents citation validation results for 20 accounting research topics, split into 10 classic topics covering established research areas and 10 modern topics covering emerging fields. We prompt each model to provide up to 10 citations per topic in structured JSON format, constraining responses to papers published in five leading accounting journals. All models use a generation temperature of 0.7 to reflect realistic research interactions. Standalone LLMs—GPT-5.2, Claude, and Gemini—generate citations from pre-trained knowledge alone. Etalia-augmented models receive up to 20 semantically retrieved papers from

the corpus as context before generation, and their prompts instruct them to cite only the retrieved papers.

The table reports four metrics. *Generated* counts the total citations produced across all topics in each category. *Verified* counts the citations confirmed to exist in the Crossref database, the authoritative registry of scholarly DOIs. We query Crossref with each citation's title and first author, computing a fuzzy title similarity score using the token set ratio algorithm from the rapidfuzz library. A citation is classified as verified if the score meets or exceeds 75%; otherwise it is classified as hallucinated. When the initial query filtered by publication year returns no match, we retry without the year constraint to accommodate minor year discrepancies in the LLM output. *Hallucination (%)* equals (Generated − Verified) / Generated × 100, the fraction of citations that do not correspond to real papers. *Relevance (%)* measures the mean cosine similarity between the OpenAI text-embedding-3-small embedding of the topic query and the embedding of each citation title. We compute this metric independently of Etalia's retrieval system to avoid circular dependence between the search engine that provides context and the evaluation that measures output quality.

For classic topics, standalone LLMs generate near the maximum of 100 citations (10 per topic) and perform reasonably well. GPT-5.2 achieves 5.1% hallucination with 93 of 98 citations verified in Crossref. Claude achieves 9.0% hallucination with 91 of 100 verified, and Gemini achieves 11.0% with 89 of 100 verified. These rates suggest that the models' training data contain substantial coverage of established accounting research, though non-trivial fractions of citations are still fabricated. All three Etalia-augmented models eliminate hallucinations entirely on classic topics: Etalia+GPT verifies all 77 generated citations, and Etalia+Claude and Etalia+Gemini each verify all 81. The lower citation counts for augmented models reflect that Etalia's retrieved context sometimes contains fewer than 10 papers per topic, and the LLMs faithfully cite only the retrieved sources rather than supplementing with pre-trained knowledge. Semantic relevance scores are comparable across all six configurations for classic topics, ranging from 55.3% to 58.3%, indicating that both

standalone and augmented models cite topically appropriate literature when the research area is well-established.

The distinction between standalone and augmented models sharpens considerably on modern topics. Gemini's hallucination rate rises from 11.0% on classic topics to 36.7% on modern topics, fabricating more than one in three references on subjects such as cryptocurrency accounting and cybersecurity disclosure. Claude's rate remains similar, declining slightly from 9.0% to 8.3%. GPT-5.2 generates only 14 citations across all 10 modern topics (versus 98 on classic topics) with 0.0% hallucination, largely declining to produce references in areas in which its training data are sparse. All three Etalia-augmented models maintain 0.0% hallucination on modern topics while producing 42 to 45 verified citations per model, demonstrating that retrieval augmentation enables reliable citation generation regardless of how established the research area is.

Relevance scores reinforce this pattern. GPT-5.2's mean relevance drops to 13.7% on modern topics, reflecting its limited and tangentially related output. Standalone Claude and Gemini achieve moderate relevance of 41.9% and 48.9%, respectively, while all three Etalia-augmented models maintain scores between 53.5% and 53.7%—comparable to their classic-topic performance and substantially higher than any standalone model on modern topics.

Panel B of Table 3 reports the distributional properties of hallucination rates across 10 independent runs of the full citation collection and validation pipeline, split by topic category. Each run re-queries all six model configurations on all 20 topics at temperature 0.7 and re-validates every generated citation against Crossref. Because the Crossref cache retains prior lookups, the validation step is deterministic for previously encountered citations; variation across runs arises entirely from the stochastic nature of LLM generation. On classic topics, GPT-5.2 averages a 6.49% hallucination rate (SE = 1.27), Claude averages 6.86% (SE = 0.88), and Gemini averages 9.00% (SE = 0.63). On modern topics, hallucination rates diverge sharply: Gemini averages 35.23% (SE = 1.44), nearly quadrupling its classic-topic

rate; Claude averages 10.35% (SE = 1.54); and GPT-5.2 averages 6.30% (SE = 1.90). All three Etalia-augmented models achieve exactly 0.00% hallucination across all 10 runs with zero variance in both topic categories, confirming that the retrieval-augmented architecture deterministically eliminates fabricated citations. Taken together, the results in Table 3 provide evidence for our first research question: domain-specific RAG eliminates citation hallucinations while preserving—and in emerging fields, substantially improving—the topical relevance of LLM-generated literature reviews.

5.2 Results for the Replication Test

Tables 4 and 5 report replication results for two seminal papers in empirical accounting. Panel A of each table reports code generation metrics; Panel B compares estimated regression coefficients against published values.

Table 4 reports results for Sloan (1996) Table 3, Panel A, which regresses earnings on the accrual and cash flow components. All six model configurations successfully generate final working code. Among baseline models, all three achieve first-iteration success: GPT-5.2 completes in 22.4 seconds, Claude in 33.0 seconds, and Gemini in 44.4 seconds. Among Etalia-augmented models, Etalia+Claude and Etalia+Gemini also succeed on the first iteration, while Etalia+GPT requires two iterations. The augmented models take longer per iteration (74.7–135.9 seconds versus 22.4–44.4 seconds for baselines), reflecting the additional time required for retrieval and context injection.

Panel B compares estimated coefficients against Sloan's published values ($\hat{\alpha} = 0.011$, $\hat{\beta}_1(\text{ACC}) = 0.765$, $\hat{\beta}_2(\text{CFO}) = 0.855$; $N = 40{,}679$; 1962–1991). Results vary substantially across baseline models. GPT-5.2 produces coefficients of 0.686 and 1.051 for the accrual and cash flow components, respectively, but uses a sample of 261,990 observations over 1988–2024—far outside the original sample period. Claude generates near-zero coefficients ($\hat{\beta}_1 = -0.000$, $\hat{\beta}_2 = 0.001$) with an R^2 of 0.000, suggesting a fundamental error in variable construction despite using a closer sample period (1963–1999). Gemini produces a strong

cash flow coefficient (0.854) but overestimates the accrual coefficient (1.111) using 195,361 observations over 1985–2024.

Etalia-augmented models more accurately match the original sample period and produce more consistent coefficient estimates. Both Etalia+GPT and Etalia+Claude closely approximate the 1963–1990 sample window, and their accrual coefficients (0.657 and 0.655, respectively) and cash flow coefficients (0.731 and 0.721) are closer to Sloan's published values than most baseline estimates. Etalia+Gemini extends the sample through 2019 but produces coefficients ($\hat{\beta}_1 = 0.849$, $\hat{\beta}_2 = 0.861$) that closely approximate the original estimates. Sample sizes across all models exceed Sloan's 40,679, ranging from 107,562 to 293,558, likely reflecting differences in data availability and filtering criteria in modern WRDS data. These results indicate that RAG-provided context—particularly the original paper's variable definitions and sample restrictions—materially improves the accuracy of LLM-generated replication code, addressing the second research question.

Table 5 reports results for Basu (1997) Table 1, Panel A, the reverse earnings-return regression $X_t/P_{t-1} = \beta_0 + \beta_1 D_t + \beta_2 R_t + \beta_3 D_t \times R_t + \varepsilon_t$. Code generation is more challenging for this specification: baseline GPT-5.2 requires three iterations, Gemini requires two, and only Claude achieves first-iteration success. Among Etalia-augmented models, Etalia+Gemini succeeds on the first iteration, while Etalia+GPT and Etalia+Claude each require two.

Panel B compares estimated coefficients against Basu's published values ($\hat{\beta}_0 = 0.090$, $\hat{\beta}_1 = 0.020$, $\hat{\beta}_2 = 0.059$, $\hat{\beta}_3 = 0.216$; $N = 43{,}320$; 1963–1990). Baseline models struggle with this specification. GPT-5.2 produces a $\hat{\beta}_3$ of 4.353 with a negative intercept (-0.264), suggesting a scaling error. Claude omits the return coefficient ($\hat{\beta}_2$) entirely, indicating an incomplete specification. Gemini produces a reasonable intercept (0.060) but overestimates the interaction coefficient ($\hat{\beta}_3 = 0.433$).

Etalia-augmented models produce estimates closer to the published values. Etalia+Claude achieves the closest match, with $\hat{\beta}_0 = 0.090$ (identical to Basu's estimate), $\hat{\beta}_1 = 0.018$, and $\hat{\beta}_3 = 0.437$, using 48,704 observations over 1963–1990—closely matching the original sam-

ple. Etalia+GPT and Etalia+Gemini produce reasonable intercepts (0.053 and 0.056) and positive or near-zero $\hat{\beta}_1$ coefficients, though their sample periods extend beyond the original (1950–2024 and 1981–2022, respectively). Combined with the Sloan (1996) results, the Basu (1997) replication reinforces our answer to the second research question: domain-specific retrieval not only helps LLMs generate working code but also guides them toward the correct variable constructions, sample windows, and regression specifications needed to approximate published findings.

6 Conclusion

Recent advances in generative AI have changed economics, finance, and accounting research. LLMs automate substantial portions of the process and produce academic papers quickly (e.g., Ciconte et al., 2025; Novy-Marx and Velikov, 2025). Although these tools may expand scholarly capacity, they also raise concerns about research integrity, replicability, and the resilience of peer-review infrastructure (Eisfeldt and Schubert, 2025; Korinek, 2023; de Kok, 2025). These risks are particularly consequential in empirical economics, finance, and accounting, in which credibility depends on transparent links between prior literature, hypothesis development, and empirical execution. Retrieval-augmented generation (RAG) systems offer a potential solution by grounding model outputs in verified domain knowledge and reducing citation hallucinations.

We introduce Etalia, a domain-specific RAG system built from 8,075 papers published in leading accounting and finance journals, 376 replication packages, and research documentation. We show that Etalia eliminates citation hallucinations—reducing rates from as high as 36.7% for standalone LLMs to 0.0% across all Etalia-augmented configurations, a result that holds with zero variance across 10 independent runs. Etalia also improves the accuracy of LLM-generated code for empirical replications of Sloan (1996) and Basu (1997): augmented models more accurately identify original sample periods and produce coefficient

estimates closer to published values. Beyond replication, the system enables researchers to trace hypothesis development by evaluating both included and omitted literature, strengthening theoretical coherence as AI lowers the cost of producing statistically significant but weakly motivated findings (Ciconte et al., 2025; Novy-Marx and Velikov, 2025). Our study contributes to the literatures on generative AI and research replication by demonstrating that domain-specific architectures enhance reliability while lowering the costs of systematic verification (Korinek, 2023; Eisfeldt and Schubert, 2025; Hail et al., 2020; Jensen et al., 2023).

References

Ankel-Peters, J., Fiala, N., and Neubauer, F. (2023). Do economists replicate? *Journal of Economic Behavior & Organization*, 212:219–232.

Ashbaugh, H., LaFond, R., and Mayhew, B. W. (2003). Do nonaudit services compromise auditor independence? further evidence. *The accounting review*, 78(3):611–639.

Bartik, A., Gupta, A., and Milo, D. (2025). The costs of housing regulation: Evidence from generative regulatory measurement. *Working Paper*.

Basu, S. (1997). The conservatism principle and the asymmetric timeliness of earnings. *Journal of Accounting and Economics*, 24(1):3–37.

Breitung, C. and Müller, S. (2025). Global business networks. *Journal of Financial Economics*, 166:104007.

Breuer, W. and Hofmann, C. (2025). Replication studies in finance and accounting. *Journal of Business Economics*, 95(2):187–196.

Brodeur, A., Sung, S. Y., Miguel, E., Vilhuber, L., and de la Guardia, F. H. (2025). Assessing reproducibility in economics using standardized crowd-sourced analysis. *Working Paper*.

Bryzgalova, S., Huang, J., and Julliard, C. (2023). Bayesian solutions for the factor zoo: We just ran two quadrillion models. *The Journal of Finance*, 78(1):487–557.

Cakici, N., Fieberg, C., Metko, D., and Zaremba, A. (2024). Do anomalies really predict market returns? New data and new evidence. *Review of Finance*, 28(1):1–44.

Chen, A. Y. and Velikov, M. (2023). Zeroing in on the expected returns of anomalies. *Journal of Financial and Quantitative Analysis*, 58(3):968–1004.

Chen, A. Y. and Zimmermann, T. (2022). Open source cross-sectional asset pricing. *Critical Finance Review*, 27(2):207–264.

Chordia, T., Goyal, A., and Saretto, A. (2020). Anomalies and false rejections. *The Review of Financial Studies*, 33(5):2134–2179.

Ciconte, W., Rozario, A., and Urcan, O. (2025). Using AI to identify exogenous shocks and conduct archival accounting research. *Working Paper*.

de Kok, T. (2025). ChatGPT for textual analysis? How to use generative LLMs in accounting research. *Management Science*.

Eisfeldt, A. L. and Schubert, G. (2025). Generative AI and Finance. *Annual Review of Financial Economics*, 17.

Frankel, R. M., Johnson, M. F., and Nelson, K. K. (2002). The relation between auditors' fees for nonaudit services and earnings management. *The accounting review*, 77(4):71–105.

Hail, L., Lang, M., and Leuz, C. (2020). Reproducibility in accounting research: Views of the research community. *Journal of Accounting Research*, 58(2):519–543.

Harvey, C. R., Liu, Y., and Zhu, H. (2016). ... and the cross-section of expected returns. *The Review of Financial Studies*, 29(1):5–68.

Heath, D., Ringgenberg, M. C., Samadi, M., and Werner, I. M. (2023). Reusing natural experiments. *The Journal of Finance*, 78(4):2329–2364.

Hou, K., Xue, C., and Zhang, L. (2020). Replicating anomalies. *Review of Financial Studies*, 33(5):2019–2133.

Huang, A. H., Wang, H., and Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Jensen, T. I., Kelly, B., and Pedersen, L. H. (2023). Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518.

Joëts, M. and Mignon, V. (2026). Slaying the undead: How long does it take to kill zombie papers? *Research Policy*, 55(2):105401.

Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317.

Li, K., Mai, F., Shen, R., Yang, C., and Zhang, T. (2026). Dissecting corporate culture using generative AI. *The Review of Financial Studies*, 39(1):253–296.

Ludwig, J. and Mullainathan, S. (2024). Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2):751–827.

Ludwig, J., Mullainathan, S., and Rambachan, A. (2025). Large language models: An applied econometric framework. *Working Paper*.

Luo, Y., Salterio, S. E., and Adamson, C. (2024). Replication of audit and financial accounting research: We do more than we think. *Accounting Horizons*, 38(4):173–189.

Maniadis, Z., Tufano, F., and List, J. A. (2017). To replicate or not to replicate? Exploring reproducibility in economics through the lens of a model and a pilot study. *The Economic Journal*.

McLean, R. D. and Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance*, 71(1):5–32.

Novy-Marx, R. and Velikov, M. Z. (2025). Ai-powered (finance) scholarship. *Working Paper*.

Salterio, S. E. (2014). We don't replicate accounting research—or do we? *Contemporary Accounting Research*, 31:1134–1142.

Shahidi, P., Rusak, G., Manning, B. S., Fradkin, A., and Horton, J. J. (2025). The coasean singularity? Demand, supply, and market design with AI agents. *Working Paper*.

Sharma, D. S. (2023). To replicate or not to replicate? that is the question. *Journal of Accounting and Public Policy*, 42(6):107151.

Sloan, R. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings? *The Accounting review*, 71(3):289–315.

Table 1: Etalia Corpus Statistics

Panel A: Papers by Journal

Journal	Papers
RFS	2,349
TAR	2,136
JAE	1,624
JAR	1,060
RAS	906
Total	**8,075**

Panel B: Code and Documentation

Category	Source	Count
Replication Packages	RFS	272
Replication Packages	JAR	104
Code	WRDS Sample Programs	93
Documentation	WRDS Manuals	143
Total		**612**

Table 2: Data Extraction Methods

Panel A: Content Extraction Methods

Content Type	Method	Output
Paper Sections (Direct Extraction)		
Title, Abstract	PDF parsing	Full text
Introduction	PDF parsing	Full text
Literature Review	PDF parsing	Full text
Hypothesis Development	PDF parsing	Full text
Research Design	PDF parsing	Full text
Results	PDF parsing	Full text
Conclusion	PDF parsing	Full text
Structured Data (LLM Extraction via Llama 3.3 70B)		
Summary Statistics	LLM	Variable, N, Mean, Std, Min, Max
Regression Tables	LLM	DV, IVs, coefficients, t-stats
Correlation Matrices	LLM	Variable pairs, correlations
Code		
Replication packages	Direct	Full code files
WRDS examples	Direct	Code snippets

Panel B: LLM Extraction Tasks

LLM Task	Description
Section Segmentation	Identify section boundaries using font size and headers
Summary Statistics Parsing	Extract variable names, N, mean, std dev from tables
Hypothesis Extraction	Identify H1, H2, H3 and their predicted relationships
Code Documentation	Parse inline comments and docstrings

Notes: Panel A describes the content types and extraction methods. Panel B lists the LLM tasks used for structured data extraction.

Table 3: Citation Validation in LLM-Generated Literature Reviews

Panel A: Citation Quality Metrics by Model

Model	Generated	Verified	Hallucination (%)	Relevance (%)
Classic Topics (10 Established Research Areas)				
GPT-5.2	98	93	5.1	56.3
Claude	100	91	9.0	55.3
Gemini	100	89	11.0	58.3
Etalia + GPT	77	77	0.0	55.3
Etalia + Claude	81	81	0.0	55.4
Etalia + Gemini	81	81	0.0	55.3
Modern Topics (10 Emerging Research Areas)				
GPT-5.2	14	14	0.0	13.7
Claude	72	66	8.3	41.9
Gemini	79	50	36.7	48.9
Etalia + GPT	45	45	0.0	53.7
Etalia + Claude	45	45	0.0	53.5
Etalia + Gemini	42	42	0.0	53.5

Notes: Panel A reports citation validation results for six LLM configurations across 20 accounting research topics. Table A.1 in the Appendix lists the topics used to prompt each model, organized by category. Each model receives a prompt requesting up to 10 citations per topic in JSON format at temperature 0.7, constrained to five leading accounting journals (TAR, JAR, JAE, RAS, CAR). Standalone LLMs generate citations from pre-trained knowledge; Etalia-augmented models receive up to 20 semantically retrieved papers as context and are instructed to cite only from the retrieved set. Generated = total citations produced across all topics in each category. Verified = citations confirmed in the Crossref database via fuzzy title matching (token set ratio $\geq 75\%$). Hallucination (%) = (Generated − Verified) / Generated × 100. Relevance (%) = mean cosine similarity between OpenAI text-embedding-3-small embeddings of the topic query and each citation title, computed independently of Etalia.

Table 3: Citation Validation in LLM-Generated Literature Reviews (continued)

Panel B: Hallucination Rate Stability (10 Independent Runs)

Model	Generated	Verified	Hallucination (%)
Classic Topics (10 Established Research Areas)			
GPT-5.2	87.3	81.0	6.49
	(9.7)	(9.1)	(1.27)
Claude	89.0	82.2	6.86
	(9.9)	(9.2)	(0.88)
Gemini	100.0	91.0	9.00
	(0.0)	(0.6)	(0.63)
Etalia + GPT	79.0	79.0	0.00
	(0.7)	(0.7)	(0.00)
Etalia + Claude	77.5	77.5	0.00
	(1.3)	(1.3)	(0.00)
Etalia + Gemini	84.7	84.7	0.00
	(5.6)	(5.6)	(0.00)
Modern Topics (10 Emerging Research Areas)			
GPT-5.2	18.8	17.5	6.30
	(2.9)	(2.7)	(1.90)
Claude	60.5	53.4	10.35
	(7.2)	(6.2)	(1.54)
Gemini	80.8	52.3	35.23
	(1.5)	(1.4)	(1.44)
Etalia + GPT	45.0	45.0	0.00
	(0.0)	(0.0)	(0.00)
Etalia + Claude	44.6	44.6	0.00
	(0.3)	(0.3)	(0.00)
Etalia + Gemini	58.3	58.3	0.00
	(10.4)	(10.4)	(0.00)

Notes: Panel B reports mean values across 10 independent runs of the full citation collection and validation pipeline, with standard errors in parentheses, split by topic category. Each run re-queries all six models at temperature 0.7 and re-validates every citation against Crossref. Gemini's hallucination rate nearly quadruples on modern topics (35.23% vs. 9.00%), while all Etalia-augmented models maintain 0% hallucination across both categories.

Table 4: Sloan (1996) Replication

Panel A: Code Generation Quality

Model	Final Code Generated	Iterations	Time/Iter (sec)	Total Time (sec)
GPT-5.2	1	1	22.4	22.4
Claude 4.5	1	1	33.0	33.0
Gemini 3	1	1	44.4	44.4
Etalia + GPT-5.2	1	2	74.7	149.4
Etalia + Claude 4.5	1	1	121.1	121.1
Etalia + Gemini 3	1	1	135.9	135.9

Panel B: Coefficient Comparison

Model	α	β_1 (ACC)	β_2 (CFO)	R^2	N	Sample Years
Richard G. Sloan (1996)	0.011	0.765	0.855	–	40,679	1962–1991
	(24.05)	(186.53)	(304.56)			
GPT-5.2	-0.049	0.686	1.051	0.616	261,990	1988–2024
	(-50.45)	(67.51)	(122.27)			
Claude 4.5	-0.259	-0.000	0.001	0.000	175,533	1963–1999
	(-3.30)	(-0.07)	(0.35)			
Gemini 3	-0.042	1.111	0.854	0.591	195,361	1985–2024
	(-23.72)	(142.83)	(465.80)			
Etalia + GPT-5.2	-0.002	0.657	0.731	0.484	107,562	1963–1990
	(-4.21)	(92.96)	(121.38)			
Etalia + Claude 4.5	-0.000	0.655	0.721	0.480	121,360	1963–1990
	(-0.57)	(216.82)	(326.14)			
Etalia + Gemini 3	-0.029	0.849	0.861	0.613	293,558	1963–2019
	(-37.41)	(185.19)	(645.06)			

Notes: This table reports single-run replication results for Sloan (1996) Table 3, Panel A. Panel A reports code generation metrics. Panel B reports estimated coefficients with t-statistics in parentheses. The benchmark row shows original values from Sloan (1996). All models receive an identical minimal prompt with no methodology hints. Etalia models use two-stage RAG retrieval from a corpus of accounting research papers. Temperature = 0.7 for all models.

Table 5: Basu (1997) Replication

Panel A: Code Generation Quality

Model	Final Code Generated	Iterations	Time/Iter (sec)	Total Time (sec)
GPT-5.2	1	3	50.5	151.5
Claude 4.5	1	1	53.7	53.7
Gemini 3	1	2	63.3	126.6
Etalia + GPT-5.2	1	2	180.8	361.6
Etalia + Claude 4.5	1	2	100.1	200.1
Etalia + Gemini 3	1	1	232.1	232.1

Panel B: Coefficient Comparison

Model	β_0	β_1 (D)	β_2 (R)	β_3 (R×D)	R^2	N	Sample Years
Sudipta Basu (1997)	0.090	0.020	0.059	0.216	0.101	43,320	1963–1990
	(68.03)	(0.86)	(18.34)	(20.66)			
GPT-5.2	-0.264	-0.308	-0.300	4.353	0.131	122,399	1963–1995
	(-1.03)	(-9.12)	(-8.69)	(38.19)			
Claude 4.5	0.136	-0.160	–	-0.000	0.089	92,694	1964–1990
	(62.98)	(-29.16)		(-3.21)			
Gemini 3	0.060	-0.004	-0.021	0.433	0.123	251,055	1970–2022
	(69.77)	(-2.40)	(-18.06)	(132.30)			
Etalia + GPT-5.2	0.053	0.018	-0.021	0.509	0.128	250,186	1950–2024
	(64.13)	(10.59)	(-13.37)	(91.42)			
Etalia + Claude 4.5	0.090	0.018	-0.002	0.437	0.142	48,704	1963–1990
	(71.73)	(7.00)	(-0.59)	(41.42)			
Etalia + Gemini 3	0.056	-0.008	-0.037	0.341	0.098	204,522	1981–2022
	(76.38)	(-6.02)	(-23.69)	(86.99)			

Notes: This table reports single-run replication results for Basu (1997) Table 1, Panel A. Panel A reports code generation metrics. Panel B reports estimated coefficients with t-statistics in parentheses. The benchmark row shows original values from Basu (1997). The Basu regression is $X_t/P_{t-1} = \beta_0 + \beta_1 D_t + \beta_2 R_t + \beta_3 D_t \times R_t + \varepsilon_t$, where X_t is earnings, R_t is annual return, and $D_t = 1$ if $R_t < 0$. All models receive an identical minimal prompt with no methodology hints. Etalia models use two-stage RAG retrieval from a corpus of accounting research papers. Temperature = 0.7 for all models.

A Research Topics for Citation Validation

Table A.1: Research Topics (20 Queries)

Panel A: Classic Topics (10 Established Research Areas)

#	Topic
1	Earnings management and accruals quality
2	Audit quality and auditor independence
3	Corporate governance and board characteristics
4	Tax avoidance and corporate tax planning
5	Internal controls and financial reporting quality
6	Voluntary disclosure and information asymmetry
7	Analyst forecasts and earnings announcements
8	Cost of capital and disclosure
9	Executive compensation and firm performance
10	Accounting conservatism and debt contracting

Panel B: Modern Topics (10 Emerging Research Areas)

#	Topic
1	Machine learning and AI in accounting
2	Workforce diversity
3	ESG disclosure
4	Social media disclosure
5	Climate risk disclosure
6	Cryptocurrency and blockchain accounting
7	COVID-19 and financial reporting
8	Cybersecurity disclosure
9	Data analytics in auditing
10	Human capital disclosure

Notes: This table lists the 20 research topics used to prompt each model in the citation validation tests reported in Table 3. Panel A reports classic topics covering established accounting research areas. Panel B reports modern topics covering emerging fields with sparser published literature.

B Prompts Used for Citation Validation (Table 3)

This appendix reproduces the exact prompts used to generate the citation results reported in Table 3. We use two prompt templates: one for standalone LLMs that generate citations from pre-trained knowledge, and one for Etalia-augmented LLMs that receive retrieved papers as context. Both prompts request structured JSON output to facilitate automated parsing and validation. All models receive the same prompt for a given configuration, with only the {topic} placeholder replaced by the research topic string (e.g., "earnings management and accruals quality"). The generation temperature is set to 0.7 for all models.

B.1 Standalone LLM Prompt

The following prompt is sent to GPT-5.2, Claude, and Gemini without any retrieval context. The model must generate citations entirely from its pre-trained knowledge. The prompt constrains the output to five specific accounting journals and requests a maximum of 10 citations in JSON format. The double curly braces in the JSON template (e.g., {{''author''}}) are Python string formatting escapes that resolve to single braces in the actual prompt sent to the model.

```
Provide up to 10 relevant academic citations on the topic: ''\{topic\}''

CRITICAL CONSTRAINTS:
1. MAXIMUM 10 CITATIONS - Do NOT provide more than 10 citations. Stop after 10.
2. Only cite papers from these top 5 accounting journals: The Accounting Review,
    Journal of Accounting Research, Journal of Accounting and Economics, Review of
    Accounting Studies, Contemporary Accounting Research

Return ONLY a valid JSON array with this exact format (no other text):
{''author'': ''Last1, First1, Last2, First2'', ''year'': ''YYYY'', ''title'': ''Full
    Paper Title'', ''journal'': ''Journal Name''\},
```

```
Requirements:
- STOP AT 10 CITATIONS MAXIMUM
- Use real, published academic papers only
- Only from the 5 specified journals above
- Include full author names (not ‘‘et al.’’)
- Use exact paper titles
- Use full journal names
- Return ONLY the JSON array, no explanations or markdown
```

B.2 Etalia-Augmented LLM Prompt (RAG)

The following prompt is sent to GPT-5.2, Claude, and Gemini via the Etalia API's LLM wrapper endpoint. Before the prompt reaches the language model, the Etalia system performs a semantic search over its corpus of 794,175 embedded paper sections, retrieves the top 20 most relevant results (with a minimum similarity score of 0.3), and injects them into the {context} placeholder. The {query} placeholder is replaced by the same topic string used in the standalone prompt. The critical difference from the standalone prompt is the instruction to cite only papers from the retrieved set, which grounds the LLM's output in verified literature and is the mechanism by which RAG eliminates hallucinations.

```
Provide up to 10 relevant academic citations on the topic: ‘‘\{query\}’’

CRITICAL CONSTRAINTS:
1. MAXIMUM 10 CITATIONS - Do NOT provide more than 10 citations. Stop after 10.
2. Use ONLY the retrieved papers from the Etalia database shown below.

Retrieved papers from Etalia:
\{context\}

Return ONLY a valid JSON array with this exact format (no other text):{‘‘author’’: ‘‘
    Last1, First1, Last2, First2’’, ‘‘year’’: ‘‘YYYY’’, ‘‘title’’: ‘‘Full Paper Title
```

```
    '', ''journal'': ''Journal Name''\},


Requirements:
- STOP AT 10 CITATIONS MAXIMUM
- Use ONLY the papers retrieved from the Etalia database shown above
- Include full author names (not ''et al.'')
- Use exact paper titles as shown in the retrieved sources
- Return ONLY the JSON array, no explanations or markdown
```

B.3 Prompt Design Rationale

Several design choices in these prompts merit discussion. We request JSON output rather than free-text citations to enable automated parsing and validation at scale. The JSON schema specifies four fields—author, year, title, and journal—which provide sufficient metadata to query the Crossref API for verification. We instruct models to use full author names rather than "et al." abbreviations because first-author last names serve as a secondary search key in the Crossref query, improving match accuracy.

The constraint to five specific journals serves two purposes. It narrows the space of valid citations to a well-defined set, making hallucination detection more precise, and it reflects a realistic use case in which a researcher seeks citations from top field journals for a literature review. We cap citations at 10 per topic to keep the validation workload manageable while providing enough citations to measure hallucination rates with reasonable precision. At 10 citations across 10 topics per category, each model produces up to 100 candidate citations per topic category, yielding sufficient statistical power to detect meaningful differences in hallucination rates across model configurations.

The temperature of 0.7 reflects the default setting in most LLM interfaces and represents the conditions under which a typical researcher would interact with these models. Lower temperatures (e.g., 0.1) would reduce output variability but would not reflect realistic usage

patterns.

C Prompt for Replication Studies

```
Generate Python code to replicate the main regression result in {BENCHMARK_PAPER['
    authors']} ({BENCHMARK_PAPER['year']}), {BENCHMARK_PAPER['journal']}.

Provide complete Python code only.
```

Build a prompt asking the LLM to fix an error in the code.

```
The following Python code for replicating the main regression result in {
    BENCHMARK_PAPER['authors']} ({BENCHMARK_PAPER['year']}) has an error.
Please fix the error and return the complete corrected code.

ERROR:
{error_msg}

ORIGINAL CODE:
```python
{original_code}
```

Please provide the complete corrected Python code. Only output the code, no
    explanations.
```