

Equity Signals and Mortgage-Backed Security Returns: An AI-Agent-Assisted Systematic Meta-Analysis

Julian Atanasov

University of Nebraska-Lincoln
julian@unl.edu

Zhipeng Jiang

New Jersey Institute of Technology
zj28@njit.edu

Yao Ma

University of Nebraska-Lincoln
yma27@huskers.unl.edu

Yanan Zhu

New Jersey Institute of Technology
yz872@njit.edu

Authors are listed alphabetically. Corresponding author: Yao Ma.

This version: September 2026

Abstract

Published estimates of how equity market signals relate to mortgage-backed security outcomes differ in sign across otherwise similar specifications. We assemble these scattered coefficients through an AI-agent-assisted systematic review in which the agent performs retrieval, screening, and extraction while a finance domain expert holds exclusive authority over every coefficient and classification judgment. Screening yields ten confirmed estimates from two studies. The apparent disagreement resolves into a moderator: the benchmark asset against which the MBS convenience premium is measured determines the sign of the volatility effect. Within one baseline regression, a one-point increase in VIX raises the premium measured against AAA corporate bonds by 3.207 basis points and lowers the premium measured against Treasury securities by 1.257, both significant at the 1% level. Because estimates measured against different benchmarks belong to different strata, we report no pooled effect size. We have identified no comparable synthesis of this literature.

AI Disclosure Statement: This paper was produced with substantial use of generative AI. A single orchestrated agent built on Anthropic's Claude model family performed literature retrieval, screening, coefficient extraction, classification proposals, state-file maintenance, and the majority of manuscript drafting. OpenAI's ChatGPT was used outside the extraction pipeline for prompt generation and refinement, critique, methodological review, and drafting support. No model was trained or fine-tuned. Every coefficient, classification, and scope decision was made by a human author and logged before entering the analysis, and the authors reviewed all content and take full responsibility for the paper. The workflow and the human and AI contributions are documented in Section 3 and Appendix D.

Keywords: mortgage-backed securities; convenience premium; equity volatility; VIX; systematic meta-analysis; AI-agent workflows; research automation; human-in-the-loop

JEL Classification: G12, G14, G21, C18

1. Introduction

The agency mortgage-backed security market is one of the largest fixed income markets in the world, and its pricing is understood to depend on prepayment risk, interest rate dynamics, government guarantees, and the demand for safe assets. A recurring empirical question is whether signals originating in equity markets carry information about MBS outcomes. The question matters for three reasons. It bears on the transmission of risk across asset classes; it bears on whether intermediary balance sheet conditions, which are observable through equity prices, propagate into mortgage credit; and it bears on the practical problem of hedging and forecasting in a market where the dominant risk factor is not equity risk.

The literature has approached this question many times, usually indirectly. Equity market variables appear in MBS pricing regressions primarily as controls rather than as the object of interest, which means the relevant coefficients are scattered across tables in papers whose stated contributions lie elsewhere. When these coefficients are collected, they do not agree. Reported effects of equity volatility on MBS spreads are positive in some specifications and negative in others, sometimes within the same paper. A reader encountering this pattern has three available interpretations: the relation is unstable, the estimates are noisy, or the specifications differ in some respect that has not been made explicit. Distinguishing among these is precisely the task a systematic meta-analysis is designed to perform, and we have identified no such synthesis of this literature.

This paper conducts that synthesis. Our finding is that the third interpretation is correct and that the relevant difference is identifiable and economically interpretable. The apparent instability is an artifact of aggregating estimates that measure the MBS premium against different benchmark assets. Once benchmark asset is coded as a moderator, the estimates are not in conflict; they are two halves of a single coherent pattern that follows from the structure of safe-asset demand.

The evidence is clearest in the baseline specification of He and Song (2026), where both coefficients are estimated in the same table, on the same monthly sample from October 2003 through December 2022, with the same controls. A one-point increase in VIX is associated with a 3.207 basis point increase in the convenience premium of newly issued Fannie Mae 30-year MBS measured against AAA corporate bonds, and with a 1.257 basis point decrease in the premium measured against Treasury securities. Both estimates are significant at the 1% level with Newey-West standard errors. The specifications differ only in the benchmark asset. The economic reading is that a flight to safety raises demand for all safe assets, so agency MBS gain relative to corporate credit, but Treasury securities gain more, so the MBS-Treasury spread compresses. This is the comparative static implied by the safe-asset demand framework the source paper

itself develops, in which a common demand shock raises the MBS convenience premium while lowering the Treasury-MBS yield spread.

The pattern is not confined to one table. When the short-term federal funds rate replaces the mortgage rate as the primary independent variable, the coefficient on the AAA-benchmarked spread remains positive and significant. In a structurally different panel of seasoned MBS coupon cohorts, estimated with coupon fixed effects and standard errors clustered by month, the contrast appears again in both directions and at the 1% level. The recurrence of the same sign structure across a change in the primary independent variable, a change in market segment, and a change in estimation framework is the strongest evidence available in our current sample that the moderator is real rather than incidental to one specification.

Alongside the benchmark result, our coding surfaces three additional distinctions that bear on any subsequent synthesis. First, equity signals operate at different levels: broad market returns, intermediary equity factors, and firm-level equity variables are not interchangeable inputs, and estimates on them should not be pooled without stratification. Diep, Eisfeldt, and Richardson (2021) report a coefficient of 3.61% on the raw monthly S&P 500 return and 2.80% on an intermediary asset pricing factor for Treasury-hedged MBS excess returns at the par coupon. These are similar in magnitude but they are answers to different questions. Second, the same equity signal can predict aggregate MBS return levels without explaining cross-sectional variation across coupons; the same paper reports a cross-sectional estimate of -0.17% with a t-statistic of -0.21. Reporting only the significant aggregate result would misstate what the literature knows. Third, supply-side outcomes such as issuance quantity are measured in different units from performance outcomes and answer a different economic question, so they belong in a separate panel.

We reach these conclusions with ten confirmed estimates drawn from two studies, which is well below the threshold at which we pre-specified that pooled effect sizes would be estimated. We therefore report no summary effect and make no claim about the population magnitude of the equity-MBS relation. What the current evidence supports is a claim about structure rather than magnitude: it identifies the coding decisions that determine whether a future pooled estimate is meaningful or meaningless. We regard this as the appropriate output of an interim synthesis and we have written the paper so that additional estimates can be incorporated without reorganizing it.

1.1 The workflow contribution

The second contribution concerns how the synthesis was produced. The paper was assembled using an AI agent operating a prompt-orchestrated pipeline that performs literature retrieval, screening, coefficient extraction, classification, and drafting, with all state maintained in external files rather than in model memory. Auxiliary AI assistance was used outside this pipeline for prompt refinement, critique, and

drafting support; Section 3.1 draws the boundary and Appendix D documents both. The design principle is that the system has no authority over finance judgments. Every coefficient value, every classification of a study into an inclusion category, and every scope decision is escalated to a human domain expert and enters the database only after an explicit ruling. The agent proposes; the expert disposes; the escalation is logged.

This division is not a safety disclaimer appended to an otherwise autonomous system. It is load-bearing, and we can demonstrate that it is load-bearing because it failed once and caught the failure. During extraction, a coefficient was recorded with the correct numerical value but the wrong dependent variable, because the verification question posed to the expert named a table column without stating what that column measured. The expert answered accurately for the column he was asked about. The error therefore passed through a correct human ruling and was detected only when the published source was re-examined directly, at which point the recorded value was found to describe an issuance quantity in billions of dollars rather than a yield spread in basis points. The correction changed the sign structure of our central result and reversed an earlier conclusion about statistical significance.

We report this in detail because the AFA session that motivates this paper asks what AI systems can and cannot do in a research setting, and the answer is more informative when it includes a failure with a diagnosed cause. The cause was not a misread number and not an incorrect expert ruling. It was an underspecified interface between the agent and the human, in which the agent asked a question whose answer could not reveal the error it contained. The protocol change that followed is narrow and mechanical: every verification query must state table, column, and dependent variable together, so that a mismatch surfaces in the question rather than propagating through the answer. Section 3 describes the pipeline and Section 8.3 returns to what this episode implies about where human effort is best spent in agent-assisted research.

1.2 Relation to the closest prior work

Our methodological template is Fernandes, Lynch, and Netemeyer (2014), who synthesize 168 papers covering 201 studies on financial literacy and downstream financial behavior and show that the headline relation depends systematically on study methodology rather than being a stable population parameter. We follow their strategy of treating methodological features as coded moderators rather than as nuisance heterogeneity, and their practice of reporting what the evidence does not support alongside what it does. Our sample is far smaller and our claims correspondingly narrower.

Relative to the MBS pricing literature, we do not produce new estimates and we do not dispute any published result. Every coefficient we analyze is taken from its published source at its published value. Our contribution is that we have identified no prior assembly of these coefficients, and that assembling them

reveals a moderator structure invisible within any single paper. Relative to the emerging literature on AI in research workflows, our contribution is a documented design in which the locus of human authority is specified in advance and its operation, including one failure, is recorded.

The remainder of the paper proceeds as follows. Section 2 positions the paper in three literatures. Section 3 describes the agent pipeline and the human escalation protocol. Section 4 describes the corpus, screening, and coding scheme. Section 5 specifies the meta-analytic framework and states why no pooled estimate is reported. Section 6 presents the ten estimates. Section 7 discusses mechanisms. Section 8 reports sensitivity analysis and workflow diagnostics. Section 9 discusses limitations and Section 10 concludes.

2. Related Literature

Three literatures bear on this paper. The first concerns the pricing of agency MBS and the role of prepayment risk. The second concerns convenience premiums and the demand for safe assets. The third concerns the use of AI systems in the research process. We take each in turn and state where this paper enters.

2.1 MBS pricing and prepayment risk

The modern literature on agency MBS pricing takes prepayment as the central risk. Gabaix, Krishnamurthy, and Vigneron (2007) show that prepayment risk is priced in the MBS market and that the pricing reflects the limited risk-bearing capacity of specialized investors rather than economy-wide risk aversion. Boyarchenko, Fuster, and Lucca (2019) decompose mortgage spreads and document how much of the variation is attributable to prepayment compensation. Chernov, Dunn, and Longstaff (2018) estimate a model in which prepayment is driven by macroeconomic factors and recover latent factors governing turnover and rate responsiveness. Diep, Eisfeldt, and Richardson (2021) examine the cross section of MBS returns and show that prepayment risk exposure organizes the cross section, with returns varying systematically across coupons relative to the current coupon. Fusari, Li, Liu, and Song (2022) show that trading organized around coupon cohorts carries its own pricing implications, which bears on our decision in Section 4.3 to record market segment as a coded field.

What this literature establishes is that the dominant priced risk in agency MBS is not equity risk. Equity variables appear in these papers, but as controls or as robustness checks, which is precisely why the coefficients relevant to our question are scattered and have not been collected. Diep, Eisfeldt, and Richardson (2021) report equity-related coefficients in a robustness table rather than in the tables supporting their main claim, and the results we extract from that paper should be read in that light: they are informative about the equity-MBS relation, but they are not the finding the paper was written to establish. We flag this as a general feature of the evidence base rather than a defect in any individual study.

2.2 Convenience premiums and safe-asset demand

A distinct literature studies the premium investors pay for assets that provide safety and liquidity services. The theoretical starting point is Gorton and Pennacchi (1990), who show why claims whose value is insensitive to private information are the ones that come to serve as money-like assets. Krishnamurthy and Vissing-Jorgensen (2012) establish the aggregate demand for Treasury debt and the convenience premium it commands. Sunderam (2014) and Nagel (2016) extend the analysis to private money-like claims and to the dependence of liquidity premiums on the opportunity cost of money, with Nagel (2016) interpreting equity volatility as a proxy for flight-to-safety demand. He and Song (2026) bring agency MBS into this literature, measuring the MBS convenience premium against both Treasury securities and AAA corporate bonds and analyzing its variation with prepayment-related demand, government guarantees, and regulatory treatment.

This second literature supplies the interpretive framework for our central result. In a model where investors derive convenience services from a portfolio of safe assets, a common demand shock raises the convenience premium of every safe asset but by different amounts, ordered by how much liquidity service each provides. Agency MBS provide less liquidity service than Treasury securities, so a rise in aggregate safe-asset demand raises the MBS premium relative to corporate credit while compressing it relative to Treasuries. The opposite-signed coefficients we document are the empirical signature of this asymmetry. What has not previously been noted is that this asymmetry makes benchmark asset a required moderator for any synthesis of the literature, because it guarantees that estimates measured against different benchmarks will carry opposite signs by construction.

2.3 Equity signals and fixed income returns

A broader literature examines whether equity market conditions predict fixed income outcomes outside the MBS market. Kwan (1996) documents a negative contemporaneous relation between individual firms' stock returns and yield changes on bonds issued by the same firms, and finds that stock returns carry predictive content for subsequent bond yield changes. Keim and Stambaugh (1986) show that predetermined variables reflecting the level of asset prices forecast returns on common stocks, long-term bonds of varying default risk, and default-free bonds. Collin-Dufresne, Goldstein, and Martin (2001) examine the determinants of credit spread changes and find that variables suggested by structural models, including equity returns, explain a limited share of the variation. Following an explicit ruling by the project domain expert, we treat this work as background motivation only and do not extract coefficients from it into the main analysis, on the ground that the dependent variables are not MBS outcomes and that the reported signs across these studies are mixed in ways that would not be informative if pooled with MBS estimates. We note the literature here because it motivates the question and because it illustrates a coding boundary that a synthesis must draw explicitly rather than by default.

2.4 AI systems in the research process

The use of large language models and agent systems in academic research has moved quickly from writing assistance toward multi-stage participation in the research process itself. Korinek (2023) surveys the use cases available to economists across ideation, background research, data analysis, coding, and writing, and argues that the productivity gains come principally from automating micro-tasks; his subsequent treatment of agentic systems (Korinek 2025) addresses workflows in which a model orchestrates multi-step tasks rather than responding to isolated prompts, which is the class of system we operate. Experimental evidence on knowledge work supports a productivity effect of meaningful size: Noy and Zhang (2023) find substantial reductions in completion time and improvements in quality on professional writing tasks. The questions the profession now faces concern which stages can be delegated, what documentation makes a delegated process auditable, and how scholarly contribution should be assessed when substantial portions of the execution are automated. Systematic review and meta-analysis are natural early candidates for delegation because the tasks are well defined, high in volume, and low in required creativity: locating candidate studies, screening against criteria, extracting numbers from tables, and coding study features.

They are also, for the same reason, the tasks where automated errors are least visible. A model that extracts a coefficient with the correct value and the wrong label produces output that passes every surface check. Dell'Acqua et al. (2025) provide the framing we find most useful here: in a field experiment with professional consultants, they document a jagged frontier of capability, in which tasks of apparently similar difficulty fall on opposite sides of what the system can do, and in which workers cannot reliably tell in advance which side a given task occupies. Their finding that subjects sometimes deferred to plausible but incorrect model output is precisely the failure mode a synthesis pipeline must be designed against, since an extraction error is both plausible on its face and consequential downstream. Our contribution to this literature is a workflow in which the human expert is positioned at the specific decision points where such errors are consequential, together with an honest account of one case in which that positioning was necessary but not sufficient because the query format was underspecified. We have identified no comparable published synthesis in finance that documents its own extraction failure and the resulting protocol change. We state this as a description of what our search found rather than as a priority claim, since the relevant literature is developing rapidly and our search of it cannot be treated as exhaustive.

2.5 Incremental contribution

Relative to the MBS pricing literature, we contribute the first assembly of the scattered equity-signal coefficients and the identification of benchmark asset as the moderator that reconciles them. Relative to the safe-asset literature, we contribute the observation that its central asymmetry has a direct methodological implication for evidence synthesis. Relative to the literature on AI-assisted research, we contribute a

documented human-authority design and a diagnosed failure case. None of these requires the claim that the underlying economic relation is new, and we do not make that claim.

3. The AI-Agent-Assisted Research Pipeline

This section describes the system that produced the synthesis. We describe it in operational terms rather than architectural ones, because what matters for assessing the results is which decisions the system made and which it did not.

3.1 What the system is

The core research pipeline is a single AI agent operating under a versioned orchestration prompt, currently version 4, which defines the sequence of tasks, the output format for each, and the conditions under which the agent must stop and escalate. That agent, built on the Claude model family, performed literature retrieval, screening, coefficient extraction, classification, state file maintenance, and manuscript drafting. It is not a multi-agent system: tasks run in sequence under one orchestration prompt rather than being distributed across specialized agents that negotiate with one another. The project involved no model training or fine-tuning of any kind. No parameters are estimated by the agent; the numerical coefficients in this paper are values published by other researchers and read from their tables. Readers should note that "agent" here denotes an orchestrated language model performing retrieval and extraction tasks, not a trained decision-making system, and that "beta" denotes an effect estimate extracted from a published regression, not a parameter we estimate.

A second class of AI use should be distinguished from the core pipeline. Auxiliary assistance from a second model family, ChatGPT, was used outside the extraction workflow for critique of proposed pipeline changes, generation and refinement of orchestration prompts, methodological review, checking of summaries and workflow logic, and drafting support for communication among the authors. This auxiliary use touched no coefficient, no classification, and no database entry; it shaped how the pipeline was built and how the work was described rather than what the evidence contains. We separate the two because a claim about what an agent pipeline can accomplish is only meaningful if the boundary of that pipeline is stated, and because the documentation in Appendix D covers both. All persistent state lives in external files rather than model memory. Four files constitute the system state: a beta database recording every extracted estimate with its full provenance, an audit flag log recording every escalated question and its ruling, an escalation queue for open questions, and a session state file. Each working session begins by re-supplying these files. This design was adopted after early sessions demonstrated that model memory could not be relied upon across context boundaries, and it has the incidental benefit that the entire decision history is a set of diffable text files rather than an unrecoverable conversation.

3.2 The human escalation node

The defining feature of the design is that the agent has no authority over finance judgments. The rule is categorical rather than discretionary: the agent may not record a coefficient, assign a study to an inclusion category, resolve an ambiguity about which column of a table corresponds to which specification, or decide the scope of a paper's contribution to the database. Each of these produces an escalation. The coordinating author relays the escalation to the domain expert, records the ruling verbatim in the audit log, and only then permits the agent to update the database.

The coordinating author does not exercise finance judgment either. This is a deliberate separation: the person operating the pipeline is not the person qualified to rule on whether a control-variable coefficient in a spread regression belongs in a meta-analysis of MBS returns. Concentrating all such rulings in a single expert makes the authority structure auditable, because every substantive decision in the database traces to a logged ruling by an identified individual.

Over the course of the project, twenty-one flags were raised and resolved. They fall into four types: classification rulings determining whether a class of estimate qualifies for inclusion, coefficient verifications confirming a specific value against a source, protocol rulings establishing general coding rules, and scope rulings determining how much of a given paper enters the database. Table 1 summarizes the rulings that shaped the current database.

Table 1. Human expert rulings governing the coding protocol

Ruling	Type	Content	Effect on database
HS-01	Classification	Control-variable coefficients qualify when the dependent variable is an observable MBS outcome	Admitted VIX coefficients from spread regressions
CDL-01	Classification	Latent structural model parameters do not qualify as dependent variables	Moved two estimates to mechanism evidence
YM-NEW-01	Protocol	Macro-, firm-, and intermediary-level signals must be coded separately and not pooled	Added signal level field to all entries
YM-NEW-04	Protocol	Statistically insignificant estimates must be included	Retained two null estimates
YM-NEW-07	Classification	Non-MBS fixed income studies are background only	Excluded three estimates from analysis
HS-04 / HS-06	Coefficient / scope	Corrected coefficient assignment; defined scope of source paper	Set final He and Song contribution at six estimates

Notes: The table lists the rulings with the largest effect on the composition of the database. The complete log of twenty-one rulings is reproduced in Appendix C. All rulings were issued by the project domain expert and recorded verbatim before any database update.

3.3 Division of labor

The agent performed literature retrieval and candidate identification, screening against inclusion criteria, extraction of coefficients and test statistics from source tables, drafting of the manuscript, and maintenance of the state files. Neither the agent nor any author estimated a new regression; the paper contains no original econometric estimation, by design.

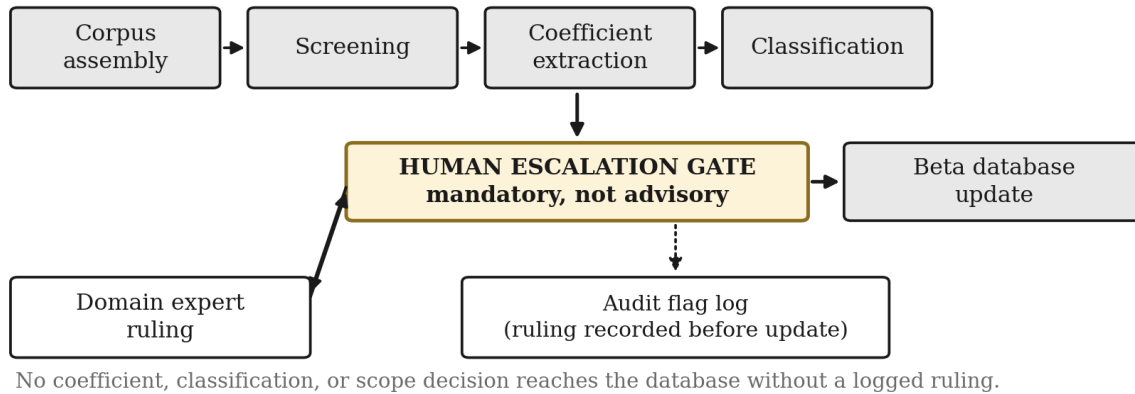
Among the human authors, the division was as follows. The finance domain expert served as final authority for all substantive finance judgments: interpretation, coefficient verification, classification, scope, mechanism judgments, methodological rulings, source and literature input, and final substantive review. A senior finance faculty author advised the domain expert on finance-domain questions. The workflow designer and principal operator defined workflow objectives, determined which task the pipeline should perform next, selected and routed source materials, transferred rulings into the workflow, sequenced successive rounds of processing and revision, maintained continuity across sessions, and organized the state files and audit materials. A further author contributed economics-oriented background research, literature identification, bibliography organization, summary-table development, reference and metadata work, manuscript review, and reference quality control.

One feature of this division is worth stating precisely, because it bears on how contribution should be assessed in workflows of this kind. Many of the detailed prompts directing the pipeline were themselves generated or refined with AI assistance rather than composed word by word, so the human contribution at the orchestration layer is not well described as prompt authorship. It is better described as task definition, sequencing, source selection, and acceptance: deciding what should be done next, what material it should be done on, and whether the result was fit to enter the record. Those decisions were made by the human authors throughout. Substantive finance rulings were made by the domain expert alone and are not attributed elsewhere.

The consequence is an unusual distribution of output. The AI system produced the majority of the research-production output measured by volume, while the human contribution concentrated in domain judgment, workflow orchestration, source provision, verification, interpretation, and final acceptance. A high production share does not confer authority: no estimate entered the database and no claim entered this manuscript without human acceptance, and the human authors retain academic responsibility for the submitted work.

Appendix D records that division in the form the session requires: the full text of the conversations through which the work was directed, a log of human activity, and an accounting of the share of written and documentary output contributed by each party.

Panel A. Task sequence and the human escalation gate
AI agent (retrieval, extraction, drafting)



Panel B. Verification query format, before and after the protocol change

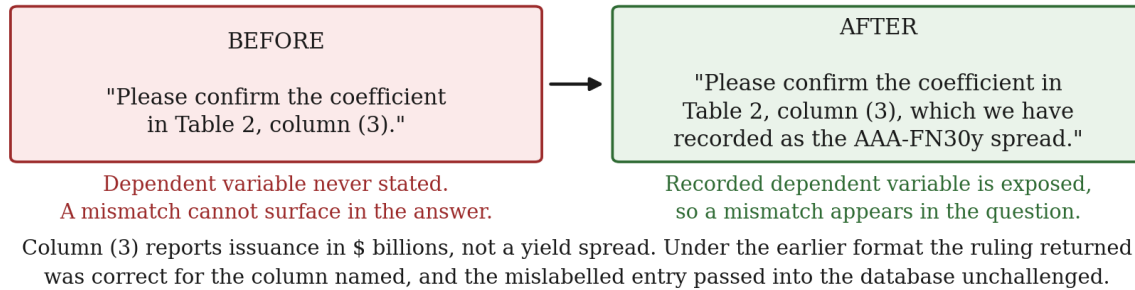


Figure 1. Pipeline architecture and the human escalation node

Notes: Panel A shows the task sequence. The agent performs corpus assembly, screening, coefficient extraction, and classification. Every coefficient-level and classification-level judgment passes through a mandatory escalation gate before any database update, and the ruling is recorded in the audit log before the update is applied. Panel B shows the verification query format before and after the protocol change described in Section 8.3. The added statement of the recorded dependent variable is what allows a mismatch between the database entry and the source table to surface in the question rather than pass silently through the answer.

4. Corpus, Screening, and Coding

4.1 Corpus construction

The candidate corpus was assembled from searches of Google Scholar, SSRN, and Semantic Scholar, supplemented by backward and forward citation tracing from the papers identified as most central. Retrieval prioritized published versions over working papers throughout, and where a published version exists it is the only version from which coefficients are drawn. This rule was adopted after a version discrepancy in one source materially affected an extracted value.

We record one departure from systematic review convention. The corpus was assembled iteratively, with search and screening interleaved rather than separated into distinct logged stages, and we did not maintain a running count of records identified and excluded at each step. We therefore do not present a screening flow diagram, and readers should treat the corpus as the product of a documented set of sources and selection criteria rather than of a reproducible search protocol. This is a limitation of the present synthesis rather than of the method, and it is the first thing we would correct in extending the sample.

4.2 Inclusion criteria

A study contributes an estimate to the main analysis if it satisfies three conditions simultaneously. The independent variable must be an equity market signal, defined as a broad equity index return, an equity volatility measure, an intermediary equity factor, or a firm-level equity variable. The dependent variable must be a directly observed, explicitly MBS-related outcome: a return, an excess return, a price, a yield spread, an option-adjusted spread, a convenience premium, or an issuance quantity. The estimate must be reported with sufficient information to recover its sign, magnitude, and statistical precision.

Two boundary rulings shaped this definition. Estimates in which the equity signal enters as a control variable rather than as the variable of interest are included, on the expert ruling that the coefficient is informative regardless of the role the source paper assigned it. Estimates in which the dependent variable is a latent parameter of a structural pricing model are excluded from the main analysis and retained separately as mechanism evidence, on the ground that a latent factor is not a directly observed MBS outcome. Both rulings are consequential: the first admits most of our sample, and the second removes estimates that would otherwise have appeared to strengthen the results.

4.3 Coding scheme

Every included estimate is coded on the dimensions in Table 2. Three of these fields exist because of rulings issued during the project rather than because we anticipated needing them, which we regard as evidence that the coding scheme was shaped by the evidence rather than imposed on it.

Table 2. Variable and coding definitions

Field	Definition	Values in current sample
Signal category	Type of equity signal	MKT (market return), VOL (volatility), CEQ (intermediary equity)
Signal level	Level of aggregation of the signal	Macro-level; Intermediary-level
Outcome type	Whether the dependent variable measures performance or supply	Performance (8 estimates); Supply (1 estimate)
Benchmark asset	Reference asset against which a spread or premium is measured	AAA corporate; Treasury; not applicable (returns)
Identification level	Whether the estimate identifies aggregate or cross-sectional variation	Aggregate; Cross-sectional
Market segment	Newly issued or seasoned MBS	Newly issued; Seasoned
Units of estimate	Units in which the coefficient is expressed	Basis points per index point; percent per unit return; \$bn per index point
Source version	Version of the source from which the value was taken	Published journal article (all estimates)

Notes: Signal level, outcome type, and benchmark asset were added to the scheme in response to expert rulings issued during coding rather than specified in advance. Units are recorded explicitly because the sample contains estimates in three distinct unit systems, which constrains which estimates may be pooled.

4.4 The included studies

Two studies satisfy the inclusion criteria and contribute the ten estimates analyzed in Section 6. Table 3 summarizes their characteristics. Approximately forty additional papers were processed and classified as MBS context, meaning they inform the interpretation but contribute no extractable estimate, and two mechanism-level estimates from a third study are retained under the classification described in Section 4.2.

Table 3. Characteristics of included studies

	Diep, Eislefeldt & Richardson (2021)	He & Song (2026)
Journal	Journal of Finance 76, 2093-2151	Review of Financial Studies 39(2), 387-426
Frequency	Monthly	Monthly
Sample period	January 1994 - June 2016	October 2003 - December 2022
MBS universe	Fannie Mae 30-year coupon-level indices, defined by coupon relative to par	Fannie Mae 30-year, production coupon and seasoned cohorts
Dependent variables used	Treasury-hedged MBS excess return	Convenience premium spreads; issuance amount
Equity signals used	S&P 500 return; intermediary factor	VIX
Inference	Coupon-level time-series regressions for factor loadings; pooled panel with time fixed effects and standard errors clustered by time for prices of risk	Newey-West; clustered by month in panel specifications
Estimates contributed	4	6

Notes: Estimate counts refer to coefficients meeting the inclusion criteria of Section 4.2, not to the number of regressions reported in each paper. Both papers report substantially more results than are extracted here; extraction is limited to specifications in which an equity signal enters and an MBS outcome is the dependent variable, subject to the scope rulings recorded in Appendix C.

5. Meta-Analytic Framework

5.1 Specification

Let i index estimates and s index source studies. Following standard practice in research synthesis (Hedges and Olkin 1985; Borenstein et al. 2009) and its adaptation to economics, where multiple estimates are typically drawn from each source and study design varies systematically (Stanley and Doucouliagos 2012), the generic random-effects meta-analytic specification for a synthesis of this literature is

$$\theta_i = \mu + \sum_k \gamma_k M_{ki} + u_s + \varepsilon_i, \quad (1)$$

where θ_i is the effect estimate reported for observation i , μ is the mean true effect, M_{ki} is the k -th coded moderator, γ_k is its coefficient, u_s is a study-level random effect capturing dependence among estimates drawn from the same source, and ε_i is sampling error with variance equal to the squared reported standard error of the source estimate. The parameters of interest are μ , the summary effect, and the vector γ , which measures how much of the observed dispersion is explained by coded study features rather than by sampling variation. Where a synthesis reports a summary effect, it is conventional to accompany it with a measure of the share of observed variation attributable to genuine heterogeneity rather than sampling error (Higgins and Thompson 2002), and with a test for asymmetry in the relation between estimates and their precision, which is the standard diagnostic for selective reporting (Egger et al. 1997).

5.2 Why no pooled estimate is reported

Equation (1) is the framework toward which this project is building, and it is not estimable on the current sample. Three obstacles are decisive and we state them plainly rather than estimating something weaker and calling it a result.

First, the sample contains ten estimates from two studies. The study-level random effect u_s cannot be separated from the residual with two clusters, and any standard error computed as though the ten estimates were independent would be badly understated, since six of them come from a single paper and several share a sample period, a data source, and a set of controls. We pre-specified a threshold of thirty estimates or ten source studies before pooled estimation would be attempted; the current sample reaches neither.

Second, the estimates are not expressed in commensurable units, and in one case are not even the same kind of object. Five are basis points of yield spread per index point of VIX. Two are factor loadings measured as percentage excess returns per unit of an equity return or factor. Two are prices of risk on those loadings, which are estimated in a different regression and answer a different question. One is billions of dollars of issuance per index point of VIX. Standardization is possible in principle but requires distributional information about the independent variables that is not uniformly reported in the sources.

Third, and most important for the substance of this paper, the estimates are not drawn from a common population. Section 6 shows that the sign of the effect is determined by benchmark asset. Pooling a positive AAA-benchmarked coefficient with a negative Treasury-benchmarked coefficient would produce a summary effect near zero that describes neither. This is not a small-sample problem that additional estimates would solve; it is a stratification requirement that would remain binding at any sample size.

Accordingly, this paper reports the structure of the evidence rather than its central tendency. We treat the identification and validation of the required stratification as the substantive output of an interim synthesis, and Section 6 is organized to establish it.

5.3 Interpretation of magnitudes

Because no pooled estimate is reported, magnitudes are interpreted individually and in the units of their source. For VIX coefficients on spread outcomes we report the basis point change in the relevant premium associated with a one-point change in the index, and we benchmark that change against the sample mean of the corresponding spread where the source reports it. For return-based estimates we report the percentage change in Treasury-hedged excess return associated with a unit change in the equity signal. We avoid describing any estimate as economically large or small on the basis of statistical significance alone.

6. Results

This section presents the ten confirmed estimates. We organize them around four empirical questions rather than by source paper, because the substantive findings cut across the sources. Table 4 reports the nine performance-outcome estimates and Table 5 reports the single supply-outcome estimate separately.

Table 4. Equity signals and MBS performance outcomes: confirmed effect estimates

ID	Source	Table / column	Equity signal	Dependent variable	Benchmark	Estimate	t-stat
B008	He & Song (2026)	Table 2, col. (2)	VIX	AAA-FN30y spread (bps)	AAA corporate	+3.207	2.673***
B011	He & Song (2026)	Table 2, col. (1)	VIX	Treasury-FN30y spread (bps)	Treasury	-1.257	-4.905***
B010	He & Song (2026)	Table 3, col. (3)	VIX	AAA-FN30y spread (bps)	AAA corporate	+2.328	2.201**
B009a	He & Song (2026)	Table 7, panel B	VIX	AAA-FN30y spread, seasoned (bps)	AAA corporate	+2.236	5.524***
B009b	He & Song (2026)	Table 7, panel B	VIX	Treasury-FN30y spread, seasoned (bps)	Treasury	-2.299	-5.165***
B004	Diep et al. (2021)	Table VI, S&P 500 panel	S&P 500 return	Treasury-hedged MBS excess return	n/a	+3.61%	4.95***
B005	Diep et al. (2021)	Table VI, Intermediary panel	HKM intermediary factor	Treasury-hedged MBS excess return	n/a	+2.80%	6.03***

ID	Source	Table / column	Equity signal	Dependent variable	Benchmark	Estimate	t-stat
B004-X	Diep et al. (2021)	Table X, S&P 500 panel	S&P 500 loading	Price of risk, MBS cross section	n/a	-0.17%	-0.21
B005-X	Diep et al. (2021)	Table X, Intermediary panel	HKM loading	Price of risk, MBS cross section	n/a	-0.41%	-0.41

Notes: All values are as published in the sources named; this paper estimates no regressions. He and Song (2026) report Newey-West standard errors with bandwidth $0.75N^{1/3}$ in time-series specifications ($N = 231$) and standard errors clustered at the month level in the seasoned MBS panel ($N = 1,550$, coupon fixed effects). VIX coefficients are in basis points of spread per index point. For Diep et al. (2021), B004 and B005 are factor loadings from coupon-level time-series regressions of Treasury-hedged monthly MBS excess returns, read at the par coupon row, defined as relative coupon equal to zero; both regressions also include the level and incentive-sensitivity prepayment factors. B004-X and B005-X are estimated prices of risk on the S&P 500 and intermediary factor loadings respectively, each from a pooled time-series cross-section regression with time fixed effects and standard errors clustered by time, estimated on 1,915 relative-coupon month observations. They are distinct empirical objects from the loadings B004 and B005 and are not comparable to them in magnitude. *** and ** denote significance at the 1% and 5% levels. Estimates are not pooled; see Section 5.2.

6.1 Benchmark asset determines the sign of the volatility effect

The first and most consequential finding concerns B008 and B011. These two estimates come from the same table of the same paper, estimated on the same monthly sample from October 2003 through December 2022, with the same control set comprising the mortgage rate, the log debt-to-GDP ratio, and the log ratio of seasoned MBS to GDP. The specifications differ in exactly one respect: the benchmark asset against which the MBS convenience premium is measured.

Measured against AAA corporate bonds, a one-point increase in VIX is associated with a 3.207 basis point increase in the premium, with a t-statistic of 2.673. Measured against Treasury securities, the same one-point increase is associated with a 1.257 basis point decrease, with a t-statistic of -4.905. Both are significant at the 1% level and the signs are opposite. Against the reported sample mean of the AAA-benchmarked spread, roughly 39 basis points, a one-standard-deviation move in VIX implies a change in the premium of the same order as the mean itself, so the magnitude is not negligible relative to the level of the series being explained.

The economic interpretation follows from the structure of safe-asset demand and is the one the source paper itself advances. When equity volatility rises, investors shift toward safe assets in aggregate. Agency MBS benefit from that shift relative to corporate credit, because their principal balances carry government backing and they receive more favorable regulatory treatment under liquidity coverage requirements. Treasury securities, however, benefit more, because they supply a higher level of liquidity services. The MBS premium therefore widens against corporate bonds and compresses against Treasuries simultaneously. The two coefficients are not in tension; they are the two margins of a single reallocation.

The implication for evidence synthesis is direct. Any procedure that pools these two estimates without conditioning on benchmark asset will average +3.207 and -1.257 toward a small positive number and report that equity volatility has a weak effect on the MBS convenience premium. That conclusion would be an artifact of aggregation. Both underlying estimates are precisely determined and economically interpretable; it is their combination that is meaningless.

6.2 The contrast recurs across specifications, segments, and estimation frameworks

A contrast observed within a single table might reflect a feature of that specification. The remaining He and Song estimates allow us to test whether it does.

B010 comes from a specification in which the short-term federal funds rate replaces the mortgage rate as the primary independent variable, with the same controls otherwise. The VIX coefficient on the AAA-benchmarked spread remains positive and significant at +2.328 with a t-statistic of 2.201. The point estimate is smaller than in the baseline, which is expected given the correlation between the two rate variables, but the sign and significance are preserved under a change in the specification's primary regressor.

B009a and B009b come from a structurally different analysis: a panel of seasoned MBS coupon cohorts with coupon fixed effects, 1,550 observations, and standard errors clustered at the month level, in which the primary independent variable is moneyness rather than a rate level. The market segment differs, the identification differs, and the estimation framework differs. The VIX coefficients are +2.236 against AAA corporate bonds and -2.299 against Treasuries, with t-statistics of 5.524 and -5.165 respectively, both significant at the 1% level.

The sign contrast therefore survives a change in the primary independent variable, a change from newly issued to seasoned securities, a change from time-series to panel estimation, and a change in the standard error construction. Within the limits of a two-study sample we regard this as meaningful evidence that benchmark asset is a structural moderator rather than a feature of one regression. We note the obvious caveat that all five volatility estimates come from a single research team using overlapping data, so this is replication across specifications, not across independent samples.

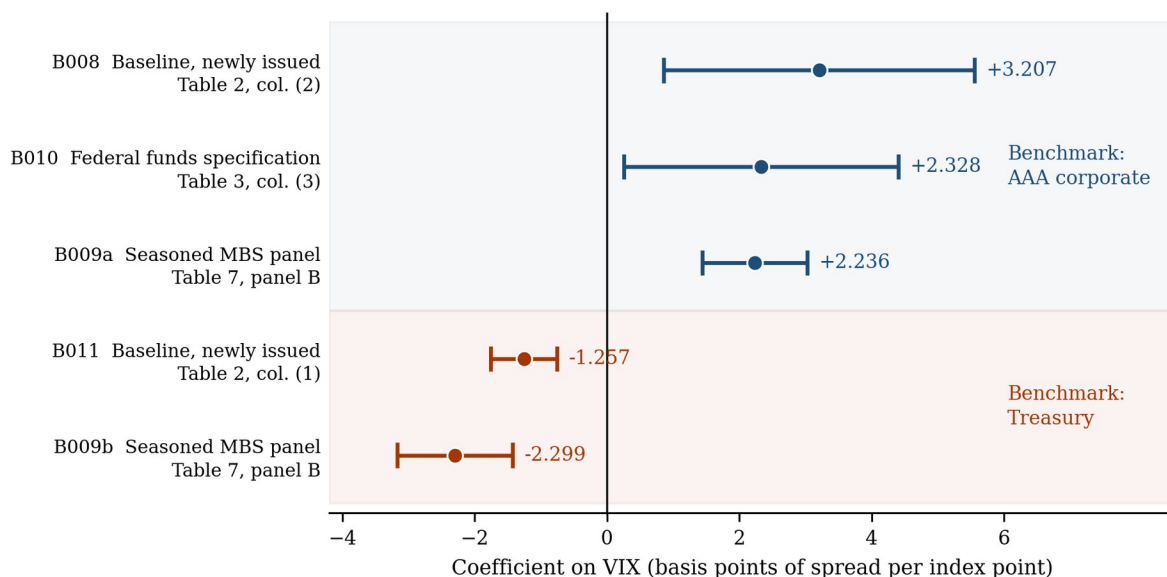


Figure 2. Volatility coefficients by benchmark asset

Notes: Each marker is a published coefficient on VIX from a regression whose dependent variable is an agency MBS convenience premium, drawn from Table 4. Horizontal bars are 95% confidence intervals implied by the reported test statistics. Estimates measured against AAA corporate bonds are shown in the upper band and estimates measured against Treasury securities in the lower band. The separation is complete: no AAA-benchmarked estimate is negative and no Treasury-benchmarked estimate is positive. The supply-outcome estimate B012 is omitted because it is denominated in billions of dollars rather than basis points.

6.3 Equity signals at different levels are not interchangeable

The Diep, Eisfeldt, and Richardson (2021) estimates address a different equity signal and a different outcome. Both are loadings from coupon-level time-series regressions of Treasury-hedged monthly MBS excess returns, estimated over January 1994 to June 2016 alongside the level and incentive-sensitivity prepayment factors. B004 reports that a unit increase in the raw monthly S&P 500 return is associated with a 3.61% increase in the excess return at the par coupon, with a t-statistic of 4.95. B005 reports that a unit increase in the intermediary asset pricing factor of He, Kelly, and Manela (2017) is associated with a 2.80% increase in the same outcome, with a t-statistic of 6.03. The par coupon, defined as a relative coupon of zero, was selected by expert ruling as the representative row because it is neutral with respect to discount and premium MBS and therefore isolates the equity signal most cleanly.

These two estimates are similar in magnitude, and it would be tempting to treat them as two measurements of one quantity. They are not. A broad market return and an intermediary equity factor capture different economic channels: the first reflects aggregate equity market conditions, the second the balance sheet capacity of financial intermediaries who are the marginal holders of MBS. Under the project coding protocol these are distinct signal levels and are reported in separate strata. The similarity of the point estimates is a substantive observation about the two channels, not a justification for combining them.

6.4 Equity loadings are not priced in the MBS cross section

B004-X and B005-X are the estimates most likely to be omitted from a less disciplined synthesis, and they are the ones whose interpretation depends most on stating precisely what the dependent variable is. Neither is a second return regression. Each is an estimated price of risk on a factor loading, obtained from a pooled time-series cross-section regression of MBS excess returns with time fixed effects, estimated on 1,915 relative-coupon month observations with standard errors clustered by time. For the S&P 500 loading the estimate is -0.17% with a t-statistic of -0.21. For the intermediary factor loading it is -0.41% with a t-statistic of -0.41. Neither is distinguishable from zero.

Read alongside B004 and B005, the pair establishes a distinction that a synthesis of this literature must preserve. Equity signals comove with MBS excess returns: both loadings are significant, at 4.95 and 6.03 respectively. But exposure to that comovement earns no compensation in the cross section of MBS once prepayment risk is accounted for. Comovement and pricing are different claims, and this literature supports the first without the second.

The source paper reaches the same conclusion independently and from its own direction. It reports that loadings on standard asset pricing factors display little cross-sectional variation and therefore add almost nothing to explaining the cross section, and it concludes that traditional intermediary asset pricing factors do not price the cross section of MBS returns, attributing the pricing of prepayment risk instead to specialized traders who may constitute only a small part of bank holding company returns.

Both estimates are retained under the expert protocol ruling that statistically insignificant results must enter the database, on the reasoning that excluding them introduces selection bias and inflates any eventual summary effect. We state what they support and no more. Two insignificant price-of-risk estimates from one specification in one paper do not establish that equity exposure is unpriced in MBS markets generally. They establish that in this sample, with these controls, the compensation is not detectable, and that a synthesis reporting only the significant loadings would misrepresent the evidence base.

6.5 Supply outcomes are a separate question

The final estimate concerns quantity rather than price. Table 5 reports it separately.

Table 5. Equity volatility and MBS supply: separate panel

ID	Source	Table / column	Signal	Dependent variable	Estimate	t-stat
B012	He & Song (2026)	Table 2, col. (3)	VIX	FN30y production-coupon monthly issuance (\$bn)	+0.217	1.481

Notes: Reported in a separate panel because the dependent variable is a quantity measured in billions of dollars, whereas all estimates in Table 4 are measured in basis points or percentage returns. The estimate is not statistically significant at conventional

levels and is retained under the full-evidence protocol. This coefficient must not be pooled with the estimates in Table 4 under any weighting scheme.

B012 is drawn from the third column of the same baseline table that produces B008 and B011, and it measures the association between VIX and the monthly issuance amount of production-coupon Fannie Mae 30-year MBS. The point estimate implies that a one-point increase in VIX is associated with an additional \$0.217 billion of monthly issuance, with a t-statistic of 1.481. The estimate is not significant at conventional levels.

Two points follow. Substantively, the demand-side effect of equity volatility on MBS pricing documented in Section 6.1 does not have a clearly detectable counterpart on the supply side within this sample, which is consistent with issuance responding to mortgage rates and origination conditions on a slower timescale than spreads respond to volatility. Methodologically, this estimate cannot be pooled with any estimate in Table 4 because its units are billions of dollars rather than basis points, and pooling it would require a standardization for which the sources do not report the necessary inputs. This is the same non-pooling logic the project applies to signal level, applied to outcome type.

We flag that this coefficient is also the origin of the extraction error described in Section 8.3. It was for a period recorded in the database as a convenience premium estimate, which had the effect of making the baseline volatility relation appear statistically insignificant when in fact it is significant at the 1% level. The correction reversed that conclusion.

7. Mechanisms

The results in Section 6 describe a pattern. This section discusses what generates it, and separates the part the estimates support from the part that remains interpretation. Three channels are candidates. They are not mutually exclusive, and Section 7.4 explains why the present evidence cannot separate them.

7.1 Reallocation within a hierarchy of safe assets

The benchmark-dependent sign structure is the empirical signature of a reallocation across assets that supply safety and liquidity services in different degrees. In a framework where investors derive convenience from safe asset holdings, a rise in aggregate safe-asset demand raises every convenience premium, but by amounts ordered by how much service each asset provides. Krishnamurthy and Vissing-Jorgensen (2012) establish the magnitude of this premium for Treasury debt, and Nagel (2016) shows that liquidity premiums on near-money assets vary with conditions that proxy for the demand for safety. Agency MBS supply less liquidity service than Treasury securities but carry government backing and favorable regulatory treatment relative to corporate credit, which places them between the two. A common demand shock therefore moves

the two spreads in opposite directions, which is the comparative static of the model in He and Song (2026) and the pattern documented in Section 6.1.

Direct evidence on the reallocation itself supports this reading. Baele, Bekaert, Inghelbrecht, and Wei (2020) identify flight-to-safety episodes across twenty-three countries using stock and bond returns alone, and find that such days are rare, comprising less than two percent of their sample, but are associated with a large bond-equity return differential and with fund flows out of equity and into government bond and money market funds. They further find that in the U.S. corporate bond market the phenomenon is primarily a flight to quality rather than to liquidity. That distinction matters for our estimates. If the reallocation into safety were driven mainly by a liquidity motive, agency MBS would be disadvantaged relative to Treasuries on the liquidity margin alone, and the compression of the MBS-Treasury spread would carry no implication for the MBS-corporate spread. That both spreads move, in opposite directions, is more consistent with a quality-driven reallocation in which MBS gain against corporate credit on the safety margin while losing to Treasuries on the liquidity margin.

We note the limit of this argument. Baele et al. identify flight-to-safety days from realized returns, whereas our estimates use a volatility index as a continuous regressor over monthly data. The two are related but not the same object, and we do not claim that our coefficients measure the effect of the discrete episodes they identify.

7.2 Prepayment uncertainty

A second channel operates through the prepayment option. Elevated equity volatility coincides with uncertainty about future interest rate paths, which changes the value of the borrower option embedded in every agency MBS. Gabaix, Krishnamurthy, and Vigneron (2007) establish that prepayment risk carries a price in this market and that the price reflects the risk-bearing capacity of specialized investors rather than aggregate risk aversion, which implies that shocks affecting those investors should move MBS prices even when economy-wide fundamentals are unchanged. Boyarchenko, Fuster, and Lucca (2019) decompose mortgage spreads and show that prepayment compensation accounts for a substantial share of their variation, so a channel operating through prepayment has a large enough base to matter for the spreads we analyze.

The mechanism estimates in Appendix B bear on this directly. Chernov, Dunn, and Longstaff (2018) estimate a model in which prepayment is driven by macroeconomic factors and recover latent parameters governing the turnover rate and borrower responsiveness to refinancing incentives. In their framework, VIX is positively associated with the change in implied turnover and negatively associated with the change in the rate response factor. These estimates are excluded from the main analysis because latent structural

parameters are not directly observed MBS outcomes, but they indicate that periods of elevated equity volatility coincide with shifts in how the market prices prepayment behavior rather than merely in the level of spreads.

We regard this channel as plausible and not established. The two latent parameters move in opposite directions, and mapping that pattern onto a prediction about the sign of a spread coefficient requires the full structure of their model rather than the two coefficients alone. A synthesis that treated these estimates as corroborating the main results would be overreading them, which is the reason for the classification ruling that keeps them separate.

7.3 Intermediary capacity

A third channel runs through the balance sheets of the institutions that hold MBS. He and Krishnamurthy (2013) model asset markets in which the marginal investor is a financial intermediary subject to an equity capital constraint, and show that risk premia rise when the constraint binds. Adrian, Etula, and Muir (2014) show empirically that a pricing kernel built from intermediary balance sheet information explains the cross section of returns across several asset classes. Because equity prices of financial institutions are observable at high frequency, equity-based measures of intermediary condition are natural proxies for the constraint, which is the logic underlying the factor of He, Kelly, and Manela (2017).

The B005 estimate speaks to this channel. A unit increase in the intermediary factor is associated with a 2.80% increase in the Treasury-hedged MBS excess return at the par coupon, with a t-statistic of 6.03, and the loading is significant across the coupon stack. The estimate is consistent with intermediary condition mattering for MBS returns, and with the finding of Gabaix, Krishnamurthy, and Vigneron (2007) that the marginal investor in this market is specialized rather than representative.

A significant loading is not, however, a priced risk, and B005-X shows directly that the two come apart here: the price of risk on the intermediary loading is -0.41% with a t-statistic of -0.41. The source paper draws the same conclusion, reporting that intermediary factor loadings show little cross-sectional variation and that traditional intermediary asset pricing factors do not price the cross section of MBS returns, attributing the pricing of prepayment risk instead to specialized traders who may represent only a small part of bank holding company returns. The channel therefore has support at the level of comovement and not at the level of compensation, and our database records both halves of that result rather than only the significant one. Separating the two would require estimates from periods in which intermediary constraints plausibly bind, and neither of our sources reports the subsample split that would provide them.

7.4 Why the channels cannot be separated here

All three channels predict a positive association between equity volatility and the AAA-benchmarked premium, so the sign of our central estimates does not discriminate among them. The difficulty is compounded by the nature of the signal itself. Bekaert, Hoerova, and Lo Duca (2013) show that the squared volatility index decomposes into a component reflecting expected stock market variance and a residual interpretable as risk aversion, and that the two components behave differently and have distinct relations to monetary policy. Our volatility estimates therefore load on both an uncertainty component, which is the natural driver of the prepayment channel, and a risk aversion component, which is the natural driver of both the safe-asset and intermediary channels. A coefficient on the composite index cannot separate them.

Three developments would make separation possible. Estimates using the decomposed components rather than the index would map more cleanly onto the channels. Estimates from subsamples in which intermediary constraints plausibly bind would isolate the third channel. And estimates in which the channels make opposing predictions, rather than concordant ones, would be far more informative than additional estimates of the same sign. We record these as design requirements for the extraction protocol as the sample grows, rather than as a defect of the present analysis, which was assembled to characterize the evidence rather than to adjudicate mechanisms.

8. Sensitivity and Workflow Diagnostics

Conventional robustness checks apply to estimates the analyst produces. Because we produce none, the analogous exercise is to ask which of our conclusions would change under different coding decisions, and to report what is known about the reliability of the extraction process itself.

8.1 Sensitivity to coding rulings

Three rulings materially affect the composition of the database, and we state the consequence of reversing each.

Had control-variable coefficients been excluded, the sample would contain four estimates rather than ten and every volatility result would disappear, since VIX enters the source specifications as a control. The central finding of this paper depends entirely on this ruling. Had insignificant estimates been excluded, B004-X, B005-X, and B012 would all be dropped. Both cross-sectional price-of-risk nulls of Section 6.4 would vanish along with the null supply-side estimate, and the remaining sample would consist exclusively of significant results. The distortion would not be uniform: it would remove precisely the estimates that qualify what the significant ones establish, leaving a literature that appears to show equity signals moving MBS outcomes with no counterweight. Had the scope of the He and Song contribution been restricted to baseline results only, B009a, B009b, and B010 would be removed, leaving seven estimates; the benchmark

contrast of Section 6.1 would survive intact because it is contained within the baseline table, but the replication evidence of Section 6.2 would be lost. The domain expert ruled that the earlier confirmations stand, and we note that the paper's central claim does not depend on that ruling having gone the way it did.

8.2 Sensitivity to inclusion of the non-MBS literature

Studies of equity signals and non-MBS fixed income outcomes were excluded by expert ruling. Their reported signs are mixed. Had they been included, the sample would be larger and more heterogeneous, and the benchmark-asset moderator identified here would not obviously extend to dependent variables that are not MBS spreads. We regard the exclusion as correct but note that it narrows the literature to which our conclusions apply.

8.3 An extraction error and its diagnosis

During coding, one coefficient was recorded with the correct numerical value and the wrong dependent variable. The estimate now labeled B012, a VIX coefficient on MBS issuance measured in billions of dollars, was recorded as a coefficient on the AAA-benchmarked convenience premium measured in basis points. The consequence was material: the baseline volatility relation appeared statistically insignificant, a mechanism narrative was drafted around that null, and the analysis was framed as documenting heterogeneity across specifications. All three were wrong.

The error is instructive because of how it survived review. The verification query sent to the domain expert named a table and a column but did not state what the column measured. The expert answered accurately for the column he was asked about. The error was therefore ratified by a correct human ruling, and no subsequent check could detect it, because every downstream artifact was internally consistent with the mislabeled entry. It was found only when the published source was re-read directly against the database, at which point the column in question was seen to report issuance rather than a spread.

A further detail makes the diagnosis sharper. The value originally extracted from the working paper version of the source had been close to the correct published value. An intermediate "correction" replaced it with the mislabeled figure. The correction, not the original extraction, introduced the error. A workflow that treats later revisions as automatically more reliable than earlier ones would have entrenched it.

The protocol change is narrow. Every verification query must now state the table, the column, and the dependent variable as recorded, so that a mismatch appears in the question itself rather than passing silently through the answer. Every database entry carries both fields and both are checked against the source before a query is issued. We report this episode in full because the value of a workflow paper lies in what it reveals

about failure modes, and because the failure mode here was not hallucination, arithmetic error, or misreading, but an underspecified interface between the automated and human components of the system.

8.4 Verification status of the current database

All ten estimates in Tables 4 and 5 have been verified against their published sources under the query protocol described above, which requires the table, the column or panel, and the recorded dependent variable to be stated together and checked against the source before an entry is admitted. The six He and Song coefficients and their test statistics were verified against the published Review of Financial Studies article. The four Diep, Eisfeldt, and Richardson coefficients were verified against the published Journal of Finance article, including the representative row, the units of both independent variables, and, for B004-X and B005-X, the distinction between a factor loading and a price of risk on that loading. No estimate carries unresolved verification status and the audit log records no open flags.

The re-verification of the Diep, Eisfeldt, and Richardson entries under the stricter protocol is worth recording, because it changed one entry without changing its value. B004-X had been carried in the database as a cross-sectional return regression. The published source reports it as a price of risk on an estimated factor loading, in a specification with time fixed effects. The coefficient and test statistic were correct as recorded; the description of what they measured was not, and the economic reading in Section 6.4 is sharper as a result. This is the same class of defect as the B008 error, caught this time by the procedure designed to catch it.

9. Discussion

9.1 What the findings imply

The primary implication is methodological and applies to anyone attempting to summarize this literature quantitatively. Benchmark asset is not a nuisance covariate; it determines the sign of the estimated effect. A synthesis that ignores it will find that equity volatility has little effect on MBS convenience premiums, and that finding will be an artifact. The same logic extends to signal level and outcome type, where pooling across categories combines estimates that answer different questions or are measured in different units.

The substantive implication is that the relation between equity volatility and MBS pricing is better described as a reallocation within a hierarchy of safe assets than as a single directional effect. Asking whether volatility raises or lowers the MBS premium is not well posed without specifying the comparison asset. This is a modest reframing rather than a discovery, since it follows from a framework already present in the source literature; what the synthesis adds is the demonstration that the reframing is necessary in practice, because the alternative produces a wrong answer.

9.2 What the findings do not imply

We do not establish a population effect size for the relation between equity signals and MBS outcomes. We report no summary estimate and the sample would not support one. We do not establish causality; every estimate analyzed comes from a predictive or associational specification, several of them with the equity signal entering as a control, and nothing in our design licenses a causal reading. We do not establish that the benchmark-asset moderator generalizes beyond agency MBS, since our sample contains no other asset class. We do not establish that the sign contrast holds outside the sample periods of the two included studies.

We also do not claim that the equity-MBS relation is stable. Five of our nine performance estimates come from one research team using overlapping data. Replication across specifications within a research design is weaker evidence than replication across independent samples, and our sample provides only the former.

9.3 Implications for AI-assisted research

The workflow evidence supports a specific and limited conclusion. The tasks the agent performed well were high-volume and well-specified: locating candidate papers, screening them against stated criteria, drafting prose, and maintaining consistent state across sessions. The task where the system failed was the one requiring a judgment about correspondence between a recorded label and a source document, which is neither high-volume nor mechanical. This is consistent with the jagged frontier documented by Dell'Acqua et al. (2025): the boundary did not track apparent task difficulty, since reading a coefficient from a table and checking that coefficient against its column label are tasks of similar surface difficulty that fell on opposite sides of it.

The design implication is that human effort in such workflows should be spent on interface specification rather than on output review. Reviewing the agent's output would not have caught our error, because the output was internally consistent. What would have caught it, and now does, is a query format that forces the mismatch into the open. This is a claim about one project and one failure, and we present it as a case rather than as a general finding.

9.4 Limitations

The limitations are substantial and we state them without qualification. The sample is ten estimates from two studies. No formal screening log exists, so the corpus cannot presently be reproduced from a documented search protocol. The included studies overlap in authorship and data for the majority of the estimates, and four of the ten come from a single table of a single paper. No pooled effect size is estimated and none could be. Publication bias cannot be assessed: funnel asymmetry tests have little power at this number of estimates, a limitation the original proponents of the test emphasize (Egger et al. 1997). The

mechanism discussion is interpretive rather than tested. And the workflow evidence is a single case study of a single pipeline, evaluated by the authors who built it.

10. Conclusion

We asked whether equity market signals predict mortgage-backed security outcomes and why the published evidence appears to disagree. Assembling the scattered coefficients into a single coded database shows that the disagreement is apparent rather than real. The sign of the estimated effect of equity volatility on the MBS convenience premium is determined by the benchmark asset against which the premium is measured: positive against AAA corporate bonds, negative against Treasury securities, with both estimates significant at the 1% level in the same baseline regression and the contrast recurring across a change in the primary regressor, a change in market segment, and a change in estimation framework. The pattern is the empirical signature of a reallocation within a hierarchy of safe assets, in which MBS gain against corporate credit while Treasuries gain more still.

The synthesis also identifies three further coding requirements: equity signals at different levels of aggregation should not be pooled, aggregate predictability should not be conflated with cross-sectional explanatory power, and supply outcomes belong in a separate panel from performance outcomes. Together these define the stratification any future quantitative synthesis of this literature must respect.

We report no summary effect size. Ten estimates from two studies fall below the threshold we set in advance for pooled estimation, and the benchmark-asset result implies that pooling across strata would remain inappropriate at any sample size. What the current evidence delivers is a validated coding protocol and a documented moderator structure, which is the appropriate output of an interim synthesis and the necessary input to a later one.

The paper was produced by an AI agent operating under a protocol in which every finance judgment was escalated to a human expert. We have described that protocol, and we have described in detail the one occasion on which it failed, the reason it failed, and the change that followed. The failure was not in the model's reading of a number or in the expert's ruling, but in the format of the question that passed between them. We think this is the most useful thing we can report about the workflow, and more useful than an account in which nothing went wrong.

The natural extensions are to expand the corpus toward the pooling threshold, to seek estimates from research teams and datasets independent of those represented here, and to design extractions capable of separating the safe-asset, prepayment, and intermediary channels that the current evidence cannot distinguish.

References

- Adrian, T., E. Etula, and T. Muir. 2014. Financial Intermediaries and the Cross-Section of Asset Returns. *Journal of Finance* 69:2557-2596.
- Baele, L., G. Bekaert, K. Inghelbrecht, and M. Wei. 2020. Flights to Safety. *Review of Financial Studies* 33(2):689-746.
- Bekaert, G., M. Hoerova, and M. Lo Duca. 2013. Risk, Uncertainty and Monetary Policy. *Journal of Monetary Economics* 60(7):771-788.
- Borenstein, M., L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein. 2009. *Introduction to Meta-Analysis*. Chichester: John Wiley and Sons.
- Boyarchenko, N., A. Fuster, and D. O. Lucca. 2019. Understanding Mortgage Spreads. *Review of Financial Studies* 32:3799-3850.
- Chernov, M., B. Dunn, and F. A. Longstaff. 2018. Macroeconomic-Driven Prepayment Risk and the Valuation of Mortgage-Backed Securities. *Review of Financial Studies* 31:1132-1183.
- Collin-Dufresne, P., R. S. Goldstein, and J. S. Martin. 2001. The Determinants of Credit Spread Changes. *Journal of Finance* 56:2177-2207.
- Dell'Acqua, F., E. McFowland III, E. R. Mollick, H. Lifshitz-Assaf, K. Kellogg, S. Rajendran, L. Kraymer, F. Candelon, and K. R. Lakhani. 2025. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality. *Organization Science*.
- Diep, P., A. L. Eisfeldt, and S. Richardson. 2021. The Cross Section of MBS Returns. *Journal of Finance* 76:2093-2151.
- Egger, M., G. Davey Smith, M. Schneider, and C. Minder. 1997. Bias in Meta-Analysis Detected by a Simple, Graphical Test. *BMJ* 315:629-634.
- Fernandes, D., J. G. Lynch Jr., and R. G. Netemeyer. 2014. Financial Literacy, Financial Education, and Downstream Financial Behaviors. *Management Science* 60(8):1861-1883.
- Fusari, N., W. Li, H. Liu, and Z. Song. 2022. Asset Pricing with Cohort-Based Trading in MBS Markets. *Journal of Finance* 77:3249-3287.
- Gabaix, X., A. Krishnamurthy, and O. Vigneron. 2007. Limits of Arbitrage: Theory and Evidence from the Mortgage-Backed Securities Market. *Journal of Finance* 62:557-595.
- Gorton, G., and G. Pennacchi. 1990. Financial Intermediaries and Liquidity Creation. *Journal of Finance* 45:49-71.
- He, Z., and A. Krishnamurthy. 2013. Intermediary Asset Pricing. *American Economic Review* 103(2):732-770.
- He, Z., and Z. Song. 2026. Agency MBS as Safe Assets. *Review of Financial Studies* 39(2):387-426.
- He, Z., B. Kelly, and A. Manela. 2017. Intermediary Asset Pricing: New Evidence from Many Asset Classes. *Journal of Financial Economics* 126:1-35.
- Hedges, L. V., and I. Olkin. 1985. *Statistical Methods for Meta-Analysis*. San Diego: Academic Press.

- Higgins, J. P. T., and S. G. Thompson. 2002. Quantifying Heterogeneity in a Meta-Analysis. *Statistics in Medicine* 21:1539-1558.
- Keim, D. B., and R. F. Stambaugh. 1986. Predicting Returns in the Stock and Bond Markets. *Journal of Financial Economics* 17:357-390.
- Korinek, A. 2023. Generative AI for Economic Research: Use Cases and Implications for Economists. *Journal of Economic Literature* 61(4):1281-1317.
- Korinek, A. 2025. AI Agents for Economic Research. NBER Working Paper 34202.
- Krishnamurthy, A., and A. Vissing-Jorgensen. 2012. The Aggregate Demand for Treasury Debt. *Journal of Political Economy* 120:233-267.
- Kwan, S. H. 1996. Firm-Specific Information and the Correlation between Individual Stocks and Bonds. *Journal of Financial Economics* 40:63-80.
- Nagel, S. 2016. The Liquidity Premium of Near-Money Assets. *Quarterly Journal of Economics* 131:1927-1971.
- Newey, W., and K. West. 1987. A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55:703-708.
- Noy, S., and W. Zhang. 2023. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. *Science* 381(6654):187-192.
- Stanley, T. D., and H. Doucouliagos. 2012. *Meta-Regression Analysis in Economics and Business*. Oxon: Routledge.
- Sunderam, A. 2014. Money Creation and the Shadow Banking System. *Review of Financial Studies* 28:939-977.

Appendix A. Coding Manual

This appendix states the operational rules applied to each coded field in Table 2.

A.1 Signal category and signal level

Signal category records the type of equity variable. MKT denotes a broad equity index return, VOL denotes an equity volatility measure, and CEQ denotes an equity-based measure of financial intermediary condition. Signal level records the level of aggregation at which the variable is measured: macro-level for economy-wide index returns and volatility indices, intermediary-level for factors constructed from the equity of financial institutions, and firm-level for individual company equity variables. The two fields are distinct because a single category can appear at more than one level. Estimates are reported in separate strata by signal level and are never pooled across levels, on the ruling that variables at different levels of aggregation answer different economic questions even when their coefficients are numerically similar.

A.2 Outcome type and benchmark asset

Outcome type distinguishes performance outcomes, which include returns, excess returns, prices, yield spreads, option-adjusted spreads, and convenience premiums, from supply outcomes, which measure issued quantity. The distinction is enforced because the two are measured in incommensurable units and because a demand-side effect on pricing has no necessary counterpart in issuance at the same horizon.

Benchmark asset applies to any estimate whose dependent variable is a spread or premium and records the reference asset against which it is measured. Values in the current sample are AAA corporate bonds and Treasury securities. Return-based estimates take the value not applicable. This field carries the weight of the paper's central finding and is required for every spread-based entry; an entry lacking it is not admissible to the database.

A.3 Identification level and market segment

Identification level distinguishes estimates that identify variation in the aggregate level of MBS outcomes over time from estimates that identify variation across securities at a point in time. The distinction matters because the two can differ in both sign and significance within a single source, as they do in the present sample. Market segment records whether the securities are newly issued or seasoned, since the relation between the current mortgage rate and the moneyness of a security runs in opposite directions across the two.

A.4 Units and source version

Units are recorded verbatim in the units of the source rather than converted, because conversion requires distributional information about the independent variable that sources do not report uniformly. The current sample contains three unit systems: basis points of spread per index point, percentage return per unit of an equity return or factor, and billions of dollars of issuance per index point. Source version records the version from which the value was drawn. Where a published version exists, it is the only admissible source; working paper values are not used even when they are more convenient, following a version discrepancy early in the project that affected an extracted value.

A.5 Admissibility

An entry entering the database must carry a source, a table and column identifier, the dependent variable as stated in the source, the coefficient, its test statistic, and every field above that applies. An entry missing the dependent variable field is inadmissible regardless of how confident the extraction is, which is the operational form of the protocol change described in Section 8.3.

Appendix B. Mechanism Evidence Retained Outside the Main Analysis

Two estimates from Chernov, Dunn, and Longstaff (2018) were classified as mechanism evidence on the ruling that their dependent variables are latent structural parameters rather than directly observed MBS outcomes. They are recorded here and are excluded from all analysis in Section 6.

Table B1. Mechanism-level estimates

ID	Signal	Latent dependent variable	Estimate	Direction
CDL-M1	VIX	Quarterly change in implied turnover rate	+0.1159	Positive, significant
CDL-M2	VIX	Quarterly change in implied rate response factor	-0.2234	Negative, significant

Notes: Excluded from the main analysis under ruling CDL-01. Significance is as characterized in the source; exact test statistics are not reproduced here because these estimates enter no calculation in this paper.

Appendix C. Log of Human Expert Rulings

Every substantive decision in the database traces to a ruling recorded here. The log is reproduced in full because the workflow claim of this paper depends on the escalation structure being auditable rather than asserted.

Table C1. Complete ruling log

Flag	Type	Ruling	Database effect
CDL-01	Classification	Latent structural parameters do not qualify as dependent variables	Two estimates moved to mechanism evidence
HS-01	Classification	Control-variable coefficients qualify when the dependent variable is an MBS outcome	Admitted all volatility estimates

Flag	Type	Ruling	Database effect
HS-02	Units	VIX in raw index points; spread outcomes in basis points	Fixed unit interpretation; re-closed after HS-04
HS-03	Interpretation	Safe-asset narrative supported; report as prose, not channel codes	Shaped Section 7.1
HS-04	Coefficient	B008 corrected to the AAA-benchmarked coefficient	Reversed an earlier correction; see Section 8.3
HS-05	Scope	Confirmed baseline entries; excluded two further specifications	Set three new entries, excluded seven candidates
HS-06	Scope	Earlier confirmations stand; restriction applies only to new candidates	He and Song contribution set at six estimates
YM-NEW-10	Classification	Include the intermediary-factor price of risk by symmetry with the S&P 500 counterpart	B005-X added; final count of ten estimates
YM-NEW-01	Protocol	Signal levels coded separately and reported in separate panels	Added signal level field
YM-NEW-02	Framing	Methodological workflow contribution to be foregrounded	Shaped Sections 1.1 and 9.3
YM-NEW-03	Terminology	Write independent variable in full; no abbreviation	Applied throughout
YM-NEW-04	Protocol	Statistically insignificant estimates must be included	Retained two null estimates
YM-NEW-05	Classification	Two macro-forecasting papers are background only	No entries created
YM-NEW-06	Format	Mechanisms described in prose, not code labels	Rewrote Section 7
YM-NEW-07	Classification	Non-MBS fixed income studies are background only	Removed three entries from analysis
YM-NEW-08	Administrative	Institutional affiliations without department names	Title page
PILOT-01	Classification	Kwan (1996) background only	Reclassified
PILOT-02	Classification	Keim and Stambaugh (1986) background only	Reclassified
PILOT-03	Classification	Collin-Dufresne et al. (2001) background only	Reclassified
PILOT-04	Coefficient	B004 confirmed; par coupon is the representative row	Entry confirmed
PILOT-05	Coefficient	B005 confirmed; factor measured in decimal return units	Entry confirmed

Notes: Type denotes the class of judgment. Classification rulings determine whether a category of estimate is admissible; coefficient rulings verify a specific value against its source; protocol rulings establish general coding rules; scope rulings determine how much of a given source enters the database. All rulings were issued by the project domain expert and recorded before any database update.

Appendix D. Documentation of Human and AI Contributions

This appendix documents the human and AI contributions to the project, as required by the special session. It reports the project timeline, the AI systems and workflow used, the human activity record, the human decision and verification record, and an accounting of the share of project output produced by each party. Detailed underlying records are held as Supporting Exhibits, listed in Section D.8 and submitted with this document.

D.1 Scope and evidentiary basis

The session requires four categories of documentation: the initial prompt and all subsequent conversations with the AI; the AI model, configuration, workflow, and data access; a log of each author's human activities; and the proportion of writing, code, and documentation contributed by humans as opposed to AI. This appendix addresses each in turn.

Throughout, we distinguish three classes of evidence and do not blur them. Timestamped records are drawn from system exports, meeting records, and dated messages, and are reported as recorded. Contemporaneous records are documents produced during the work itself, including the beta database, the audit flag log, and successive manuscript versions. Retrospective estimates are figures reconstructed after the fact by the authors, and are labeled as estimates wherever they appear. We have not converted any retrospective estimate into an apparently logged time.

D.2 Project initiation and timeline

On May 26, 2026, Yao Ma contacted Zhipeng Jiang regarding the special session and proposed exploring an AI-agent-assisted finance paper. Later the same day, Jiang contacted Yanan Zhu, described the session and the proposed project, and invited her participation, including a prospective economics-oriented background research role.

These communications were preliminary. They concerned recruitment, project scoping, feasibility, and a prospective division of labor. No substantive investigation, literature review, data analysis, or AI-assisted research was conducted on May 26, 2026. Substantive project work began only after June 1, 2026, consistent with the eligibility requirement for the session.

The earliest clearly substantive AI-assisted research interaction identified in the conversation archive occurred on June 5, 2026, at approximately 9:35 p.m. Eastern Time, when Jiang submitted background research materials prepared by Zhu for AI-assisted evaluation and the resulting literature set was discussed. On June 6, 2026, at 10:30 a.m. Eastern Time, the project team held its first joint meeting, which marked the beginning of coordinated team work. The division of labor described in Section 3.3 of the paper was established following that meeting.

All dates and times are reported in Eastern Time and are supported by system, meeting, or message records held in Exhibits D-A and D-E.

D.3 AI systems, configuration, and workflow

The core research pipeline used a single orchestrated agent built on the Claude model family, operating under a versioned orchestration prompt. It performed literature retrieval, screening, extraction of regression evidence, classification proposals, state file maintenance, and manuscript drafting. Auxiliary assistance from a second model family, ChatGPT, was used outside that pipeline for prompt generation and refinement, critique of proposed workflow changes, methodological review, checking of summaries and logic, and drafting support for communication among the authors.

The distinction matters for interpreting the results. The auxiliary system touched no coefficient, no classification, and no database entry; it shaped how the pipeline was built and how the work was described rather than what the evidence contains. The project is not a multi-agent system: the core pipeline runs tasks in sequence under one orchestration prompt rather than distributing them across specialized agents. No model was trained or fine-tuned at any point. Source materials were supplied to the systems as uploaded files by the human operator; no system had autonomous access to project data.

The realized workflow followed a fixed sequence: human authors identified and supplied candidate papers and background materials; those materials were submitted to the agent with project-specific extraction instructions; the agent extracted and organized study characteristics, coefficients, and candidate classifications; usable estimates entered a provisional set or were flagged; questions requiring finance-domain or methodological judgment were routed to the domain expert; rulings were recorded and then incorporated into the beta database, audit log, and manuscript; drafting and revision followed once the underlying evidence and open flags had been resolved; and the authors conducted final review before submission.

An earlier plan contemplated a broader sequence involving coding-agent-assisted analysis and later multi-agent refinement. The realized workflow for this paper centered on AI-assisted literature synthesis, evidence extraction, structured human verification, and manuscript drafting. We describe what was implemented rather than what was contemplated.

D.4 Human activity record

Table D.1 records the human activities for which a dated supporting record exists. Where elapsed working time was not recorded contemporaneously, the table says so rather than supplying a reconstructed duration.

Table D.1. Human activity record

Date	Author	Activity	Supporting record	Duration
May 26, 2026	Ma	Initial recruitment and project scoping (preliminary; no research conducted)	Message record	Not separately recorded
May 26, 2026	Jiang	Reviewed proposal; discussed feasibility; contacted Zhu	Message record	Not separately recorded
May 26, 2026	Zhu	Discussed prospective background research role	Message record	Not separately recorded
June 5, 2026	Jiang	Submitted background materials to the agent; reviewed output	System timestamp	Not separately recorded
June 6, 2026	Jiang, Ma, Zhu	First joint author meeting; division of labor established	Meeting record	Not separately recorded
June-August 2026	Zhu	Background research, literature identification, bibliography and metadata work	Files and messages	See Table D.2
June-August 2026	Ma	Coefficient verification, classification, scope, and methodological rulings	Audit log (Appendix C); email	See Table D.2
June-August 2026	Jiang	Workflow orchestration, routing of rulings, state file and audit maintenance, revision coordination	AI and audit records	See Table D.2
2026	Atanassov	Advising the domain expert on finance-domain questions	Messages	Not separately recorded
August 2026	Jiang, Ma, Zhu	Manuscript review, correction, and final quality control	Drafts and email	See Table D.2

Notes: Dates and supporting records are as held in Exhibits D-B and D-E. The project did not maintain contemporaneous duration records for individual sessions. Author-level totals are reported in Table D.2 as retrospective estimates; we do not decompose them into session-level durations, since any such decomposition would be constructed rather than recovered.

Based on retrospective estimates, Zhipeng Jiang, Yao Ma, and Yanan Zhu contributed approximately 75 person-hours of human labor in total: Jiang approximately 30 hours, Ma approximately 30 hours, and Zhu approximately 15 hours. Julian Atanassov contributed in an advisory capacity to the domain expert; that advisory time was not separately estimated. These figures are retrospective estimates based on documented project communications, AI workflow sessions, human verification activities, literature work, manuscript review, file and state management, and submission preparation.

Table D.2. Human labor by author (retrospective estimates)

Author	Principal human contribution	Approximate hours
Julian Atanassov	Advising the domain expert on finance-domain questions	Not separately estimated
Zhipeng Jiang	Workflow design and orchestration, AI operation, task definition and sequencing, routing of rulings, state file and audit maintenance, revision and submission coordination	30
Yao Ma	Finance-domain judgment, coefficient verification, classification, scope and methodological rulings, source provision, final substantive review	30
Yanan Zhu	Economics-oriented background research, literature identification, bibliography and metadata work, summary tables, manuscript review and quality control	15
Total		approximately 75 (excluding advisory time)

Notes: Retrospective estimates, not logged times. Hours are stated to the nearest five and should not be read as precise measurements.

One feature of this accounting deserves emphasis. Human effort was concentrated in judgment rather than production. The domain expert wrote no manuscript prose and maintained no project files, yet issued the twenty-one rulings that determine what the database contains, and roughly 30 of the 75 hours were spent on that judgment. A short ruling can determine whether an AI-generated result is accepted, corrected, reclassified, or discarded. Conversely, the AI contribution should not be understated merely because its outputs were reviewed: where the system extracted evidence, organized results, drafted prose, or revised text, that work is reflected in the estimates in Section D.6.

D.5 Human decision and verification record

Human judgment was required whenever the agent encountered an issue that could materially affect the evidence base or its interpretation. Recurring categories included whether a coefficient belonged in the main evidence set; whether a dependent variable was directly MBS-related; whether a control-variable coefficient qualified; whether two estimates represented distinct outcomes; whether units or scaling had been read correctly; whether a source belonged in the main analysis or only in mechanism or background discussion; whether estimates could appropriately be pooled; and how identified extraction errors should be corrected and documented.

The complete record of these rulings is Appendix C, which lists twenty-one rulings with their content and database effect. Zero flags remain open.

The B008 extraction error, described in Section 8.3, is retained in the record as the clearest illustration of how the verification mechanism performed and where it initially failed. The agent extracted a coefficient with the correct value and the wrong dependent variable; the verification query named a table column without stating what that column measured; the ruling returned was accurate for the column named; and the mislabeled entry entered the database unchallenged. It was found when the published source was re-read directly against the database. The corrective protocol, requiring table, column or panel, and dependent variable to be stated together, subsequently caught a second defect of the same class in B004-X. We document both because the value of this record lies in what it shows about failure modes, not in an unblemished account.

D.6 Human and AI contribution accounting

The project was designed as an AI-agent-assisted workflow rather than a conventionally human-drafted paper with incidental AI editing. Table D.3 reports estimated shares of project output by category. These percentages measure production of output. They do not measure academic responsibility, which rests entirely with the human authors.

Table D.3. Estimated share of project output

Output category	Human	AI	Basis of estimate
Manuscript writing	5%	95%	The agent generated the majority of draft prose and revisions; humans supplied objectives, domain feedback, corrections, and final approval.
Literature processing and extraction	20%	80%	The agent performed most bulk reading, extraction, summarization, and organization; humans supplied sources, defined tasks, and verified evidence.
Code and analysis	30%	70%	The agent performed most structured evidence analysis and analytic support; humans determined acceptable interpretations, checked outputs, and made final rulings. No material standalone human coding contribution is identified in the record.
Documentation and appendices	20%	80%	The agent drafted and organized most documentation and archive structure; humans supplied source records, verified chronology, and approved the final account.
Overall project output	25%	75%	Estimate based on the AI archives, communication records, audit history, workflow orchestration, and documented human verification and judgment.

Notes: Estimates of output production, not of academic responsibility or authorship authority. Human authors retained final authority over source verification, methodological rulings, interpretation, error correction, inclusion and exclusion of evidence, and acceptance of material into the manuscript. The shares are stated at coarse intervals to avoid false precision and would be revised if the archive revealed material activity not represented in the reviewed records.

Table D.4 records where authority rested at each stage of the research, which is the dimension the percentages above do not capture.

Table D.4. Division of labor and final authority by research stage

Research stage	Human role	AI role	Final authority
Project conception	Selected opportunity, domain, and objectives	None or limited brainstorming	Human
Project scoping	Defined feasible directions and constraints	Generated and evaluated alternatives	Human
Literature collection	Identified and supplied relevant papers	Assisted with processing and screening	Human
Literature synthesis	Reviewed relevance and framing	Extracted, summarized, organized, compared	Human
Evidence extraction	Verified source evidence and open items	Performed initial structured extraction	Human
Classification	Made inclusion and interpretation rulings	Proposed classifications; generated flags	Human
Analysis support	Determined acceptable methodology and interpretation	Assisted with structured analysis	Human
Error detection and correction	Rechecked sources; approved corrections	Surfaced inconsistencies; revised outputs	Human
Manuscript structure	Directed argument; accepted or rejected revisions	Generated outlines, organization, draft text	Human
Manuscript drafting	Reviewed, corrected, refined	Generated a substantial share of draft prose	Human
References and formatting	Verified references and presentation	Assisted with organization and formatting	Human
Final submission	Reviewed and approved the complete submission	No independent authority	Human

D.7 Record limitations

This appendix rests on AI conversation exports, message and meeting records, email, manuscript drafts, bibliography files, the beta database, and the audit log. Exact system timestamps are used where available. Human working time was not recorded contemporaneously for every activity, and reconstructed durations are labeled as estimates throughout.

The project did not maintain a formal literature-screening log from the outset. Search and screening were interleaved rather than separated into logged stages, and running counts of records identified and excluded were not kept. We therefore report the number of context papers as approximate and present no screening flow diagram. This limitation is disclosed in Section 4.1 of the paper and is not reconstructed retrospectively here. The surviving records do not permit recovery of an exact classification count.

The contributor referred to as "Ruby" in informal project communications is Yanan Zhu. There was no additional contributor. All work on this project was performed by the four named authors.

D.8 Supporting exhibits

The following exhibits are submitted with this appendix. The conversation archive is retained as an exhibit rather than reproduced inline, given its length; the remaining exhibits hold the underlying records summarized above.

Exhibit	Contents
D-A	Complete AI conversation archive. Assembled Claude export records retaining conversation identifiers and Eastern-Time system timestamps, together with the ChatGPT project transcript, with metadata preserved where present in the source record. Project-related conversations are included in full rather than selectively.
D-B	Human activity time log. Chronological record of human activity by author, with supporting record type for each entry.
D-C	Human decision and verification audit. Cross-reference among audit flags, AI outputs, source evidence, human rulings, corrections, and final treatment in the manuscript.
D-D	Human and AI contribution accounting. Category estimates and the basis on which each was derived.
D-E	Supporting communication records. Dated records establishing project initiation, the June 1 compliance boundary, coordination, author roles, and major workflow decisions.