

Mapping Economic Theory with AI*

Elliott Ash¹, Andrea Ciccarone¹, and Sergio Galletta²

¹ETH Zurich

²Sapienza University of Rome

October 22, 2026

Abstract

We develop a method for mapping and measuring how economic theory papers reach their conclusions. An agentic AI system represents each paper as a claim graph, which shows how model assumptions and intermediate claims support the reported results. In a random sample of 190 theory papers from four leading economics journals, removing the most consequential starting input leaves an average of 91 percent of weighted headline and major claims unsupported. Next, we benchmark error discovery against published corrections from 100 papers and use claim graphs to assess the consequences of repair. The graphs trace which conclusions depend on mistaken derivations and measure how much of their economic content is lost after correction. While errors in theory papers tend to affect support for a large share of a paper's conclusions, they rarely change the main economic conclusions after the smallest valid repair.

*Contact: Elliott Ash (ashe@ethz.ch), Andrea Ciccarone (corresponding author, aciccarone@ethz.ch), and Sergio Galletta (sergio.galletta@uniroma1.it). We thank Wei Pin and David Zollikofer for helpful research assistance.

1 Introduction

Economic theory provides predictions and welfare comparisons that economists use to interpret evidence, evaluate policies, and develop new models. Theorists assess these contributions through shared norms concerning assumptions, logical coherence, and economic interpretation (Rubinstein, 2006; Gilboa et al., 2014). Assessing these contributions requires understanding which assumptions support the main conclusions and whether those conclusions survive a correction to the argument. Yet these judgments have been difficult to analyze consistently and at scale, not least because of the substantial cost of expert assessment.

Recent advances in agentic AI make such assessments feasible at larger scale. In empirical economics, recent advances in AI-assisted data analysis offer a path to expanding reproduction and verification (Kohler et al., 2026; Willner and Yanagizawa-Drott, 2026; Schwartz et al., 2026). In mathematics, recent applications demonstrate the potential of AI to analyze complex formal arguments, including work on Navier–Stokes singularities and zeros of the Riemann zeta function (OpenAI, 2026; Claude, 2026). Can these tools also help assess the arguments supporting economic theory?

This paper develops a framework for tracing how economic conclusions depend on their supporting arguments and assessing how those conclusions change after correction. Using frontier language models available as of September 2026, our agent system reconstructs each paper as a *claim graph* connecting assumptions, intermediate results, and economic conclusions through their supporting derivations. The graph records which premises are required jointly and which derivations provide alternative support, allowing us to measure how conclusions depend on shared inputs and chains of reasoning. Separate automated audits identify potential errors, which are linked to the graph to trace which conclusions lose their justification. A separate repair stage assesses which conclusions survive correction, distinguishing network fragility (i.e., the share of weighted claims that lose their recorded support) from substantive severity (i.e., the share of weighted qualitative content assessed as lost after repair).

We apply the framework to two samples with distinct purposes. A random sample of 190 theory papers published from 1995 onward in four leading economics journals allows us to describe argument structure and assess the credibility of central premises. A separate sample of 100 papers with published corrections provides documented errors against which to benchmark discovery and examine the consequences of repair. Both applications use graphs reconstructed from the original articles, with published corrections withheld from graph extraction and error discovery.

The random-sample graphs reveal layered arguments with extensive dependence on shared inputs. The longest chain to a headline claim averages 5.3 derivation steps, and removing the most consequential starting input leaves an average of 91 percent of weighted headline and major claims

unsupported.¹ Only 23 of the 190 papers record any alternative derivation. These patterns reveal how dependence on shared inputs can leave many economic conclusions vulnerable to a single error.

Dependence on shared inputs need not imply reliance on unusually weak premises. To examine this possibility, a separate language model assessment provides a qualitative judgment on how well justified substantive premises are for a paper’s question. For example, a relatively justifiable premise could be weak assumptions on preferences, such as transitivity. A less justifiable premise could be a strong assumption on functional form, such as a log utility function. We find that, in general, the most load-bearing premises are not outliers in terms of justifiability and if anything are somewhat more justifiable. This offers some reassurance that economic theories do not systematically rest their main conclusions on their least defensible premises.

Next we explore the importance of errors in economic theory by asking how much of a paper’s economic contribution changes when a faulty step is corrected. An example of a paper where repairs do not affect the main conclusions is [Rothschild and Stiglitz \(1970\)](#). Our error repair agent provides a proof correction similar to [Leshno et al. \(1997\)](#), which preserves the economic conclusions despite the loss of their original support. An example of a paper where the errors are severe, in the sense that headline claims do not survive repair, is [Brown \(1983\)](#). Our agent identifies the distribution-preserving rotation documented by [Benkard and Berry \(2006\)](#), and by using the claim graph we show that the derivative-based identification conclusions do not survive the model-suggested repair. These examples illustrate why exposure to a mistaken step must be distinguished from the qualitative content lost after repair.

Published corrections provide a benchmark for assessing whether the audit can recover documented errors. After graph extraction and error discovery are complete, findings are compared with the corrections under exact and partial matching rules. The primary audit fully recovers 79 of 168 documented errors, or 47 percent, with recovery rising to 115 errors, or 69 percent, when partial matches are included. Our model finds many more errors in this sample however – 473 in total – most of which are less significant than those identified in the published correction. AI error findings that affect a larger share of the claim graph are more likely to match a correction. That suggests our measure of fragility could be a promising way to prioritize the claims that peer reviewers should examine.

The presence of errors need not undermine a paper’s scientific contribution if the arguments can be repaired while preserving the main economic conclusions. To understand the impacts of errors, we measure their severity – comparing the share of conclusions that lose support because of an error with the substantive changes required to repair it. Among findings matched to published

¹Headline claims receive weight one and major results one-half, following the prominence assigned to them by the authors. Starting inputs can include definitions and imported results as well as substantive assumptions.

corrections, median fragility is 28 percent, while median severity after repair is only 4 percent. A similar pattern emerges in the random sample of 190 theory papers, where median fragility is 14 percent and median severity after repair is just 2 percent. These results suggest that while errors in economic theory can undermine large parts of the original argument, they rarely change the headline conclusions.

Finally, we look descriptively at what features of argument structure are associated with the severity of errors. We run regressions using structural features as covariates with a binary outcome variable equaling one for the presence of a severe error, defined as having a post-repair loss of 10 percent or more. We find that a higher maximum load (a larger weighted share of headline and major claims losing support when one starting input is removed) is associated with a higher probability of a severe error. An increase in result yield (formal results per substantive input) is associated with a lower probability of severe error. Derivation depth has no clear association. These patterns suggest that reviewers could prioritize scrutiny of premises supporting a large share of a paper’s economic conclusions.

To support further research on the structure and reliability of economic theory, we introduce [theorize.ink](#), an interactive platform for exploring theoretical arguments. Readers can navigate a paper’s argument, trace a conclusion back to its supporting assumptions and intermediate results, and identify which further conclusions depend on a particular claim. Each statement is linked to its source passage, making the reconstruction directly inspectable against the original paper. Researchers can request graphs of additional papers and provide feedback on existing reconstructions. We also release the dataset of claim graphs in a common format for systematic comparisons of argument structure and further research on the reliability of economic theory.

This paper contributes to the new literature on measuring and evaluating economic theories. [Fudenberg et al. \(2022\)](#) measure how much predictable variation economic models capture, while [Fudenberg et al. \(2026\)](#) quantify how much hypothetical behavior they rule out. This quantitative approach complements earlier methodological work on the role of assumptions, interpretation, and reasoning in economic theory ([Rubinstein, 2006](#); [Gilboa et al., 2014, 2018](#)). We examine the internal structure of theoretical arguments, measuring how conclusions depend on supporting premises and derivations and assessing what survives when an argument requires correction.

We also contribute to the science-of-science literature studying research practices in economics. [Hamermesh \(2013\)](#) and [Angrist et al. \(2017\)](#) document changes in research styles, while [Currie et al. \(2020\)](#) use text mining to track methods and data sources. [Oliveira and Dávila-Fernández \(2020\)](#) study the evolution of mathematical and quantitative methods in leading economics journals. Most closely related in method, [Garg and Fetzer \(2026\)](#) use LLMs to construct graphs of relationships between economic concepts, distinguishing claims by their evidentiary basis, and study how graph structure relates to publication and citation outcomes. We contribute to this literature by measuring

the dependencies within theoretical arguments, connecting starting premises and intermediate results to the economic conclusions they support. Our graphs distinguish inputs required jointly from alternative derivations that can preserve support when one argument fails.

Finally, the paper contributes to the literature on AI for science, particularly the reproduction, verification, and generation of research. [Kohler et al. \(2026\)](#) study whether agents can reproduce empirical results from methods descriptions and original data, while [Schwartz et al. \(2026\)](#) develop an LLM workflow to reproduce, improve, and extend published economics research. [Willner and Yanagizawa-Drott \(2026\)](#) generate 1,000 policy-evaluation papers using AI and identify reliable verification as the bottleneck to research production at scale. They benchmark automated verifiers against controlled error injections to assess whether verification can keep pace with generation. For theoretical research, [Toda \(2026\)](#) examines the identification and correction of errors in published papers, while [Koren \(2026\)](#) compares protocols for producing and checking theoretical arguments.² We contribute a framework for tracing the consequences of an identified error and distinguishing a failure of the original argument from a loss of economic content after repair. Because graph construction is independent of error discovery, the representation can be used with findings from automated audits or expert review.

Section 2 presents the claim-graph representation and the measures of argument structure and consequences of errors and repairs. Section 3 describes the samples and graph construction, then examines argument structure and premise credibility. Section 4 presents the audit procedure, validation against published corrections, and results on errors and repair. The final section concludes.

2 Framework

We use claim graphs to describe how economic conclusions depend on premises and derivations. The framework distinguishes structural dependence from premise credibility, and the support threatened by an error from the qualitative content lost after repair. We first define the representation and structural measures, then introduce premise credibility, error linkage, and repair.

2.1 Representation

We record which assumptions and intermediate results support each economic conclusion in a theory paper. The representation distinguishes premises required together from alternative derivations that can independently establish the same result. The graph records the support presented in the paper. It does not establish that a premise is logically necessary or that no alternative proof exists.

²Related work examines LLM-based peer review in economics ([Pataranutaporn et al., 2025](#)) and error detection and repair in other fields ([Son et al., 2025](#); [Bianchi et al., 2025](#)).

Formally, we represent the paper’s argument as a typed directed hypergraph $G = (V, E)$. Each node $v \in V$ is an atomic statement, such as an assumption, definition, lemma, proposition, or substantive claim. Each hyperedge $e = (S_e, t_e) \in E$ records a derivation, which we call a route. Its source set $S_e \subseteq V$ contains the statements jointly used to derive its target $t_e \in V$. Using a hypergraph allows us to represent a derivation that requires several statements together. Only derivations used to establish a statement enter E . Links that merely illustrate, motivate, or frame a statement do not.

For each statement v , let $\text{In}(v) = \{e \in E : t_e = v\}$ denote its incoming routes. A premise is a statement with no incoming route that the argument uses as a starting point, taking it as given rather than deriving it. Statements without an incoming route that only motivate or frame the analysis, such as research questions, or that the paper asserts without support are not premises. A derived statement has at least one incoming route; when several routes are recorded, each represents an alternative derivation sufficient to establish the statement over its full stated domain. We require the graph to be acyclic, so that no statement depends, directly or indirectly, on itself.

Figure 1 illustrates the representation with a fragment from [Rothschild and Stiglitz \(1970\)](#). The paper establishes the equivalence of three criteria for ranking distributions by riskiness: every risk-averse agent prefers F to G ; G can be obtained by adding noise to F ; and G places more weight in the tails. Lemma 1 is an intermediate step in this argument. Its derivation uses integral conditions, two definitions, and a scope condition. These four statements jointly support Lemma 1. The lemma then combines with an approximation result to derive a proposition on approximation by spread pairs.

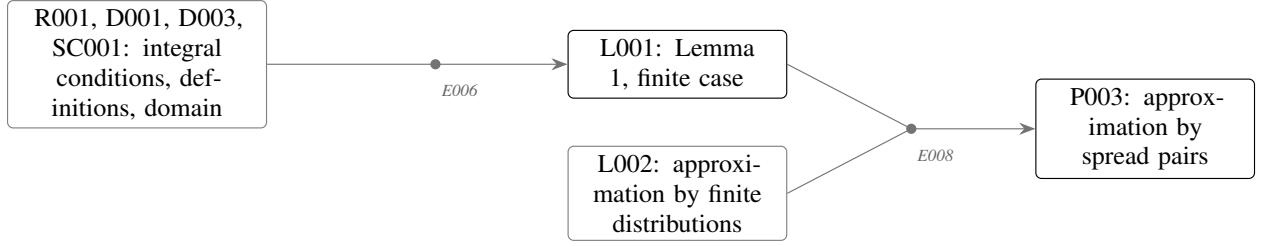
The example shows how a result can depend on several statements together, but a paper may also provide more than one derivation of the same result. Each derivation requires its source statements jointly (AND), while multiple incoming hyperedges represent alternative derivations, any one of which is sufficient to establish the conclusion (OR). If a required premise loses support, that derivation can no longer support its conclusion. A recorded alternative derivation may nevertheless preserve support for the conclusion.

Each node has a semantic type, a contribution role, and a premise status. Type identifies the kind of statement, such as an assumption, definition, lemma, or welfare result. We distinguish 23 types.³ A formal proposition and the economic claim derived from it are separate nodes linked by a derivation.

Contribution role records a statement’s place in the paper’s advertised contribution. Headline claims express the paper’s main contributions, major claims are important results supporting them, and supporting nodes provide assumptions, definitions, and intermediate steps. Auxiliary nodes

³The complete list: research question, primitive, definition, assumption, parameter restriction, equilibrium concept, formal condition, lemma, proposition, corollary, existence result, uniqueness result, equilibrium selection, comparative static, welfare result, mechanism claim, substantive claim, interpretation claim, policy claim, scope condition, extension result, illustration, external input. Figure 9 in Appendix B reports how often each occurs in our sample of papers.

Figure 1: Rothschild–Stiglitz claim-graph fragment



Note: One fragment of the 30-statement graph for [Rothschild and Stiglitz \(1970\)](#), with statements as nodes and derivations as hyperedges. R001 denotes the integral conditions, D001 and D003 definitions, and SC001 the scope condition. Hyperedge E006 derives L001 (Lemma 1, finite case) from these four statements. E008 combines L001 with L002 (approximation by finite distributions) to derive P003 (approximation by spread pairs). Black dots join sources used jointly, and arrows point to derived statements. Data source: Claim graphs extracted from the original articles.

contain illustrations and extensions outside the main contribution, while contextual nodes motivate or frame the analysis. Premise status records whether a statement is taken as given rather than derived. These classifications need not coincide: a lemma may be a headline contribution or a supporting result, and a starting premise need not be an assumption. The structure statistics use types to distinguish formal results from primitive inputs.

2.2 Argument structure

Claim graphs allow us to compare how papers build their contributions from assumptions and intermediate results. We use this structure to count the formal results derived per substantive primitive input and to trace which inputs and derivations support the main claims. This subsection defines these descriptive measures, then introduces interventions that remove part of the argument and the fragility measure built on them.

We use the following notation. Let $n = |V|$ and $m = |E|$ denote the numbers of statements and derivations respectively, and let H denote the set of headline claims. Let $\mathcal{T}_P$ contain the semantic types primitive, assumption, parameter restriction, and scope condition. We define the set of primitive inputs as

$$P = \{v \in V : \text{type}(v) \in \mathcal{T}_P\},$$

so P contains the paper’s substantive model inputs, assumptions, restrictions, and scope conditions. A premise, as defined in Section 2.1, is any statement the paper takes as given rather than derives. Definitions and imported results may therefore be premises in the graph, but they do not enter P unless they are substantively typed as primitive inputs. Let $R \subseteq V$ collect formal results (lemmas, propositions, corollaries, existence and uniqueness results, comparative statics, welfare and extension results).

Definition 1 (Yield). Result yield is $Y = |R|/|P|$, the number of formal results per primitive input. Headline yield is $Y_H = |H|/|P|$, the number of headline claims per primitive input.

A high yield indicates that a paper develops many formal results from relatively few substantive inputs. This can reflect several deductions from a common model, but also differences in how authors divide an argument into lemmas and propositions. Yield therefore describes the organization of the argument, rather than the novelty or economic importance of its conclusions.

Definition 2 (Density and fan-in). Derivation density is $\delta = m/n$, the number of derivations per statement. Mean fan-in is $\bar{\phi} = m^{-1} \sum_{e \in E} |S_e|$, the average number of source statements used jointly in a derivation.

Fan-in describes the number of statements used jointly in a single step. For example, a comparative-static result might jointly use an equilibrium condition, a parameter restriction, and a lemma. This differs from depth, defined next, which counts successive derivation steps.

Definition 3 (Depth). Let $d(v)$ be the number of hyperedges on the longest path from a premise to v . Derivation depth is $D_{\max} = \max_{h \in H} d(h)$, and mean headline depth is $\bar{D} = |H|^{-1} \sum_{h \in H} d(h)$.

Depth measures how many successive derivations connect a starting premise to a headline claim along the longest path. A welfare conclusion, for example, may follow from an equilibrium characterization that itself depends on an existence result. Depth records these layers of reasoning, without measuring the mathematical difficulty of any individual step.

For each statement v , let $\text{Anc}(v)$ contain the statements that precede it along at least one directed path in the graph. We refer to these as its ancestors. The union $\text{Anc}(H) = \bigcup_{h \in H} \text{Anc}(h)$ collects the ancestors of all headline claims.

Definition 4 (Redundancy and coverage). Route redundancy ρ is the share of derived nodes with at least two derivations that each establish the full statement on its stated domain. Headline coverage is $\kappa = |\text{Anc}(H) \setminus H|/(n - |H|)$, the share of non-headline statements that are ancestors of at least one headline claim.

Redundancy records whether a claim has more than one presented derivation. If one derivation fails, another can preserve support for the claim, provided its own premises and intermediate steps remain supported. Two derivations that share a faulty premise may therefore fail together. Coverage measures the share of non-headline statements that contribute, directly or indirectly, to deriving the headline claims. Statements outside these paths may develop examples, extensions, or other results. The measure describes how much of the recorded argument supports the advertised main contribution.

Interventions. We ask which conclusions would lose their recorded support if a premise were withdrawn or a derivation found invalid. We represent this intervention by removing a set of statements $D \subseteq V$ and a set of derivations $X \subseteq E$ from the graph. A statement with no incoming route remains supported unless it is removed. A derived statement remains supported if at least one of its derivations survives and all statements used in that derivation remain supported.

Let $s_{D,X}(v) \in \{0, 1\}$ denote the support of statement v after the intervention. Define

$$(1) \quad s_{D,X}(v) = \begin{cases} 0 & v \in D, \\ 1 & v \notin D, \text{In}(v) = \emptyset, \\ \max_{e \in \text{In}(v) \setminus X} \min_{u \in \mathcal{S}_e} s_{D,X}(u) & \text{otherwise,} \end{cases}$$

where the maximum over an empty set is zero. Acyclicity makes the recursion well defined. The set $\text{In}(v)$ refers to the original graph, so deleting every incoming derivation leaves a derived statement unsupported. With no intervention, every statement has support one. After intervention, zero denotes loss of recorded support, rather than falsity.

Figure 2 illustrates the removal of the derivation of Lemma 1 in [Rothschild and Stiglitz \(1970\)](#). The lemma loses support, as do the propositions that use it to establish that the integral conditions imply the noise criterion. Without this part of the proof, the equivalence theorem and the three substantive claims derived from it are no longer supported. The result linking riskiness to variance remains supported because its derivation does not use the lemma. Overall, seven of the graph’s 30 statements lose support, including three of its four headline claims. These losses would be smaller if the paper provided an alternative derivation that did not rely on the lemma.

Definition 5 (Network fragility). Let

$$(2) \quad w_v = \begin{cases} 1 & \text{headline,} \\ \alpha & \text{major,} \\ 0 & \text{otherwise,} \end{cases} \quad W = \sum_v w_v,$$

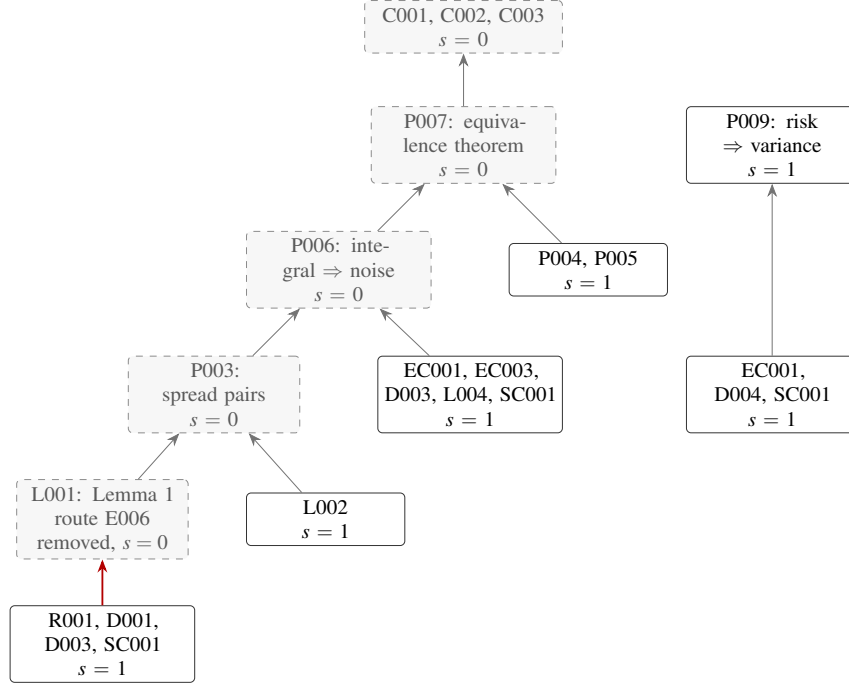
where $W > 0$ by construction and $\alpha \in (0, 1)$ is the weight of a major result relative to a headline claim; we set $\alpha = \frac{1}{2}$ throughout. The network fragility of an intervention (D, X) is

$$(3) \quad F(D, X) = \frac{100}{W} \sum_{v: w_v > 0} w_v [1 - s_{D,X}(v)],$$

the weighted share of headline and major claims that lose support.

Fragility ranges from zero, when every headline and major claim remains supported, to 100,

Figure 2: Support after removing the Lemma 1 derivation



Note: One intervention on the 30-statement Rothschild–Stiglitz graph. Removing E006, the derivation of L001 (Lemma 1), leaves seven statements unsupported, including three of four headline claims. Support $s = 0$ means a statement loses recorded support relative to the baseline, and $s = 1$ means it retains that support under the intervention. Identifiers refer to integral conditions (R001), definitions (D001, D003, D004, EC001, EC003), a scope condition (SC001), lemmas (L001, L002, L004), propositions (P003–P007, P009), and substantive claims (C001–C003). P009, linking riskiness to variance, remains supported. The red arrow marks the removed derivation. Dashed grey boxes are unsupported and solid boxes supported. Other arrows show dependencies. Appendix M shows the same intervention on the full graph. Data source: claim graphs extracted from the original articles.

when none does. Only headline and major claims carry weight. Other statements matter through the weighted claims that depend on them. The weights reflect prominence in the paper, rather than economic value. Fragility therefore measures how much of the advertised contribution loses its recorded support when an assumption or proof step is withdrawn (Rubinstein, 2006; Gilboa et al., 2014). In Figure 2, the unsupported claims carry weight 4.5 out of a total of 10, giving $F = 45$. This is exposure before repair. A corrected argument may recover some or all of these conclusions.

For the load statistics, we separately define the active primitives

$$P_A = \{v \in V : \text{In}(v) = \emptyset \text{ and } v \in S_e \text{ for some } e \in E\}.$$

An active primitive has no incoming derivation and enters at least one downstream derivation. This is a structural category: unlike P , it can include definitions and imported results as well as substantive assumptions. We therefore refer to these nodes as starting inputs when interpreting load.

Definition 6 (Active-primitive load). For each active primitive $a \in P_A$, let $F_a = F(\{a\}, \emptyset)$ as defined

in (3). Maximum load is $L_{\max} = \max_{a \in P_A} F_a$, and mean load is $\bar{L} = |P_A|^{-1} \sum_{a \in P_A} F_a$. The load share λ is the fraction of active primitives whose removal leaves at least one headline claim unsupported. Load concentration is the Herfindahl index of active-primitive loads:

$$\sum_{a \in P_A} \left(\frac{F_a}{\sum_{b \in P_A} F_b} \right)^2,$$

defined when total load is positive.

A starting input’s load is the weighted share of headline and major claims that lose support when it is removed. Maximum load is the largest of these shares. A value of 100 means that removing one starting input leaves every headline and major claim unsupported. Load share measures the fraction of starting inputs whose removal leaves at least one headline claim unsupported, and so whether headline support depends on a few inputs or on nearly all of them. Concentration measures how unevenly load is distributed across starting inputs. If every headline and major claim requires each of several inputs, those inputs have equal loads. Concentration is then low, even though removing any one of them leaves all these claims unsupported.

2.3 Premise credibility

Structural dependence and premise credibility capture different properties. We ask whether the premises carrying the most load are less defensible than other premises in the same paper. Local nonsatiation, for example, is a general and often credible premise about preferences. A specific functional form such as logarithmic utility is less defensible as a basis for a general claim, unless it is used for illustration and the authors also provide a more general derivation. We assess the defensibility of substantive premises on a 1–10 scale, considering their coherence, motivation, and scope. We call this rating the credibility score. Definitions, identities, imported theorems, and derived results are excluded.

2.4 Error linkage

A finding is a proposed error in the paper. Each finding must identify the relevant passage and provide supporting evidence, such as a counterexample, a rederivation, or an internal contradiction. For each finding f , we identify the smallest set of statements or derivations that would be directly affected if the proposed error were valid. Statements assessed as false and derivations assessed as invalid are removed. Those assessed as valid only on a narrower domain are marked as qualified. Findings that cannot be mapped unambiguously to the graph are excluded from graph-based consequence scoring, rather than assigned zero loss.

Let (D_f, X_f) denote the removed nodes and hyperedges, and (D_f^q, X_f^q) the qualified nodes and hyperedges. Using the fragility measure in (3), define

$$(4) \quad F_f^- = F(D_f, X_f), \quad F_f^+ = F(D_f \cup D_f^q, X_f \cup X_f^q).$$

The lower endpoint retains qualified statements and derivations, while the upper endpoint removes them. For example, a comparative-static result may remain valid only under a stronger parameter restriction. The interval records the support at stake when such qualified parts of the argument are retained or withdrawn.

When the evidence supports a scope-loss fraction $\phi_f \in [0, 1]$, we summarize the interval by

$$F_f = F_f^- + \phi_f(F_f^+ - F_f^-).$$

This interpolation assigns a fraction ϕ_f of the additional loss associated with the qualified statements and derivations. When no such fraction can be justified, we report the interval $[F_f^-, F_f^+]$.

Several findings may remove support from the same claim. To compute paper-level fragility, we apply all findings' removals and qualifications to the graph together and recompute which claims lose support. This counts each affected claim once and captures losses that arise only when findings act jointly.

2.5 Repair and substantive severity

Fragility identifies the claims exposed to an error. To measure its substantive consequences, we ask whether those claims can be recovered after correction. We fix the set of claims to reassess before considering a repair:

$$(5) \quad R_f = \{v \in V : w_v > 0, s_{D_f \cup D_f^q, X_f \cup X_f^q}(v) = 0\}.$$

The target set includes every headline and major claim exposed by the upper intervention, including claims that lose support only when qualified statements or derivations are withdrawn. Fixing it before repair ensures that each potentially affected conclusion is assessed, even when the repair ultimately preserves it.

The protocol seeks the weakest valid repair, meaning the least extensive justified change that repairs the argument. It considers a local formula correction first, then a proof rewrite, a justified change in assumptions, and finally a change to the model. This ordering guides the search without guaranteeing the smallest possible repair. An additional assumption must have a justification beyond recovering the desired conclusion. For each affected claim, we record whether the repair preserves

it, changes it, restricts its domain, or leaves it unresolved.

Let $q_f(v)$ denote the share of claim v 's qualitative content lost after repair. Qualitative changes include changes in sign, direction, existence, ordering, or mechanism. An unchanged or purely quantitatively changed claim receives zero loss. A change in the magnitude of a comparative-static effect alone is one such case. A reversed or false claim receives full loss unless confined to a justified fraction of its domain. A justified claim-level fraction of scope lost after repair is denoted ϕ_v , distinct from the finding-level fraction ϕ_f used for fragility. A claim surviving on a narrower domain receives this fraction, or $[0, 1]$ if no fraction is justified. An unresolved claim receives $[0, \phi_v]$ if its potentially affected scope is bounded, or $[0, 1]$ if that fraction is unknown. Appendix Table 5 gives the full rule.

Substantive severity is

$$(6) \quad Q_f = \frac{100}{W} \sum_{v \in R_f} w_v q_f(v),$$

where the weights and total weighted claim count W are defined in (2). Claims outside R_f contribute zero. When claim-level losses are intervals, lower and upper severity scores use their respective endpoints. These intervals record unresolved repair or scope judgments, rather than sampling uncertainty. Because $0 \leq q_f(v) \leq 1$ and R_f is fixed by the upper intervention, every severity endpoint satisfies $0 \leq Q_f \leq F_f^+$.⁴

Severity measures the assessed loss of qualitative content after repair as a percentage of the paper's total weighted claims. It can be zero even when fragility is high, if a corrected proof recovers every affected conclusion. A score of $Q = 10$ represents loss of one-tenth of this weighted content, and $Q = 25$ one-quarter. These shares describe the paper's qualitative contribution, rather than welfare losses or the fraction of statements containing errors. We do not sum severity across findings because their repairs may interact. For paper-level summaries, we report the maximum severity across scored findings, set to zero when none is scored, and whether any finding reaches the stated severity threshold.

3 Structure of published theory

We use claim graphs to examine how published theory papers build their economic conclusions. We then assess whether the premises on which these conclusions depend most heavily are well justified. We first describe the samples and graph construction, then compare argument structures and assess premise credibility.

⁴For deletion-only interventions, $F_f^+ = F_f$. The bound need not hold relative to an interpolated or midpoint fragility score.

3.1 Samples

We map theories in two samples of economics papers. First, we have 190 theory papers published from 1995 onward in the American Economic Review, Journal of Political Economy, Quarterly Journal of Economics, and Review of Economic Studies. We draw from EconLit records classified as theory with full text available, stratifying by journal and publication period.⁵ We screen titles, abstracts, and numbered results to retain papers that develop and analyze a formal model. Papers that do not meet this criterion are replaced by draws from the same journal-period cell.⁶

The corrected sample contains 100 theory papers with published corrections, spanning publication years from 1953 to 2022. These are the first 100 entries in the correction catalogue for which publisher versions of both the article and correction are available (Appendix D). The sample does not overlap with the random sample, and Econometrica and the Journal of Economic Theory are its largest sources. We use the original article for extraction and error discovery and reserve the correction for subsequent evaluation.

3.2 Graph construction

We extract claim graphs for both samples. A separate exercise assesses the credibility of substantive premises in the random sample. Producing these measurements requires extensive model reasoning and large token budgets, with a full workflow taking several hours per paper. The computational demands reflect the detailed reconstruction and scrutiny of arguments that would otherwise require sustained work by human experts. Although intensive in tokens and elapsed time, the workflow offers a way to produce these assessments at a small fraction of the cost of equivalent expert labor. This cost advantage makes it feasible to apply a common protocol across a corpus of theory papers and to revisit the measurements as models and protocols improve.

Graph extraction. Following the representation in Section 2.1, extraction starts from the paper’s main claims and traces their supporting arguments backward through intermediate results to the starting inputs. A policy conclusion, for example, may follow from a welfare result that in turn relies on an equilibrium characterization and assumptions about preferences or information. Each statement and derivation cites the passage, equation, or numbered result where it appears in the paper.

⁵The sample contains 49 AER papers, 46 JPE papers, 48 QJE papers, and 47 ReStud papers. The final random sample contains 22 papers for 1995–1999, 62 for the 2000s, 56 for the 2010s, and 50 for 2020 onward.

⁶Replacement draws require at least two numbered formal results and exclude comments and replies. When a journal-period cell is empty, we draw from the neighboring period. The final-run manifest, `corpus/final_run_manifest_190.json`, records one validated graph and one complete audit–linkage–repair–score artifact set for each retained paper.

The extraction is done by prompting an agentic AI system with the paper as input. The agent system – OpenAI Codex running GPT 5.6 Sol – works in an isolated working environment containing only the paper’s PDF and a separate file containing the extracted plain text. An automated session of the agent is prompted with the instruction from Appendix N.1 for each paper.

The model records the paper’s stated derivations without certifying their validity. Missing derivations and unclear dependencies are flagged rather than filled in, and no alternative proofs are added. Outputs are checked against the data format and acyclicity requirements in Appendix B.

Premise-credibility assessment. A separate GPT-5.6 Sol pass scores each eligible substantive premise—such as assumptions about preferences, information, or institutions—on the 1–10 credibility scale defined in Section 2.3. Each premise is evaluated relative to the paper’s question, so a tractable simplification can receive a high score when its analytical purpose is clear and the paper’s conclusions respect its limitations. For each premise, the adjudicator records the authors’ stated justification, explains its rating, and identifies the main limitation, citing the relevant passage. The adjudicator receives the paper and premise statements in randomized order, but is not shown the extracted graph or its computed load measures.

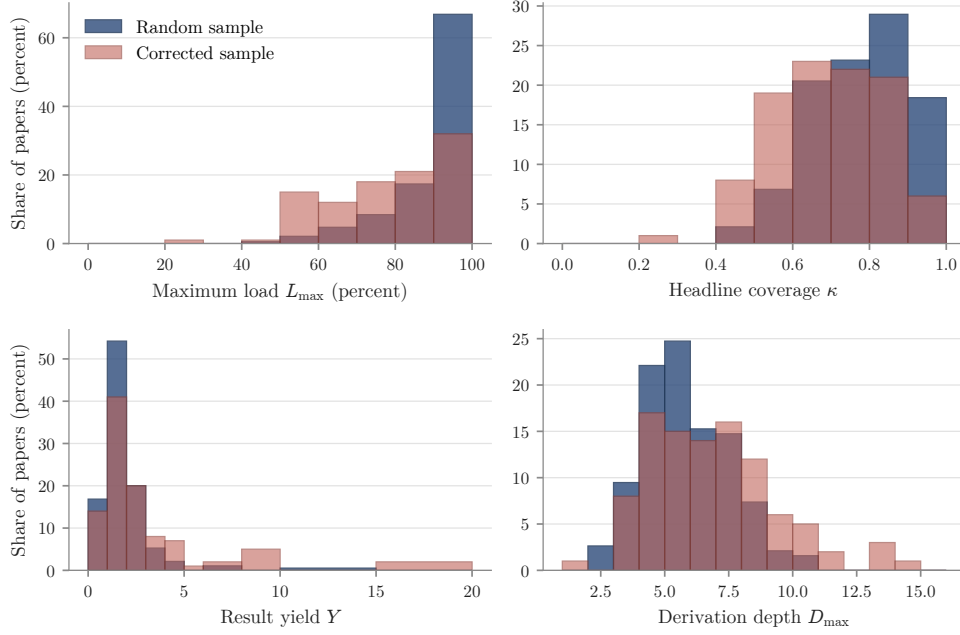
3.3 Argument structure

Figure 3 and Appendix Table 13 summarize how the papers build their arguments. In the random sample, the average graph contains 28.6 statements and 16.9 derivations, and the longest chain to a headline claim averages 5.3 steps. Seventy-seven percent of non-headline statements support at least one headline claim. Removing the most consequential starting input leaves 91 percent of weighted headline and major claims unsupported on average, while only 23 of the 190 papers record an alternative derivation. The conclusions are reached through several layers of reasoning, but they often share the same foundations. A challenge to a central premise can therefore require reassessing several economic conclusions together.

Corrected papers derive more formal results per substantive primitive input than random-sample papers (2.70 versus 1.71) through longer chains of reasoning (6.4 versus 5.3 steps). They have lower headline coverage (69 versus 77 percent) and maximum load (79 versus 91 percent). All four differences are statistically significant at the 1 percent level. The samples also differ in journal composition, publication period, and selection into a catalogue of published corrections. The differences in graph structure may reflect these structural differences rather than having a correctable error.

See Appendix Table 14 and Figure 12 for the corresponding breakdowns by journal and publication period. Mean maximum load is high across the four random-sample journals, ranging from 87

Figure 3: Distributions of argument structure



Note: Paper-level histograms for 190 random-sample and 100 corrected-sample papers. Result yield Y is formal results per substantive primitive input (primitives, assumptions, parameter restrictions, and scope conditions). Derivation depth $D_{\max}$ is the longest chain of derivations from a premise to a headline claim. Headline coverage κ is the share of non-headline statements that are ancestors of at least one headline claim. Maximum load $L_{\max}$ is the largest percentage of weighted headline and major claims that lose support after removing one starting input (an active primitive with no incoming derivation that enters a downstream derivation). Headline claims carry weight 1, major claims weight $\alpha = \frac{1}{2}$, and other statements weight 0. Bars report within-sample shares, with the random sample in navy and the corrected sample in red. Data source: claim graphs extracted from the original articles.

to 93 percent. We next examine how well the premises carrying the greatest load are justified.

3.4 Load-bearing premises and credibility

We compare the credibility of the most load-bearing premises with that of other premises in the same paper, asking whether greater structural dependence is associated with weaker foundations. The blinded assessment in Section 3.2 yields scores for 1,533 rateable substantive premises from the 1,954 starting inputs in the 190 random-sample papers. Eligibility follows the criteria in Section 2.3.⁷

Table 1 illustrates the assessment. In *Declining Search Frictions, Unemployment, and Growth*, discounting faster than productivity growth receives 9/10 because it keeps present values finite. In *Rational Debate and One-Dimensional Conflict*, selecting informative communication receives 4/10

⁷The inventory stores each graph’s SHA-256 hash; all 190 hashes match the canonical final-run manifest. In two independent passes on a balanced 20-paper calibration sample, rateability agreement is 98.3 percent ($\kappa = 0.956$), and 98.9 percent of jointly rateable scores differ by at most one point. Scores remain ordinal model assessments rather than expert judgments.

Table 1: Premise credibility and network load: four illustrations

	Higher credibility: 9/10	Lower credibility: 4/10
Higher load	Finite present values. In <i>Declining Search Frictions, Unemployment, and Growth</i> (2020), discounting exceeds productivity growth. Why 9/10? Future payoffs must be discounted fast enough for their present value to be finite. The restriction has a direct economic and mathematical purpose in the stated balanced-growth exercise. <i>Load: 83.3%; credibility: 9/10.</i>	Selecting informative debate. In <i>Rational Debate and One-Dimensional Conflict</i> (2000), informative communication is selected whenever sustainable. Why 4/10? Agents can also sustain uninformative talk. The paper assumes they choose revelation whenever possible without explaining what coordinates that choice; the assumption selects whether debate changes beliefs. <i>Load: 87.5%; credibility: 4/10.</i>
Lower load	Allowing correlated valuations. In <i>Rational Inattention, Optimal Consideration Sets, and Stochastic Choice</i> (2019), an extension allows arbitrary correlation across valuations. Why 9/10? Valuations can move together. Allowing this avoids an independence restriction, and the paper explains how correlation changes attention and choice. The broader domain fits the question. <i>Load: 19.2%; credibility: 9/10.</i>	Assuming a unique crisis equilibrium. In <i>Liquidity and the Threat of Fraudulent Assets</i> (2012), crisis comparative statics assume uniqueness even though the extension permits multiplicity. Why 4/10? The model itself permits multiple equilibria, but the paper does not show when they disappear. Assuming uniqueness sets aside an alternative crisis mechanism rather than explaining why it can be ignored. <i>Load: 20.0%; credibility: 4/10.</i>

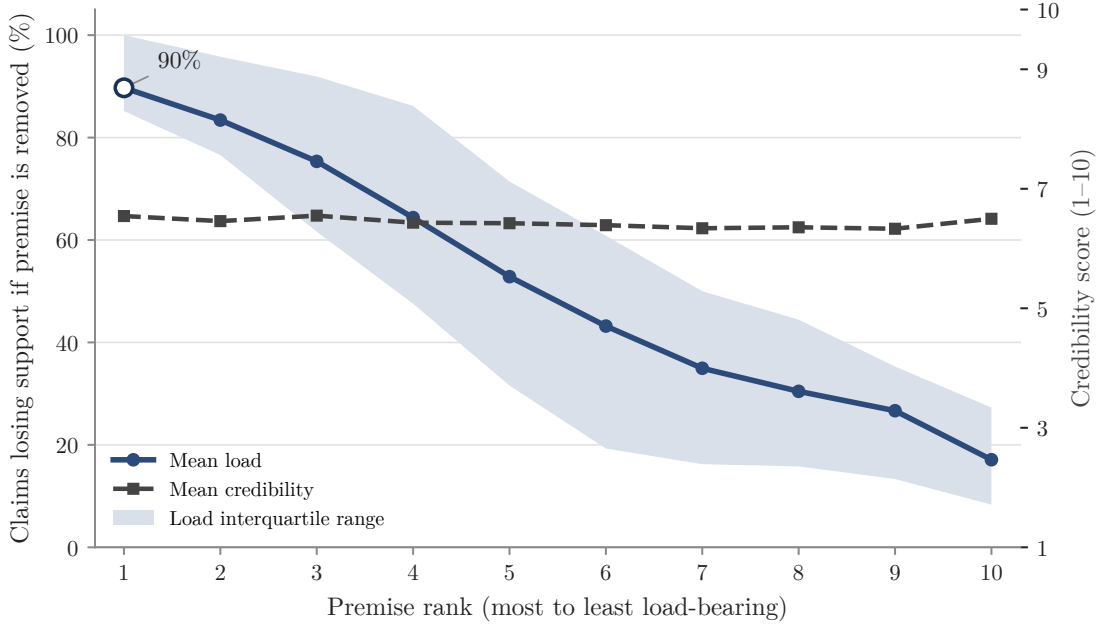
Note: Four examples from the random-sample blinded assessments, organized by network load and credibility score. Higher and lower load refer to the displayed examples, each of which carries positive load. Load is the percentage of weighted headline and major claims that lose support when the premise is removed (weights 1 for headline claims, $\alpha = \frac{1}{2}$ for major claims, and 0 otherwise). Credibility is scored on a 1–10 scale independently of the graph. A score of 4 denotes a restrictive, contestable simplification usable for a narrow exercise rather than an internally inconsistent premise. Source passages in the original articles are Section II.B (pp. 4399–4402) and Theorem 1 (p. 4406) for top-left, Section IV (p. 190) for top-right, Section 3.3 (p. 1079) for bottom-left, and Sections V.B–V.C (pp. 839–841) for bottom-right.

because uninformative talk is also an equilibrium and the selection of revelation is left unjustified. Both premises carry high loads (83 and 88 percent), but differ sharply in defensibility.

Figure 4 plots the aggregate pattern when each paper’s rateable premises are ranked from highest to lowest network load. Across ranks 1 to 10, mean load falls from 90 to 17 percent, whereas mean credibility remains between 6.3 and 6.6. Premises carrying much more of the argument therefore receive similar average credibility scores.

Because the number of premises varies across papers, we also compare highest-load and lower-load premises within the same paper. For each paper with both a maximum-load set and other rateable premises, we compare the mean credibility scores of the two groups, dividing their difference by the pooled within-paper standard deviation. Across the 188 papers included in the comparison, maximum-load premises score 0.18 pooled within-paper standard deviations higher

Figure 4: Load and credibility of central premises



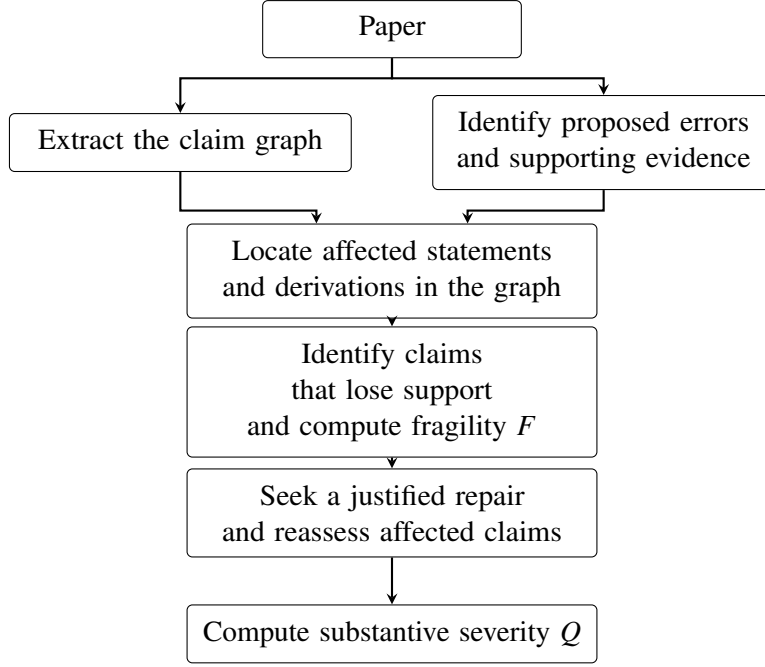
Note: Random sample of 190 papers. The figure ranks each paper’s rateable premises by network load and reports mean load (solid navy line, left axis) and mean credibility (dashed charcoal line, right axis, scored 1–10) at each of the first ten ranks. The shaded band shows the cross-paper interquartile range of load. Each paper contributes one observation at each available rank, and the number of contributing papers falls from 190 at rank 1 to 45 at rank 10. Within a load tie, credibility is averaged across the tied premises and assigned to each occupied rank, so node ordering cannot affect the profile. In the 188 papers with a within-paper comparison, maximum-load premises score 0.18 pooled within-paper standard deviations higher than other rateable premises (95 percent confidence interval 0.05 to 0.31, obtained by resampling papers). Load is the percentage of weighted headline and major claims that lose support when the premise is removed (weights 1 for headline claims, $\alpha = \frac{1}{2}$ for major claims, and 0 otherwise). The adjudicator receives the paper and randomized premise statements with source anchors, without graph edges, contribution roles, downstream claims, loads, or ranks. Data source: claim graphs and blinded premise-credibility assessments.

on average than the paper’s other rateable premises (95 percent confidence interval 0.05 to 0.31). This pattern mitigates general concerns that authors tend to hang main conclusions on less credible assumptions.

4 Errors and repairs

An error in a shared step can affect several main claims at once. We use the claim graphs to assess these consequences and how much of the affected contribution survives repair. We describe the audit and evaluation procedure, assess detection against published corrections, and examine repair outcomes and consequential findings in the random sample.

Figure 5: From proposed errors to substantive severity



Note: Schematic of the procedure for a paper and its proposed errors. Boxes show graph extraction, independent error discovery, linkage to affected statements and derivations, loss-of-support scoring, repair, and reassessment. Fragility F measures weighted support lost before repair; severity Q measures qualitative content lost after the weakest valid repair. Both are percentages of advertised weight W . Weights are 1 for headline claims, $\alpha = \frac{1}{2}$ for major claims, and 0 otherwise; W is their sum. Arrows indicate the flow of inputs and outputs. Graph extraction and error discovery run independently. The procedure uses the original article and claim graph.

4.1 Audit procedure

Figure 5 summarizes error discovery, linkage to the claim graph, and reassessment after repair. Discovery identifies potential mistakes in the original article independently of graph extraction. Linkage traces which conclusions lose support, while repair assesses which of those conclusions can be recovered.

Error discovery. An LLM agent audits each paper’s formal arguments using the original PDF and its text transcription. Each finding identifies the relevant passage, provides supporting evidence, such as a counterexample or rederivation, and considers alternative interpretations and possible corrections. Severity is assessed separately. Discovery is independent of graph extraction. For papers with published corrections, the correction is withheld, so the agent works from the original article alone.

Linkage and repair. A language model links each finding to affected statements or derivations in the claim graph, following Section 2.4. The recorded interventions determine network fragility

and the potentially affected claims, R_f . A separate repair assessment seeks the weakest valid correction and records which of these claims survive, require qualification, or are lost, following Section 2.5. Both scores are computed deterministically from these recorded interventions and claim-level assessments. All outputs are fixed before comparison with published corrections.

4.2 Validation criteria and metrics

Published corrections benchmark detection of the errors they document. After the audit outputs are fixed, a model compares findings with these errors and maps the documented errors to graph objects (Appendix D). This comparison evaluates detection, without independently validating repairs or severity scores. Because corrections are not a complete inventory of errors, an unmatched finding may be false or may identify an undocumented error. The documented-error match rate therefore cannot directly establish precision.

Matching and evaluation. We evaluate detection separately for each documented error, since a single published correction may describe more than one. After the graph extraction, audit findings, and scores are fixed, we compare the findings with each error described in the correction. An exact match identifies both the affected result or proof step and the central mistake described in the correction. A partial match identifies the affected result or step but misses or misstates the central mistake. A documented error is classified as missed if no finding provides an exact or partial match. One finding can match multiple documented errors only if it identifies each error separately. A matched finding identifies a documented error under this exact-or-partial matching rule. We also assess whether findings and their matched documented errors receive the same or compatible error-type labels.

Let $\mathcal{E}$ denote the documented errors, $\mathcal{F}_k$ the findings from run k , and $\text{hit}_k(\epsilon)$ indicate an exact or partial match. Recall is the share of documented errors detected under this matching rule. The documented-error match rate, denoted Match_k , is the share of findings matching at least one documented error:

$$(7) \quad \text{Rec}_k = \frac{\sum_{\epsilon \in \mathcal{E}} \text{hit}_k(\epsilon)}{|\mathcal{E}|}, \quad \text{Match}_k = \frac{|\{f \in \mathcal{F}_k : f \text{ matches at least one documented error}\}|}{|\mathcal{F}_k|}.$$

To distinguish detection of minor errors from detection of errors affecting a larger share of the argument, we also compute fragility-weighted recall:

$$(8) \quad \text{Rec}_k^F = \frac{\sum_{\epsilon \in \mathcal{E}} F_\epsilon \text{hit}_k(\epsilon)}{\sum_{\epsilon \in \mathcal{E}} F_\epsilon},$$

where F_ϵ is the fragility of documented error ϵ . This measure gives greater weight to errors whose

removal leaves more of the paper’s claims unsupported.

We report recall and documented-error match rate for all errors and findings and for the subset relevant to headline claims. An error or finding is claim-relevant if its affected statements or derivations can alter support for at least one headline claim. Errors that cannot be mapped to the graph remain in ordinary recall but receive zero weight in fragility-weighted recall, as do mapped errors with zero fragility. Table 2 also reports a fragility-weighted match rate, weighting each finding by its own deletion-based validation fragility. Each run is evaluated separately.

Fragility of documented errors. We compute fragility for every documented error that can be linked to the graph, regardless of whether the model detects it. For each error, we identify the smallest affected set of nodes and hyperedges O and apply (3) with $D = O \cap V$ and $X = O \cap E$. This assigns fragility independently of detection and allows us to ask whether the model is more likely to find errors with larger consequences for the argument. We refer to this as validation fragility.⁸

4.3 Detection of documented errors

Published corrections provide an external reference for assessing error discovery. Using the 100 corrected papers, we ask how often the audit recovers documented errors and whether graph fragility helps identify findings more likely to match them. We apply the matching criteria and metrics in Section 4.2.

Of the 168 documented errors, 158 can be mapped to graph objects. Their median fragility is 36, and seven have zero fragility. The remaining ten cannot be mapped; two of these are detected by the model. They enter the unweighted detection rate but not the fragility-weighted measure.

Detection and candidate review ordering. Table 2, panel (a) reports recovery of documented errors and the share of findings matching them. The audit returns 477 findings, of which 473 are compared with corrections and 404 receive both fragility and severity scores. Detection counts documented error instances, while match rates count compared findings.

The audit detects 115 of 168 documented errors (68.5 percent), including 79 exact matches (47.0 percent). It therefore locates roughly two-thirds of known errors but identifies the central mistake in fewer than half. Recall is similar for errors affecting headline claims (71.3 percent). It also remains at 68.5 percent when each documented error is weighted by the share of weighted claims that lose support because of it. Appendix Table 7 reports detection by error type.

⁸Validation fragility uses deletion-based mappings. Full-pipeline finding fragility instead uses removed and qualified objects from linkage and can be interpolated or interval-valued. The two procedures can therefore assign different interventions to the same error.

Table 2: Detection and documented-error matching

(a) Detection and matching: deletion-based validation fragility

	All	Claim-relevant	Fragility-weighted
Documented error instances	168	136	
Recall, exact or partial match	115/168 (68.5%)	97/136 (71.3%)	68.5%
Recall, exact matches only	79/168 (47.0%)	68/136 (50.0%)	
Findings	473	306	
Match rate, exact or partial	101/473 (21.4%)	87/306 (28.4%)	29.5%
Match rate, exact only	75/473 (15.9%)		
Type exact or compatible, given detection	109/115		

(b) Detection and matching by validation fragility band

Fragility band	Findings	Matched	Match rate (%)	Recall (detected/errors)
Unmapped	74	2	2.7%	2/10
$F = 0$ (mapped, no weighted claim lost)	30	4	13.3%	6/7
$0 < F \leq 20$	137	27	19.7%	33/43
$20 < F \leq 50$	123	34	27.6%	37/56
$F > 50$	109	34	31.2%	37/52
All findings	473	101	21.4%	115/168

Note: The audit uses the original articles alone, withholding corrections until evaluation. Both panels use 100 corrected-sample publisher originals, 168 documented errors, and 473 compared findings. In panel (a), recall is the share of documented errors detected, and documented-error match rate is the share of findings matching a documented error. Exact and partial matches count as detections in the first recall row, while the second counts exact matches only. Claim-relevant means that a headline claim depends on the error or finding in the graph. Fragility-weighted metrics weight each error or finding by the network fragility (3) of deleting its graph objects (Section 4.2). Weights are 1 for headline claims, $\alpha = \frac{1}{2}$ for major claims, and 0 otherwise; W is their sum. Ten errors and 74 findings map to no graph object, while seven errors and 30 findings map to objects supporting no weighted claim. All carry zero fragility weight. Two unmapped errors are detected. Two unmapped and four mapped zero-weight findings match a documented error. Type exact or compatible counts detected errors with a matching or compatible type label. The exact-only row counts findings with at least one exact match; nine also partially match a different documented error. Panel (b) uses the same deletion-based validation fragility from the structural mapping pass, not full-pipeline interval scores. Recall bands use each documented error’s own fragility. Unmapped findings and errors enter the Unmapped row and the total. Bands contain 158 errors and 113 detections, giving 168 errors and 115 detections overall. Cumulative documented-error match rate is 25.7 percent for $F > 0$, 30.4 for $F \geq 25$, 31.9 for $F \geq 50$, and 30.6 for $F \geq 75$. Data source: published corrections, claim graphs, and blind-audit comparison records.

Detecting known errors is only part of the task. An audit that returns many unconfirmed findings can still leave substantial work for a human reviewer. Here, 101 of the 473 findings match at least one documented error (21.4 percent); together, they identify the 115 detected errors because a single finding can identify several errors. Of the matched findings, 99 receive both fragility and severity scores; we use this same set throughout the following comparisons of their consequences.

Table 2, panel (b), asks whether prioritizing findings by their potential consequences also selects findings more likely to match documented errors. The match rate rises from 13.3 percent among mapped findings with zero fragility to 31.2 percent among findings threatening more than half of weighted claim support. Graph fragility therefore provides a natural way to prioritize expert review, directing attention toward findings with greater potential consequences and a higher documented-error match rate.

4.4 Illustrative repairs

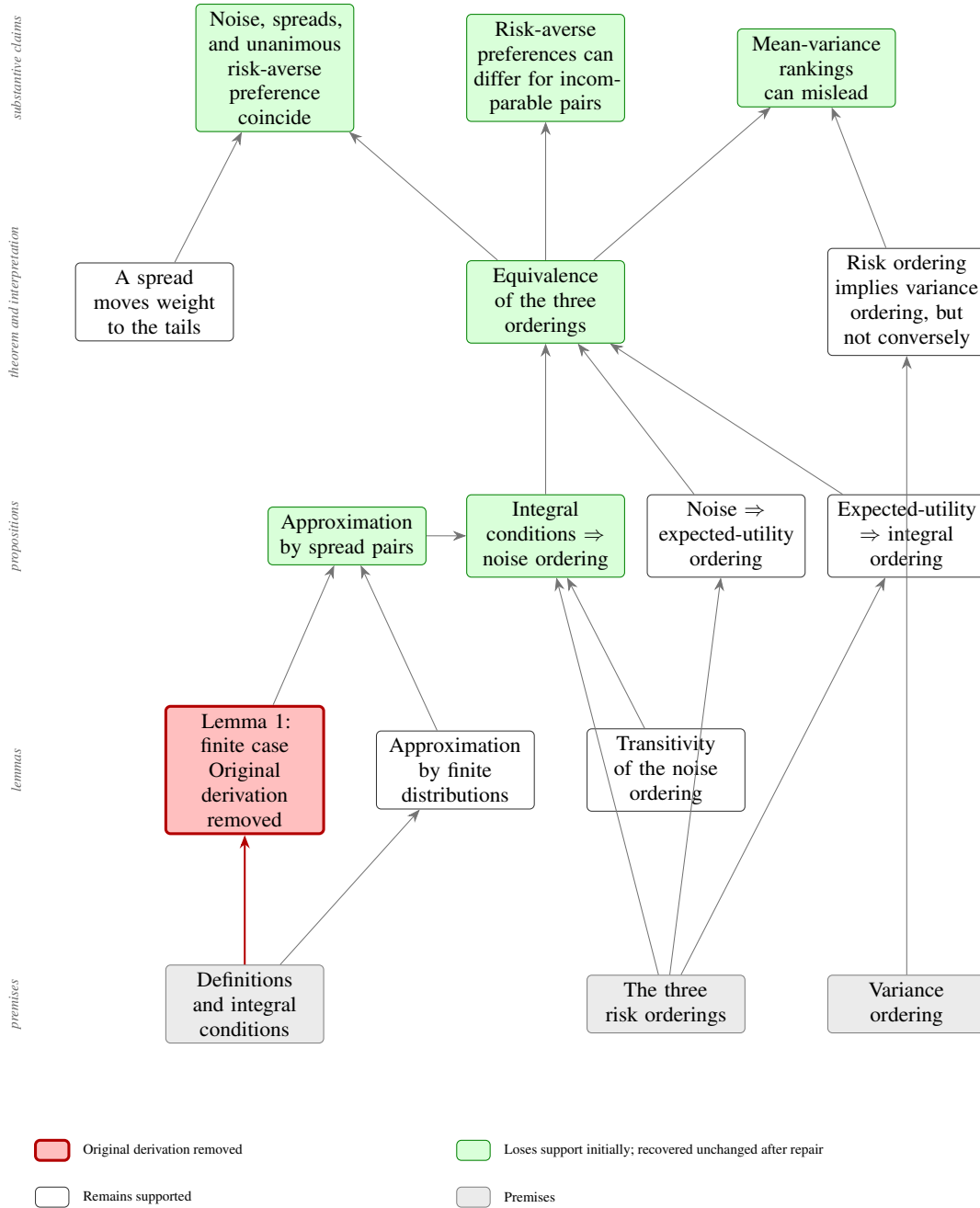
Two cases show when correcting a faulty argument requires an economist to change its conclusions. In [Rothschild and Stiglitz \(1970\)](#), a proof correction preserves the economic conclusions. In [Brown \(1983\)](#), central identification conclusions do not survive the model-assessed repair intact. These contrasting cases illustrate the distinction between network fragility, the weighted claim support initially lost, and substantive severity, the qualitative content assessed as lost after repair.

Figure 6 follows the finding in [Rothschild and Stiglitz \(1970\)](#) through repair. The construction used to prove Lemma 1 can produce an intermediate function that is not a valid distribution function. Removing this derivation leaves six weighted claims unsupported, including the equivalence theorem and claims derived from it. Their combined weight is 4.5 out of a total of 10, giving $F = 45$.

A proof rewrite that preserves valid distributions at each step recovers all six claims without changing their conclusions, giving $Q = 0$. The published correction reaches the same conclusion through a different construction ([Leshno et al., 1997](#)). The example shows why exposure before repair can be substantial even when the economic conclusions survive unchanged.

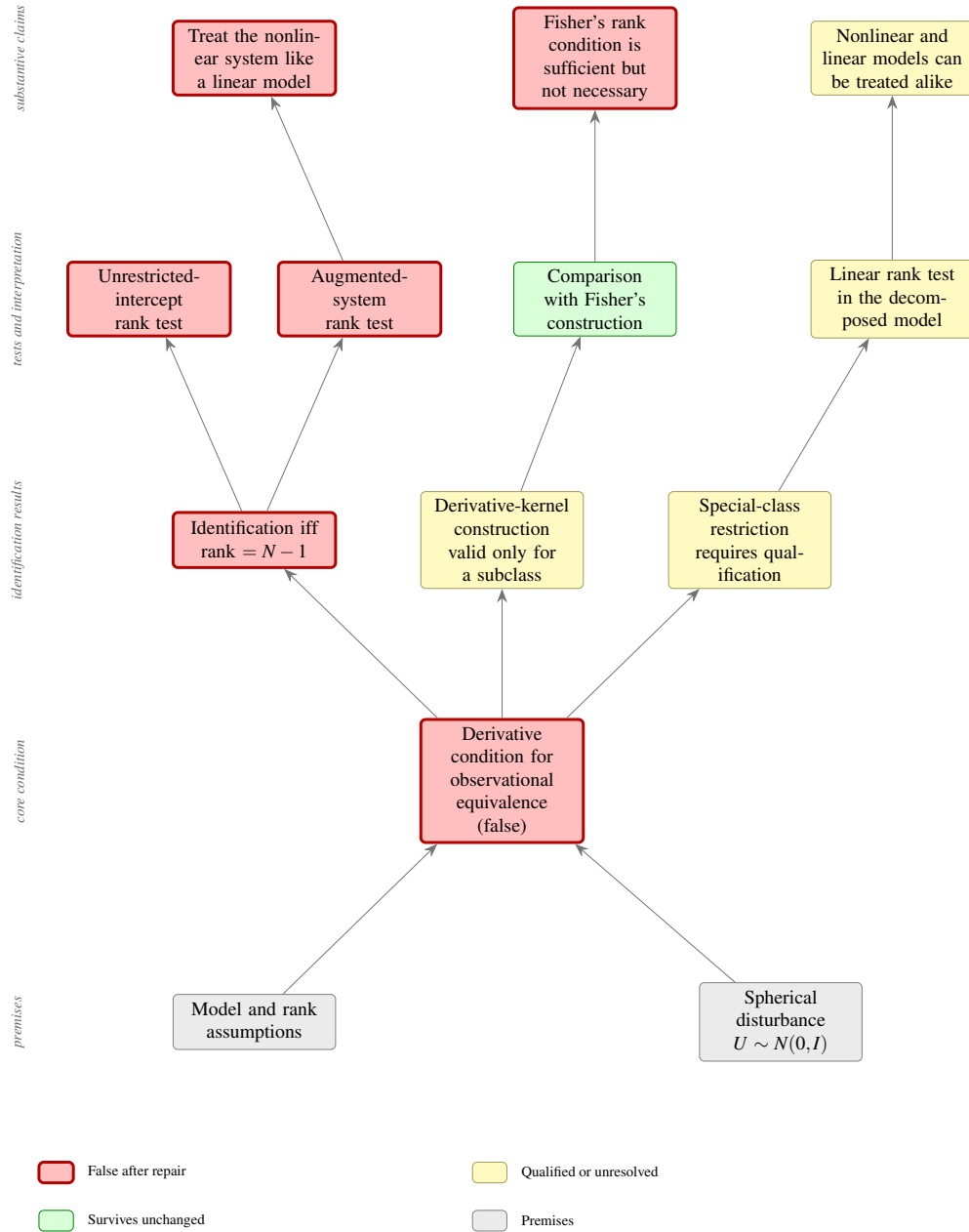
[Brown \(1983\)](#) gives the contrasting case. The paper claims that a transformed vector disturbance $\tilde{U} = T(X, U)$ is independent of X if and only if $\partial T / \partial X = 0$. With the published Comment withheld, the audit rejects the converse by constructing an X -dependent orthonormal rotation, $\tilde{U} = \Gamma(X)'U$. If U is independent of X and $U \sim N(0, I)$, then $\tilde{U} \mid X = x \sim N(0, I)$ for every x , even though its derivative with respect to X is generally nonzero. An X -dependent transformation can therefore leave the disturbance distribution unchanged. Economically, the problem is that different structural models can have the same observable implications, so the data need not single out the structure claimed to be identified. The audit links this failure to Brown's derivative-based rank conditions and identification conclusions. After the audit and graph were completed, a model-based comparison

Figure 6: Fragility and repair in the Rothschild–Stiglitz graph



Note: Removing the derivation of Lemma 1 leaves six weighted claims unsupported: weight 4.5 out of 10, giving $F = 45$. A proof rewrite recovers all six unchanged, giving $Q = 0$. Both measures are percentages; headline claims have weight 1, major results $\alpha = \frac{1}{2}$, and other statements zero. Red marks the removed derivation, green the recovered claims, white statements that remain supported, and grey premises. Incoming arrows jointly represent a supporting derivation. Unaffected results are omitted. Source: the Rothschild–Stiglitz (1970) graph and its audit, linkage, and repair outputs.

Figure 7: Fragility and repair in the Brown graph



Note: The rotation counterexample rejects the derivative condition at the base of the argument. Removing it leaves every weighted output unsupported ($F = 100$). After repair, red claims are false, yellow claims are qualified or unresolved, and green claims survive unchanged. Model-assessed severity is $Q \in [57.1, 92.9]$; unresolved scope determines the interval. Its midpoint of 75 is the highest among the corrected sample's 404 scored findings and represents loss of three-quarters of weighted qualitative content. Grey nodes are premises and arrows show dependencies. The published Comment was withheld from the audit's input. The subsequent model-based comparison classified the finding as an exact match to the central error. Source: graph and audit outputs, and [Benkard and Berry \(2006\)](#).

with [Benkard and Berry \(2006\)](#) classified the finding as an exact match. Their Comment gives the same rotation mechanism, calls the finding substantive, and states that the counterexample contradicts the main theorems.

Figure 7 shows how this failure affects the identification argument. Deleting the derivative condition removes support from every weighted output, so $F = 100$. The model-assessed repair replaces pointwise invariance with invariance of the transformed disturbance distribution. Under this assessment, the original finite-dimensional rank conditions are no longer sufficient: six weighted outputs are assessed as false after repair, while four are qualified or unresolved. The resulting model-assessed severity is $Q \in [57.1, 92.9]$, with midpoint 75. This is the highest midpoint severity among all 404 scored findings in the corrected sample, placing it in the top 1 percent; it also exceeds every scored random-sample finding. The assessment therefore implies a substantial failure of the identification argument: even the lower endpoint represents loss of more than half of the paper’s weighted qualitative contribution, and the midpoint represents loss of three-quarters.

4.5 Consequences after repair

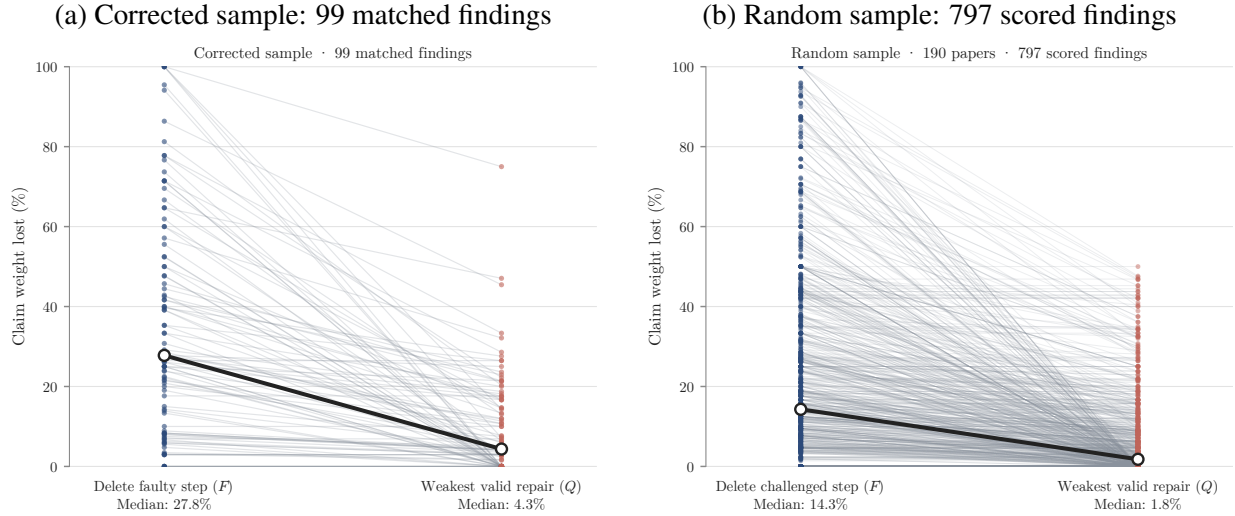
The examples show how repair can separate the loss of a derivation from the loss of an economic conclusion. We now examine how often repair recovers the contribution initially exposed by a finding. We begin with findings matched to documented errors, then extend the comparison to the random sample.

Consequences of assessed repairs. Figure 8 follows each finding from fragility before repair to severity after repair. Panel (a) leads with the 99 matched findings. Their median fragility is 27.8 and median severity is 4.3. Thirty-two percent have zero severity, while 10 percent imply losses of at least one-quarter of weighted qualitative content. Severity is below fragility in 87 of 99 cases, with 12 unchanged scores. Repair therefore recovers much of the initially affected contribution for the typical matched finding, while a smaller group retains substantial qualitative loss.

Panel (b) extends the comparison to 797 scored findings from 190 random-sample papers. Median fragility is 14.3 and median severity is 1.8. Severity is below fragility in 522 of 797 findings, roughly two-thirds. The paired declines in both groups show that much of the contribution exposed by a faulty step can be recovered through repair. Each comparison follows an individual finding from its initial consequences to the qualitative loss remaining after correction. Appendix Table 8 reports further cases, including unresolved repairs for which severity remains an interval.

Comparison by documented-error match. Table 3, panel (b), compares 99 scored findings matched to documented errors with 305 unmatched scored findings from the corrected sample.

Figure 8: Claim loss before and after repair



Note: Thin lines join fragility F before repair to severity Q after repair for individual findings. The thick black line with open circles joins the group medians at the two stages. Panel (a) uses 99 unique scored findings matched to documented errors, with 87 decreases and 12 unchanged scores. Panel (b) uses all 797 scored findings from the 190 random-sample papers, with 522 decreases, 274 unchanged scores, and one increase. Both scores are percentages of weighted claims or qualitative content, respectively; interval scores enter at their midpoints. An increase relative to midpoint F is compatible with the bound $Q_f \leq F_f^+$. Each finding appears once, even when it matches multiple documented errors. Headline claims have weight 1, major claims $\alpha = \frac{1}{2}$, and other statements zero. Source: published corrections and the graph, audit, linkage, and repair assessments.

Appendix Figure 10 shows their distributions. Matched findings have significantly higher mean fragility (35.5 versus 20.9, $p < 0.001$) and severity after repair (9.6 versus 5.8, $p = 0.011$). They are also nearly twice as likely to imply a loss of at least one-tenth of weighted qualitative content (38.4 versus 20.3 percent). Because scores are fixed before comparison with corrections, the mean differences associate greater assessed consequences with findings subsequently linked to documented errors. Matching evaluates detection of those errors, rather than the accuracy of the repair assessments.

The strict thresholds distinguish exposure before repair from substantial loss after it. More than half of matched findings have $F > 25$, compared with 31.1 percent of unmatched findings. After repair, the shares with $Q > 25$ are 9.1 and 6.6 percent, respectively ($p = 0.433$). Only one of the 404 scored findings has $Q > 50$, and none has $Q > 75$. Thus the higher average assessed consequences of matched findings do not establish a clear difference in the most severe tail.

4.6 Findings in published theory

We use the random sample to assess how frequently published theory papers have findings with substantial consequences after repair. The paper-level outcome is whether at least one finding crosses a specified severity threshold. We compare incidence across severity thresholds and interval

endpoints to show how the assessment depends on the size and uncertainty of the losses.

Frequency of findings. The severity thresholds of 10 and 25 percent select roughly the upper quarter and upper 8 percent of scored findings, respectively, in both samples when intervals enter at their midpoints. Table 3, panel (a), reports incidence among the 190 random-sample papers. At $Q \geq 10$, 8 percent have at least one finding that crosses the threshold using lower endpoints, 58 percent using midpoints, and 75 percent using upper endpoints. Severity remains an interval when a repair outcome is unresolved or the fraction of scope lost cannot be justified. The spread across these conventions shows how uncertainty about recoverable claims affects measured incidence. The outcome tracks the presence of at least one consequential finding in a paper, so a qualifying paper can still retain other economic conclusions.

Under the midpoint convention, 28 percent of random-sample papers have a finding with $Q \geq 25$. Among the 100 corrected-sample papers, the corresponding shares are 58 percent at $Q \geq 10$ and 24 percent at $Q \geq 25$. The audit returns a finding for every random-sample paper and one with positive fragility for 97 percent. Appendix Table 9 provides the full breakdown for both samples, with Wilson confidence intervals describing sampling uncertainty conditional on each endpoint convention.

Comparison across samples. Appendix Tables 10 and 11 compare assessed consequences across samples, and Appendix Figure 11 shows their distributions. Mean severity is 6.8 in corrected papers and 6.6 in random papers ($p = 0.84$), and mean paper-level maximum severity is 16.2 versus 16.5 ($p = 0.85$). Within corrected papers, findings matched to documented errors have higher mean severity than random-sample findings (9.6 versus 6.6, $p = 0.029$), whereas unmatched findings do not differ significantly ($p = 0.33$). The small aggregate mean difference therefore coexists with greater assessed consequences among matched findings.

4.7 Structural correlates of consequential findings

We finally examine which features of an argument are associated with more consequential findings. This links the earlier evidence on structural dependence to the qualitative losses remaining after repair. Table 4 pools the 190 random-sample and 100 corrected-sample papers. The first column examines each paper's maximum severity after repair. The second examines whether at least one finding reaches severity of 10 percent. Both outcomes use midpoint severity when scores are intervals. Both specifications include maximum load, result yield, derivation depth, and an indicator for the corrected sample. The structural measures are standardized, so their coefficients describe differences associated with a one-standard-deviation increase, holding the other included variables constant.

Table 3: Audit coverage and assessed consequences

Panel (a): Paper-level coverage in the random sample	
	Random sample
Papers	190
Findings per paper (mean)	5.3
Findings scored per paper (mean)	4.2
Papers with at least one finding	100%
Papers with a finding with $F > 0$	97%
Papers with $Q \geq 10$, lower endpoints	8%
Papers with $Q \geq 10$, midpoints	58%
Papers with $Q \geq 10$, upper endpoints	75%
Papers with $Q \geq 25$, midpoints	28%

Note: Panel (a) covers 190 random-sample papers. The first row counts papers, the next two report mean findings per paper, and all remaining rows report percentages of papers. A paper crosses a severity threshold if at least one finding crosses it under the stated interval convention. Appendix Table 9 reports both samples and Wilson 95 percent confidence intervals.

Panel (b): Finding-level scores from 100 corrected papers					
	Total	Matched	Unmatched	Difference	p
Scored findings	404	99	305	—	—
Mean F	24.5	35.5	20.9	+14.6	< 0.001
Mean Q	6.8	9.6	5.8	+3.8	0.011
Share with $F > 0$ (percent)	84.2	94.9	80.7	+14.3	< 0.001
Share with $F > 25$ (percent)	37.1	55.6	31.1	+24.4	< 0.001
Share with $F > 50$ (percent)	15.6	26.3	12.1	+14.1	0.003
Share with $F > 75$ (percent)	6.2	11.1	4.6	+6.5	0.056
Share with $Q > 0$ (percent)	55.4	67.7	51.5	+16.2	0.005
Share with $Q > 25$ (percent)	7.2	9.1	6.6	+2.5	0.433
Share with $Q > 50$ (percent)	0.2	1.0	0.0	+1.0	0.319
Share with $Q > 75$ (percent)	0.0	0.0	0.0	+0.0	—
Share with $Q \geq 10$ (percent)	24.8	38.4	20.3	+18.1	0.002

Note: Panel (b) covers 404 scored findings from 99 of the 100 corrected-sample papers, a different population and unit from panel (a). One corrected-sample paper has no scored finding. Matched and Unmatched partition these findings by documented-error match, not the 190 random-sample papers. Total uses all corrected-sample scored findings. Fragility F measures weighted support lost before repair and severity Q measures qualitative content lost after the weakest valid repair, both as percentages of advertised weight W . Headline claims have weight 1, major claims $\alpha = \frac{1}{2}$, and other statements zero. Interval scores enter at their midpoints. Shares are percentages and their differences are percentage points. Difference is Matched minus Unmatched. Threshold comparisons use unrounded scores and strict $>$ cutoffs, except for the retained $Q \geq 10$ row. The p -values use OLS difference-in-means regressions with standard errors clustered by paper. A dash denotes undefined inference for the all-zero $Q > 75$ outcome. Source for both panels: claim graphs and their audit, linkage, and repair outputs.

Table 4: Structural correlates of consequential findings

	Maximum Q	$1[\max Q \geq 10]$ (percent)
Depth $D_{\max}$ (SD)	0.16 (0.79)	-0.1 (3.1)
Maximum load $L_{\max}$ (SD)	3.61*** (0.80)	9.9*** (3.0)
Result yield Y (SD)	-3.24*** (0.61)	-8.3*** (2.0)
Corrected sample	4.28** (1.98)	12.7* (6.6)
Papers	290	290
R^2	0.10	0.05

Note: Paper-level ordinary least squares (OLS) regressions pooling 190 random-sample and 100 corrected-sample papers. Both outcomes use midpoint Q when severity is an interval. Column 1 uses maximum substantive severity Q per paper, with 0 for no scored finding. This zero records the absence of a scored finding, rather than an assessed loss of zero for an unscored finding. Column 2 is a linear probability model for at least one finding with $Q \geq 10$, reported in percentage points. $1[\max Q \geq 10]$ is the corresponding indicator. Result yield Y is formal results per substantive primitive input: nodes typed as primitives, assumptions, parameter restrictions, or scope conditions. Derivation depth $D_{\max}$ is the longest chain of derivations from a premise to a headline claim. Maximum load $L_{\max}$ is the largest percentage of weighted claims that lose support after removing one active primitive, which has no incoming derivation and enters a downstream derivation. Q is qualitative content lost after the weakest valid repair, as a percentage of advertised weight. Weights are 1 for headline claims, $\alpha = \frac{1}{2}$ for major claims, and 0 otherwise; W is their sum. Continuous regressors are standardized to unit variance over the pooled sample, with SD denoting standard deviation. Corrected sample is an unstandardized indicator for that sample. R^2 is the share of outcome variation explained. Parentheses contain HC1 heteroskedasticity-robust standard errors. Stars denote * $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$.

Maximum load measures how much weighted claim support depends on the paper’s most consequential starting input. A one-standard-deviation increase is associated with 3.6 points higher maximum severity and a 9.9-percentage-point higher probability of a finding reaching the severity threshold. The association with maximum load also persists when controlling for textual topic (Appendix K). Thus, greater dependence on a single starting input is associated with greater assessed loss even after repair. One possible explanation is that repairing a challenged input becomes more consequential when many conclusions depend on it.

Result yield shows the opposite association. Papers deriving more formal results per substantive input have lower maximum severity and are less likely to cross the threshold. A one-standard-deviation increase in yield is associated with reductions of 3.2 severity points and 8.3 percentage points, respectively. Derivation depth shows no clear association with either outcome. These results suggest that dependence on a single starting input and the number of results derived per substantive input are more closely associated with assessed consequences than the length of its longest derivation. The specifications explain 10 percent of variation in maximum severity and 5 percent in threshold incidence. Across both outcomes, the main pattern is that concentrated dependence is associated with greater assessed loss, while higher result yield is associated with lower assessed loss.

5 Conclusion

This paper develops claim graphs to trace how economic conclusions depend on premises and derivations and assess what survives when an argument requires correction. Conclusions often share supporting premises, but the most load-bearing premises are not systematically less credible. Assessed repairs recover much of the economic content initially exposed by a finding, although some findings retain substantial losses. Concentrated dependence therefore identifies where scrutiny matters, while the importance of an error depends on what can be recovered after repair.

Automated assessments also require scrutiny of their own. Published corrections provide a benchmark for recovering documented errors, but they do not establish the validity of unmatched findings. Likewise, the frequency of findings in the random sample should not be interpreted as confirmed error prevalence. The framework makes these assessments inspectable by linking each finding to the challenged argument, its downstream consequences, and a proposed repair.

These results suggest that evaluating theory requires distinguishing the validity of a particular argument from the durability of its economic conclusions. A failed proof creates a need for repair, but the substantive question is what can still be established under defensible premises. Conversely, a repair that restores formal validity may narrow a result's scope or change its economic interpretation. An informative assessment should make both the logical defect and these substantive consequences explicit.

Claim graphs could support this approach by helping authors and reviewers identify consequential dependencies, record alternative derivations, and compare proposed repairs. Their role is to organize the evidence on which judgments about a theory's contribution depend. Making that evidence explicit could help disagreements turn on identifiable assumptions and arguments, while clarifying which conclusions survive correction.

References

- Joshua Angrist, Pierre Azoulay, Glenn Ellison, Ryan Hill, and Susan Feng Lu. Economic research evolves: Fields and styles. *American Economic Review*, 107(5):293–297, 2017. doi: 10.1257/aer.p20171117.
- C. Lanier Benkard and Steven Berry. On the nonparametric identification of nonlinear simultaneous equations models: Comment on Brown (1983) and Roehrig (1988). *Econometrica*, 74(5): 1429–1440, 2006. doi: 10.1111/j.1468-0262.2006.00711.x.
- Federico Bianchi, Yongchan Kwon, Zachary Izzo, Linjun Zhang, and James Zou. To err is human: Systematic quantification of errors in published AI papers via LLM analysis, 2025. URL <https://arxiv.org/abs/2512.05925>. Preprint.
- Bryan W. Brown. The identification problem in systems nonlinear in the variables. *Econometrica*, 51(1):175–196, 1983. doi: 10.2307/1912253.
- Claude. More than two thirds of the zeros of the Riemann zeta function are simple and on the critical line. Anthropic research manuscript, 2026. URL <https://www-cdn.anthropic.com/95c246936988e43127bc6b2ceb7077c1dad2d68e.pdf>. Released August 10, 2026; revised August 13, 2026.
- Janet Currie, Henrik Kleven, and Esmée Zwiers. Technology and big data are changing economics: Mining text to track methods. *AEA Papers and Proceedings*, 110:42–48, 2020. doi: 10.1257/pandp.20201058.
- Drew Fudenberg, Jon Kleinberg, Annie Liang, and Sendhil Mullainathan. Measuring the completeness of economic models. *Journal of Political Economy*, 130(4):956–990, 2022. doi: 10.1086/718371.
- Drew Fudenberg, Wayne Gao, and Annie Liang. How flexible is that functional form? quantifying the restrictiveness of theories. *Review of Economics and Statistics*, 108(1):194–209, 2026. doi: 10.1162/rest_a_01401.
- Prashant Garg and Thiemo Fetzer. Causal claims in economics, 2026. URL <https://arxiv.org/abs/2501.06873v2>. Working paper, version 2.
- Itzhak Gilboa, Andrew Postlewaite, Larry Samuelson, and David Schmeidler. Economic models as analogies. *The Economic Journal*, 124(578):F513–F533, 2014. doi: 10.1111/ecoj.12128.

- Itzhak Gilboa, Andrew Postlewaite, Larry Samuelson, and David Schmeidler. Economics: Between prediction and criticism. *International Economic Review*, 59(2):367–390, 2018. doi: 10.1111/iere.12273.
- Daniel S. Hamermesh. Six decades of top economics publishing: Who and how? *Journal of Economic Literature*, 51(1):162–172, 2013. doi: 10.1257/jel.51.1.162.
- Benjamin Kohler, David Zollikofer, Johanna Einsiedler, Alexander Hoyle, and Elliott Ash. Read the paper, write the code: Agentic reproduction of social-science results, 2026. URL <https://arxiv.org/abs/2604.21965>. Preprint, version 2; forthcoming at EMNLP.
- Moran Koren. Theorist toolbox: Tools for agent based LLM-assisted economic theory research, 2026. URL <https://arxiv.org/abs/2606.22337v2>. Working paper, version 2.
- Moshe Leshno, Haim Levy, and Yishay Spector. A comment on Rothschild and Stiglitz’s “Increasing Risk: I. A Definition”. *Journal of Economic Theory*, 77(1):223–228, 1997. doi: 10.1006/jeth.1997.2302.
- Thiago Dumont Oliveira and Marwil J. Dávila-Fernández. From modelmania to datanomics? the rise of mathematical and quantitative methods in three top economics journals. *Scientometrics*, 123(1):51–70, 2020. doi: 10.1007/s11192-020-03375-y.
- OpenAI. Finite time blowup for Navier–Stokes, 2026. URL <https://cdn.openai.com/pdf/32d9f210-8b73-45e0-91bc-82a30aef8a9a/navier-stokes.pdf>. Research manuscript released September 8, 2026.
- Pat Pataranutaporn, Nattavudh Powdthavee, Chayapatr Achiwaranguprok, and Pattie Maes. Can AI solve the peer review crisis? a large scale cross model experiment of LLMs’ performance and biases in evaluating over 1000 economics papers, 2025. URL <https://arxiv.org/abs/2502.00070v2>. Working paper, version 2.
- Michael Rothschild and Joseph E. Stiglitz. Increasing risk: I. A definition. *Journal of Economic Theory*, 2(3):225–243, 1970. doi: 10.1016/0022-0531(70)90038-4.
- Ariel Rubinstein. Dilemmas of an economic theorist. *Econometrica*, 74(4):865–883, 2006. doi: 10.1111/j.1468-0262.2006.00689.x.
- Matthew Schwartz, Isaiah Andrews, and Jesse M. Shapiro. An LLM workflow that reproduces, improves, and extends published economics research. Working Paper 35782, National Bureau of Economic Research, September 2026. URL <https://www.nber.org/papers/w35782>.

Guijin Son, Jiwoo Hong, Honglu Fan, Heejeong Nam, Hyunwoo Ko, Seungwon Lim, Jinyeop Song, Jinha Choi, Gonçalo Paulo, Youngjae Yu, and Stella Biderman. When AI co-scientists fail: SPOT—a benchmark for automated verification of scientific research, 2025. URL <https://arxiv.org/abs/2505.11855>. Preprint.

Alexis Akira Toda. Can AI refute economic theory? evidence from beyond the knowledge cutoff, 2026. URL <https://arxiv.org/abs/2606.05383>. Working paper.

Olaf Willner and David Yanagizawa-Drott. Verifying the verifiers: Towards autonomous policy evaluation. Working paper, University of Zurich, 2026. URL <https://ape.socialcatalystlab.org/verify>. Version dated September 25, 2026.

Table 5: Claim-level loss after repair

Outcome for claim v	$q_f(v)$
Unchanged	0
Quantitatively changed only	0
Valid on a narrower domain	ϕ_v if justified; otherwise $[0, 1]$
Reversed or false	ϕ_v if the loss is confined to a justified fraction of the domain; otherwise 1
Unresolved	$[0, \phi_v]$ if the potentially affected scope is bounded; otherwise $[0, 1]$

Note: Scoring rule, with one row per repair outcome for a claim v affected by finding f . Claim-level loss $q_f(v)$ is the share of the claim's qualitative content lost after the weakest valid repair. Scope loss ϕ_v is the justified fraction of the original domain lost. Numerical changes alone receive zero loss. Intervals record unresolved content or an unjustified scope fraction, with endpoints set by the outcome shown. The rule is applied to repair assessments attached to claim graphs.

A Claim-level loss scoring

Table 5 maps each repair outcome to the share of a claim's qualitative content lost.

B Schemas

B.1 Graph contract (`graph.json`)

A graph file carries `schema_version`, `framework`, `paper` (title, identifier, source), `graph_conventions`, `target_claims`, `nodes`, `edges`, and `unresolved_mappings`. Every node and edge carries at least one anchor: PDF page, printed page, section, result number, equation number, and a source phrase of at most 200 characters. Schema validation checks the enumerations below, requires every edge source and target to exist, and requires the propagating subgraph to be acyclic.

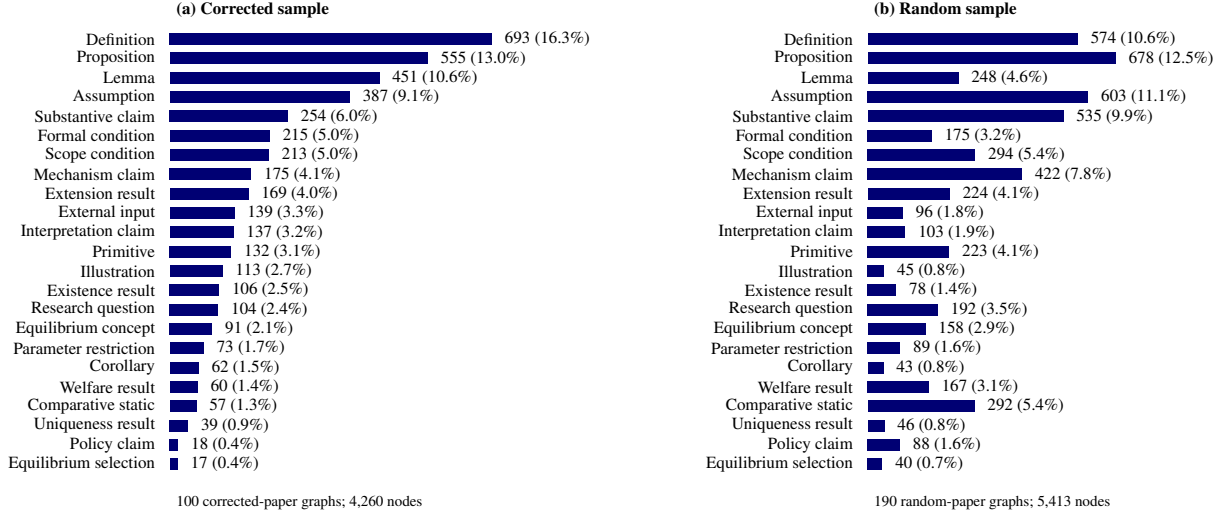
Node fields. `id`, `type`, `layer`, `statement` (4 to 2000 characters), `mathematical_form`, `contribution_role`, `source_status`, `root_status`, `anchors`, `extraction_confidence`, `mapping_note`.

Node types (23). research question, primitive, definition, assumption, parameter restriction, equilibrium concept, formal condition, lemma, proposition, corollary, existence result, uniqueness result, equilibrium selection, comparative static, welfare result, mechanism claim, substantive claim, interpretation claim, policy claim, scope condition, extension result, illustration, external input.

The distribution of semantic node types in the 100 corrected-sample graphs and the 190 random-sample graphs is shown in Figure 9.

Layers (3). Formal, interpretive, contextual. **Contribution roles (5).** Headline, major, supporting, auxiliary, contextual; only headline and major carry scoring weight. **Source status (4).** Explicit, formalized, reconstructed, attributed external. **Root status (4).** Premise, derived, asserted without support, contextual.

Figure 9: Node-type frequencies across samples



Note: Node types in 190 current random-sample graphs (5,438 nodes, left) and 100 corrected-sample graphs (4,260 nodes, right), over the same 23 semantic types. Each node has one type, distinct from contribution roles such as headline or supporting. Both panels use the same category order. Bars show within-sample percentages; labels report counts and percentages. Navy denotes the random sample and red the corrected sample. Data source: claim graphs extracted from the original articles.

Edge fields. id, sources (one or more node ids), target, relation, dependency_role, propagates_support, source_status (explicit or reconstructed), anchors, extraction_confidence, mapping_note.

Edge relations (17). Defines, restricts scope, enters equilibrium, formally derives, establishes existence, establishes uniqueness, selects equilibrium, produces comparative static, produces welfare result, supports mechanism, supports substantive claim, supports interpretation, motivates policy claim, qualifies, extends, illustrates, imports external support.

Dependency roles (6). Complete inferential route, corroborative, interpretive, scope display, motivational, unclear. Only complete inferential routes propagate support.

B.2 Audit contract (audit.result.json)

Top level: paper_id, claim_ledger, findings, scope, blocked_reason. Each ledger entry names a numbered proposition, theorem, lemma, corollary, remark, or load-bearing equation and gives one verdict: survives, survives with qualification, unproved as written, false as stated. Every verdict other than survives cites at least one supporting finding. False as stated requires counterexample or rederivation evidence and cannot rest on a boundary-case finding alone.

Finding fields. `finding_id`, `claim_id`, `code`, `quote_anchor`, `location`, `severity`, `fixability`, `evidence_kind`, `evidence_detail`, `propagation`, `status` (definite or possible), `confidence`, `explanation`. The pipeline consumes only the core fields (identifier, quote anchor, location, evidence kind and detail, status, explanation); `code`, `severity`, and `fixability` are optional on ingestion and never enter a score.

Codes (25). Derivation error, sign or index error, differentiation error, unjustified step, local to global, quantifier error, circular argument, missing assumption, hidden normalization, stronger than stated, domain violation, boundary case, objective inconsistency, dimensional error, existence or uniqueness, selection unstated, comparative-static scope, envelope or first-order-condition misuse, statement-proof mismatch, overclaim, false as stated, appendix-main-text mismatch, reference error, math typo, other. Each code has a minimum admission evidence stated in the taxonomy (Appendix N).

Severity (4). Fatal, major, minor, cosmetic. **Fixability (5).** Trivial fix, local fix, restatement needed, substantial rework, no known fix. **Evidence kind (5).** Counterexample, rederivation, numerical scan, internal contradiction, reasoning only (always status possible).

Neutral-discovery output. Under `neutral-discovery-direct-v1.4` the model writes `formal_map.json` (paper identity with a rendered title-page check; primitives; domains; timing and information; assumptions and quantifiers; definitions; solution concepts; numbered claims; proof dependencies; load-bearing equations; identification, estimation, and asymptotic statements; formal appendices), `candidates.json` and `rejected_candidates.json` (records with candidate id, status in {candidate, unresolved, rejected}, title, sources, location, rendered source check, claim being tested, intended argument, first potentially invalid step, reproducible evidence, alternative interpretation, weakest repair, directly affected claims, open checks, evidence files, check files), executable checks under `checks/`, rendered page images under `evidence/`, and a `neutral_summary.txt`. `import_neutral_review.py` converts a review into the audit contract’s minimal profile; the discoverer’s weakest-repair hints are written to a separate hints file that the repair stage never reads.

B.3 Linkage contract (`linkage.json`)

One mapping per finding with `finding_id`, `match_status` (matched, no structural effect, ambiguous, unmatched), `targets` (each an intervention in {delete node, delete edge, qualify node, qualify edge} on a named object), `candidate_object_ids`, `consequence_kind` (false, unsupported, qualified, reversed, no structural effect, uncertain), `scope_region_lost`, `scope_loss_fraction`, `directly_affected_claim_ids`, `mapping_confidence`, `mapping_rationale`. The validator checks the mapping jointly against the graph and audit and enforces the epistemic guard: consequence false or reversed requires counterexample, rederivation, or internal-contradiction evidence in the finding.

B.4 Repair contract (repair.result.json)

One analysis record per linkage finding, including unresolved or unassessable cases, with `analysis_status` (resolved, partially resolved, unresolved, not needed, not assessable), `weakest_valid_repair` (a precise statement of the corrected object), `repair_burden` (none, local formula, proof rewrite, assumption change, model change, unresolved, not applicable), `changed_graph_object_ids`, `output_assessments`, `overall_reasoning`, `repair_confidence`, and `adjudication_status` (model provisional or human adjudicated). Each output assessment names an output in the deterministic target set R_f and gives `repaired_status` (survives unchanged, survives quantitatively changed, survives with scope qualification, qualitatively reversed, false after repair, unresolved after repair), `affected_scope_fraction`, `scope_description`, and `derivation_summary`. The validator rejects any omission from or addition to R_f .

B.5 Score file (severity.json)

Produced only by `src/claim_graph.py`: role weights, scored output ids, per-finding scores (network fragility bounds and point, substantive severity bounds and point, network failure class, substantive class, single points of failure, joint-cut outputs, repair burden, adjudication status), combined network fragility over the union of matched deletions, no combined substantive score, reference-distribution size, and notes. Percentile ranks are attached only when a human-adjudicated reference distribution is supplied.

C Reproducibility and source separation

Graph extraction and error discovery use only the original article. The graph is fixed before findings are linked to it; linkage and repair then identify the affected statements and assess what conclusions survive. Fragility and severity are computed deterministically from these assessments. Reports and visualizations display the resulting records without altering them.

Source documents, prompts, model identifiers, and intermediate assessments are recorded so that each score can be traced to its inputs. Outputs are checked against the schemas in Appendix B. Corrections to output format do not introduce information from published corrections. The prompts used for extraction, discovery, linkage, repair, and validation are reproduced in Appendix N.

D Validation protocol

Catalogue. The catalogue contains 118 rows, covering papers from 1953 to 2022; 115 rows include a DOI, and each records the paper and correction bibliography, the documented error in prose, the outcome class, and a coder confidence. Publisher versions of both documents were obtained for 107 pairs and stored with a manifest (DOI, source URL, file hash, page count, extraction check). The validation exercise uses the first 100 complete pairs in catalogue order. The other 18 rows comprise seven complete pairs held outside the analysis and 11 rows without a complete publisher-document pair.

Unit and blindness. The unit is the documented error instance. Pipeline stages receive only the original PDF under a generic filename and a neutral path. Bibliographic information printed inside the paper remains visible. Correction PDFs, source manifests, case metadata, answer keys, and the evaluation specification stay outside every model workspace.

Evaluation. Each original article is audited once, independently of graph extraction. After the graph and findings are fixed, a model-based comparison reads the published correction, identifies its documented errors, and matches them to the audit findings using the exact and partial criteria in Section 4.2. It also compares error types, maps documented errors to graph objects, and identifies claim relevance. Published corrections are withheld from all preceding assessments. The comparison and mapping prompts appear in Appendix N.

Metrics and coverage. Recall counts documented errors matched exactly or partially; documented-error match rate counts findings matching at least one documented error. Claim-relevant variants restrict the corresponding units to those affecting headline claims. Fragility-weighted recall uses each documented error’s own mapped consequences, including errors the audit misses. The audit returns 477 findings, of which 473 are evaluated against corrections and 404 receive fragility and severity scores. Detection statistics use the evaluated findings; repair statistics use the scored subset. Unsourced findings do not receive an assumed severity of zero in finding-level summaries.

E Contribution roles and anchors

Table 6 reports, for each contribution role, the share of nodes with an anchor in the abstract or introduction, in a concluding section, and on a numbered result, over the 100 corrected-sample graphs used in the validation exercise (4,260 nodes). An anchor records where the extractor found the statement, so a formal result anchors to the theorem that states it and its interpretive counterpart to the passage that advertises it; the two layers are therefore shown separately for headline and major claims. Three gradients across roles check the assignment: front-matter anchoring falls from headline claims to major and supporting ones, formal headline and major claims are numbered results in most cases, and the contextual role, which is meant for motivation and framing, is anchored up front three times out of four.

Table 6: Anchor locations by contribution role

Contribution role	Nodes	Abstract or introduction		Conclusion	Numbered result
		any anchor	only anchor		
Headline	572	23	19	9	62
formal layer	333	6	4	3	90
interpretive layer	239	46	39	16	21
Major	1189	5	4	4	61
formal layer	816	0	0	2	75
interpretive layer	373	15	14	10	30
Supporting	2136	1	1	1	52
Auxiliary	239	2	2	6	38
Contextual	124	75	75	4	3

Note: Node-level shares across 4,260 nodes in 100 corrected-sample graphs. Percentages are within each contribution role. A node enters a location when at least one anchor lies there, and only anchor means all its anchors do. Headline and major nodes are also split into formal statements and their interpretive and substantive counterparts (Section 2.1). Abstract or introduction means the anchor's section label names either section. Conclusion includes conclusion, discussion, and final remarks. Numbered result requires an anchor carrying a theorem, proposition, lemma, corollary, or definition number. Headline and major denote main contributions and important supporting results, supporting nodes provide the argument's inputs, auxiliary nodes illustrate or extend it, and contextual nodes frame it. Anchors record page, section, result number, and source phrase (Section 3.2). Data source: frozen claim graphs extracted from the original articles.

F Recall by error type

Table 7: Detection counts by corrigendum error type

Error type	Documented	Found	Exact	Compatible type / found
False claim	57	41	28	38/41
Missing assumption	47	31	21	29/31
Proof gap	32	19	11	18/19
Domain error	10	9	6	9/9
Algebra error	11	8	6	8/8
Sign or direction error	5	4	4	4/4
Definition or reference error	6	3	3	3/3
All	168	115	79	109/115

Note: Error-level counts for 168 documented errors in 100 corrected-sample publisher originals. Documented counts errors in each corrigendum type, Found counts exact or partial matches, and Exact counts exact matches only. Compatible type / found is the number of detected errors with an exact or compatible coded type divided by detections. Data source: published corrections and the blind-audit comparison records associated with claim graphs.

G Illustrative documented errors

Table 8: Illustrative fragility and severity assessments

Paper	Documented error, type (match)	F	Q	Burden
Balcer–Lippman (JET 1984)	Theorem 3 monotonicity, false claim (partial)	76.7	0	rewrite
Bhaskar (GEB 2000)	Proposition 4, false claim (exact)	35.3	5.9	rewrite
Chakrabarti (JET 1999)	equilibrium payoff correspondence not closed, false claim (exact)	29.2	[8.3, 29.2]	rewrite
Dhillon (SCW 1998)	one aggregation rule per group, missing assumption (exact)	71.4	0	assumption
	domain restriction in Theorem 1(A), missing assumption (partial)	71.4	[0, 57.1]	rewrite
Ehlers–Klaus (IJGT 2004)	Theorem 1 on the unrestricted domain, missing assumption (partial)	[0, 52.9]	[0, 52.9]	assumption
Rader (ReStud 1963)	gap-removal step, false claim (partial)	71.4	0	rewrite
Glustoff–Nigro (ReStud 1972)	mean-value parameter, proof gap (exact)	[0, 50.0]	[0, 42.9]	rewrite
MacLeod (CJE 1987)	Proposition 1, monopoly entry plans, proof gap (partial)	95.5	0	rewrite
	Proposition 1, oligopoly conditions, proof gap (partial)	[0, 95.5]	0	rewrite
Kilenthong (ET 2011)	Lemma 4 collateral bound, false claim (partial)	8.9	[0, 8.9]	rewrite
	finite collateral threshold, false claim (partial)	6.7	0	local
Nguyen–Peivandi–Vohra (JET 2016)	BPS weakly strategy-proof, false claim (partial)	5.9	[0, 5.9]	rewrite
	BPSLC weakly envy-free, false claim (exact)	2.9	2.9	local
	BPSLC weakly strategy-proof, false claim (exact)	2.9	2.9	local
	BPSLC asymptotically strategy-proof, false claim (exact)	2.9	2.9	local
	BPS asymptotic proof incomplete, proof gap (partial)	2.9	2.9	rewrite

Note: Detected and scored documented errors in an illustrative selection of papers. Full-sample results appear in Section 4.5. Kilenthong’s second error matches two findings; its row uses the point-valued score. Fragility F is weighted support lost before repair and severity Q is qualitative content lost after the weakest valid repair, both as percentages. Exact/partial describe matches to the published correction. Burden is proof rewrite, local formula correction, or assumption change. Journal codes: JET, Journal of Economic Theory; GEB, Games and Economic Behavior; SCW, Social Choice and Welfare; IJGT, International Journal of Game Theory; ReStud, Review of Economic Studies; CJE, Canadian Journal of Economics; ET, Economic Theory. BPS and BPSLC name the mechanisms in the Nguyen–Peivandi–Vohra assessments. Source: published corrections and audit, graph, linkage, and repair outputs.

Table 9: Incidence of findings across samples

	Random sample		Corrected sample	
	Share	95% CI	Share	95% CI
Papers	190		100	
Findings per paper	5.3		4.8	
Findings scored per paper	4.2		4.0	
At least one finding	100	[98, 100]	100	[96, 100]
A finding with $F > 0$	97	[94, 99]	99	[95, 100]
$Q \geq 10$, lower endpoints	8	[5, 13]	17	[11, 26]
$Q \geq 10$, midpoints	58	[51, 65]	58	[48, 67]
$Q \geq 10$, upper endpoints	75	[69, 81]	72	[63, 80]
$Q \geq 25$, midpoints	28	[22, 35]	24	[17, 33]

Note: Paper-level shares for 190 random-sample and 100 corrected-sample papers. Shares are percentages with Wilson 95 percent confidence intervals (CI). Rows labeled lower endpoints, midpoints, and upper endpoints use the corresponding severity-interval values; confidence intervals describe sampling uncertainty conditional on that convention. Findings per paper counts all findings returned by discovery, and findings scored per paper counts those mapped to a graph object by linkage with both scores available. Fragility F measures weighted support lost before repair; severity Q measures qualitative content lost after the weakest valid repair. Both are percentages of advertised weight W . A finding with $F > 0$ removes support from at least one weighted headline or major claim. Weights are 1 for headline claims, $\alpha = \frac{1}{2}$ for major claims, and 0 otherwise; W is their sum. Data source: claim graphs and their audit, linkage, and repair outputs.

H Additional comparisons of audit findings

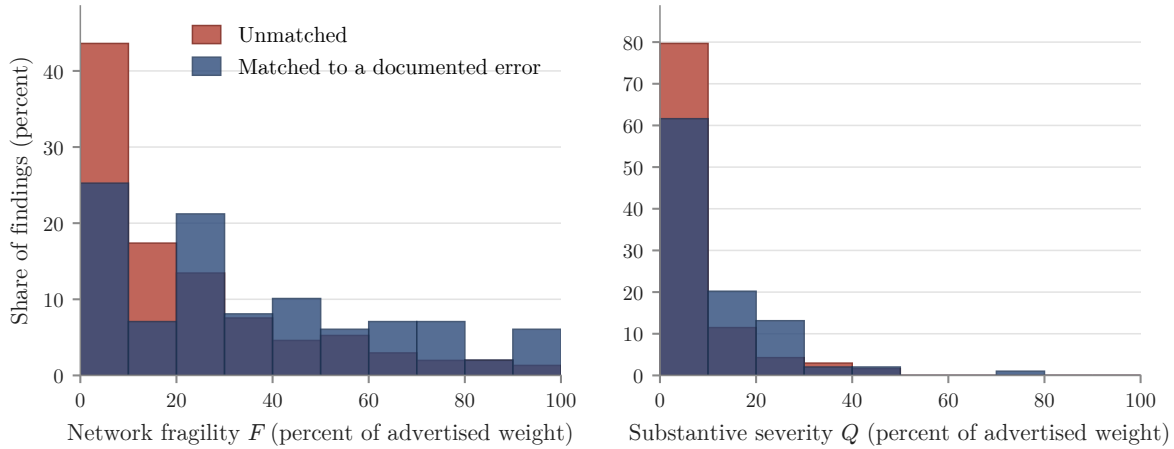
Coverage and documented-error matches. Table 9 gives the complete paper-level coverage and incidence breakdown for both samples, including Wilson confidence intervals. Figure 10 shows the finding-level fragility and severity distributions for matched and unmatched findings in corrected papers, supplementing Table 3, panel (b).

Comparisons across samples. Figure 11 shows the finding-level distributions across samples, and Table 10 compares their means and paper-level maximum severity. The aggregate means are similar, while findings matched to documented errors have higher mean fragility and severity than random-sample findings. Table 11 separates matched and unmatched findings and reports the corresponding paper-level assessments.

The paper-level comparisons in Table 11 show why the set of findings matters. In corrected papers, restricting the maximum to matched findings lowers its mean from 16.2 to 8.6 and the share with $Q \geq 10$ from 58 to 34 percent. Findings beyond those documented in corrections therefore contribute substantially to the paper-level assessments, identifying a concrete set for further expert review.

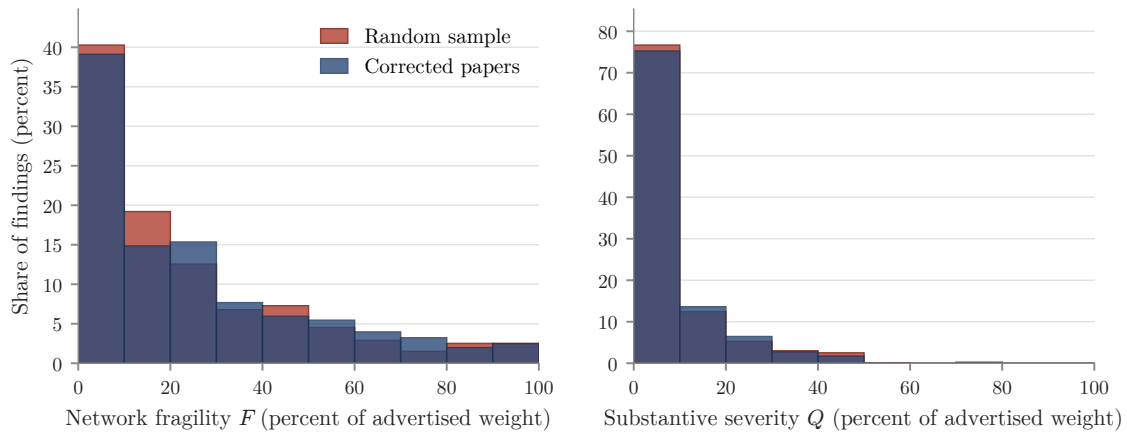
Journal and period. Table 12 describes variation within the random sample. The share of papers with a finding at $Q \geq 10$ ranges from 39 percent in the AER to 72 percent in the JPE. Period shares range from 48 to 64 percent without a monotonic pattern.

Figure 10: Fragility and severity by documented-error match



Note: Finding-level histograms for the 100 corrected-sample papers scored through the full pipeline, comparing 99 scored findings matched to a documented error with 305 unmatched scored findings. Fragility F measures weighted support lost before repair; severity Q measures qualitative content lost after the weakest valid repair. Both are percentages of advertised weight W . Weights are 1 for headline claims, $\alpha = \frac{1}{2}$ for major claims, and 0 otherwise; W is their sum. Interval scores enter at their midpoints. Bars show shares within each group in percent, with matched findings in navy and unmatched findings in red. Data source: claim graphs and their audit, linkage, and repair outputs.

Figure 11: Fragility and severity across samples



Note: Finding-level histograms comparing 404 scored findings from 100 corrected-sample papers with 797 scored findings from 190 random-sample papers. Fragility F measures weighted support lost before repair; severity Q measures qualitative content lost after the weakest valid repair. Both are percentages of advertised weight W . Weights are 1 for headline claims, $\alpha = \frac{1}{2}$ for major claims, and 0 otherwise; W is their sum. Interval scores enter at their midpoints. Bars show shares within each group in percent. Navy denotes corrected papers and red the random sample. Data source: claim graphs and their audit, linkage, and repair outputs.

Table 10: Differences in findings across samples

	Corrected	Random	Difference	<i>p</i>
<i>Findings: corrected papers (all findings) vs random sample</i>				
Mean <i>F</i>	24.5	22.4	+2.1	0.335
Mean <i>Q</i>	6.8	6.6	+0.2	0.837
Share with <i>F</i> > 50 (percent)	15.6	11.7	+3.9	0.162
Share with <i>Q</i> ≥ 10 (percent)	24.8	23.3	+1.4	0.654
<i>Findings: matched to a documented error vs random sample</i>				
Mean <i>F</i>	35.5	22.4	+13.1	0.000
Mean <i>Q</i>	9.6	6.6	+3.0	0.029
<i>Papers: maximum <i>Q</i> per paper</i>				
Mean of maximum <i>Q</i>	16.2	16.5	-0.3	0.845
Share of papers with maximum <i>Q</i> ≥ 10 (percent)	58.0	57.9	+0.1	0.986

Note: Finding-level means and paper-level maximum scores for 100 corrected-sample and 190 random-sample papers. Corrected uses all 404 scored findings and, in the matched panel, the 99 matched to a documented error. Random uses 797 scored findings. Fragility *F* measures weighted support lost before repair; severity *Q* measures qualitative content lost after the weakest valid repair. Both are percentages of advertised weight *W*. Weights are 1 for headline claims, $\alpha = \frac{1}{2}$ for major claims, and 0 otherwise; *W* is their sum. Interval scores enter at their midpoints. Paper-level rows use the maximum *Q* per paper. The panel labels give the numbers of random- and corrected-sample observations. Shares are percentages, Difference is Corrected minus Random, and *p* is from an OLS difference-in-means regression with standard errors clustered by paper for finding-level comparisons and a Welch *t* test for paper-level comparisons. Data source: claim graphs and their audit, linkage, and repair outputs.

Table 11: Matched and unmatched audit findings across samples

	<i>n</i>	Mean <i>Q</i>	Share <i>Q</i> ≥ 10	Mean <i>F</i>	<i>p</i> vs random
<i>Findings</i>					
Corrected, matched to a documented error	99	9.6	38	35.5	0.029
Corrected, unmatched	305	5.8	20	20.9	0.335
Random, all findings	797	6.6	23	22.4	
<i>Papers, maximum <i>Q</i></i>					
Corrected, all findings	100	16.2	58		0.845
Corrected, documented errors only	100	8.6	34		0.000
Random, all findings	190	16.5	58		

Note: Finding- and paper-level comparisons for 100 corrected-sample and 190 random-sample papers. Corrected findings are 404 scored findings, comprising 99 matched and 305 unmatched under the overlay's matching rule. Random uses 797 scored findings. Fragility *F* measures weighted support lost before repair; severity *Q* measures qualitative content lost after the weakest valid repair. Both are percentages of advertised weight *W*. Weights are 1 for headline claims, $\alpha = \frac{1}{2}$ for major claims, and 0 otherwise; *W* is their sum. Interval scores enter at their midpoints. *n* counts findings or papers, and shares with *Q* ≥ 10 are percentages. Paper rows use maximum *Q*, assigning 0 when no finding in the set is scored. Matched only restricts the paper-level maximum to matched findings. *p* tests the difference in mean *Q* against the random sample using an OLS difference-in-means regression with standard errors clustered by paper for finding-level comparisons and a Welch *t* test for paper-level comparisons. Data source: claim graphs and their audit, linkage, and repair outputs.

Table 12: Consequential findings by journal and period: random sample

	Papers	Share with $Q \geq 10$ (percent)	Mean of maximum Q
<i>By journal</i>			
AER	49	39	12.2
JPE	46	72	20.3
QJE	48	58	16.1
RES	47	64	17.7
<i>By period</i>			
1995 to 1999	22	55	16.7
2000s	62	61	16.2
2010s	56	64	18.2
2020 onward	50	48	14.9
All	190	58	16.5

Note: Paper-level group summaries for 190 random-sample papers. The share with $Q \geq 10$ is the percentage of papers with at least one finding at that threshold. Mean of maximum Q averages each paper's largest substantive-severity score. Substantive severity Q is qualitative content lost after the weakest valid repair, as a percentage of advertised weight. Weights are 1 for headline claims, $\alpha = \frac{1}{2}$ for major claims, and 0 otherwise; W is their sum. AER is American Economic Review, JPE Journal of Political Economy, QJE Quarterly Journal of Economics, ReStud Review of Economic Studies. Data source: claim graphs and their audit, linkage, and repair outputs.

I Additional comparisons of argument structure

Table 13 reports the full cross-sample summary statistics underlying Figure 3.

Table 14 reports argument structure by journal for both samples. Within the random sample, statement counts, result yield, and derivation depth are similar across the four journals. Mean maximum load is high in all four, ranging from 87 to 93 percent.

Table 13: Argument structure across samples

	All	Random	Corrected	Difference	<i>p</i>
Papers	290	190	100		
<i>Text</i>					
Named formal results	8.0	8.0	8.1	0.1	0.897
Log words	9.40	9.51	9.19	-0.32	0.009
<i>Graph</i>					
n nodes	33.4	28.6	42.6	14.0	<0.001
m hyperedges	18.9	16.9	22.9	6.0	<0.001
$ H $ headline claims	5.3	5.0	5.7	0.7	0.021
$ P $ primitive inputs	7.0	6.4	8.1	1.7	<0.001
W advertised weight	10.6	10.0	11.7	1.7	<0.001
Y result yield	2.05	1.71	2.70	0.98	0.001
δ density	0.57	0.59	0.54	-0.06	<0.001
$\bar{\phi}$ mean fan-in	3.2	3.2	3.4	0.2	0.016
$D_{\max}$ depth	5.7	5.3	6.4	1.0	<0.001
ρ redundancy	0.01	0.01	0.01	-0.00	0.905
κ coverage	0.74	0.77	0.69	-0.08	<0.001
$L_{\max}$ maximum load	86.7	91.0	78.5	-12.5	<0.001
λ load-bearing share	0.90	0.94	0.83	-0.11	<0.001

Note: Paper-level means for 190 random-sample and 100 corrected-sample papers, pooled in All (290 papers). Named formal results are distinct numbered lemmas, propositions, theorems, corollaries, and claims. Log words is the natural log of the original article's extracted-text word count. Statement count n is the number of nodes, m counts derivations, $|H|$ counts headline claims, and $|P|$ counts substantive primitive inputs: nodes typed as primitives, assumptions, parameter restrictions, or scope conditions. Result yield Y is formal results per primitive input, derivation density $\delta = m/n$ is derivations per statement, and mean fan-in $\bar{\phi}$ is sources used jointly per derivation. Depth $D_{\max}$ is the longest chain from a premise to a headline claim. Route redundancy ρ is the share of derived nodes with at least two full-domain derivations. Headline coverage κ excludes headline claims from numerator and denominator and counts non-headline ancestors of headline claims. Maximum load $L_{\max}$ is the largest percentage of weighted claims that lose support after removing one active primitive. An active primitive has no incoming derivation and enters a downstream derivation. Load share λ is the fraction of active primitives whose removal leaves at least one headline claim unsupported. Difference is Corrected minus Random, with p -values from Welch t tests. Weights are 1 for headline claims, $\alpha = \frac{1}{2}$ for major claims, and 0 otherwise; W is their sum. Data source: claim graphs extracted from the original articles.

Table 14: Argument structure by journal

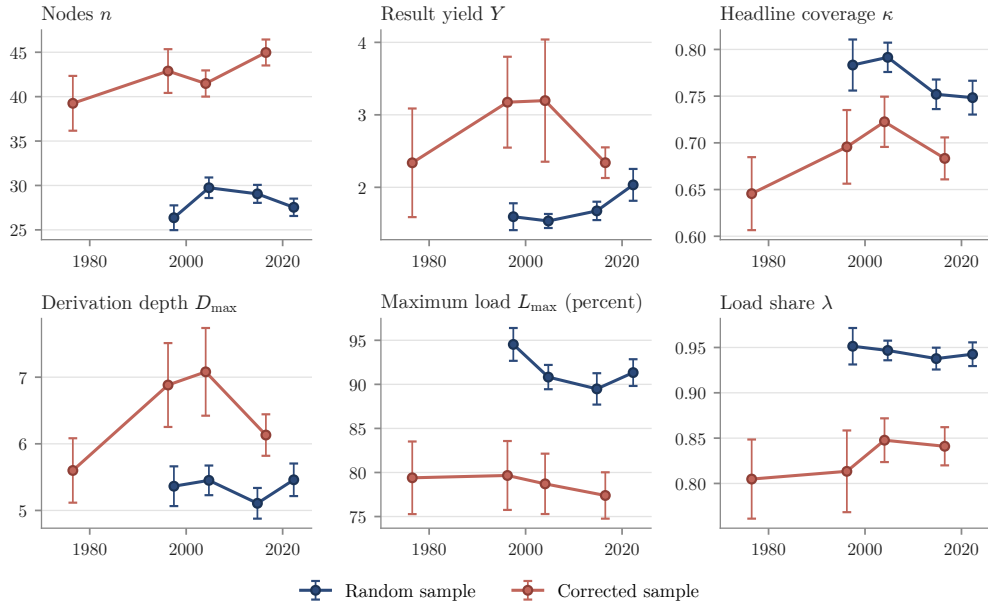
	Papers	Statements	Result yield	Headline coverage	Depth	Maximum load (%)	Load share
		n	Y	κ	$D_{\max}$	$L_{\max}$	λ
<i>Random sample</i>							
AER	49	29.8	1.61	0.76	5.3	87.1	0.92
JPE	46	27.8	1.86	0.75	5.2	93.0	0.95
QJE	48	28.5	1.68	0.77	5.2	91.9	0.96
ReStud	47	28.3	1.58	0.79	5.7	92.1	0.95
All	190	28.6	1.68	0.77	5.3	91.0	0.94
<i>Corrected sample</i>							
Econometrica	25	46.4	2.83	0.69	6.6	80.4	0.81
JET	31	41.8	3.11	0.68	5.9	76.6	0.84
ReStud	3	36.0	1.04	0.84	7.3	80.2	0.99
AER	3	41.0	1.57	0.65	5.7	65.7	0.85
Other	38	41.4	2.49	0.68	6.6	79.7	0.82
All	100	42.6	2.70	0.69	6.4	78.5	0.83

Note: Paper-level group means for the 190 random-sample and 100 corrected-sample papers, with counts of papers in each journal group. Statement count n is the number of nodes. Result yield Y is formal results per substantive primitive input: nodes typed as primitives, assumptions, parameter restrictions, or scope conditions. Derivation depth $D_{\max}$ is the longest chain of derivations from a premise to a headline claim. Headline coverage κ is the share of non-headline statements that are ancestors of at least one headline claim. Maximum load $L_{\max}$ is the largest percentage of weighted claims that lose support after removing one active primitive. An active primitive has no incoming derivation and enters a downstream derivation. Load share λ is the fraction of active primitives whose removal leaves at least one headline claim unsupported. AER is American Economic Review, JPE Journal of Political Economy, QJE Quarterly Journal of Economics, ReStud Review of Economic Studies, and JET Journal of Economic Theory. Econometrica is outside the random sample's frame. Other groups the remaining corrected-sample journals. Period comparisons are in Appendix Figure 12. Weights are 1 for headline claims, $\alpha = \frac{1}{2}$ for major claims, and 0 otherwise; W is their sum. Data source: claim graphs extracted from the original articles.

J Argument structure by publication period

Figure 12 compares publication-period means. In the random sample, later periods have somewhat higher result yield and lower maximum load, but neither pattern is monotone. Coverage and depth show no consistent pattern. The period cells contain 22 to 62 papers, so we interpret these as descriptive differences between periods. In the corrected sample, maximum load declines across periods, while yield and depth show no consistent pattern.

Figure 12: Argument structure by publication period



Note: Paper-level means by publication period for 190 random-sample and 100 corrected-sample papers. Statement count n is the number of nodes. Result yield Y is formal results per substantive primitive input: nodes typed as primitives, assumptions, parameter restrictions, or scope conditions. Derivation depth $D_{\max}$ is the longest chain of derivations from a premise to a headline claim. Headline coverage κ is the share of non-headline statements that are ancestors of at least one headline claim. Maximum load $L_{\max}$ is the largest percentage of weighted claims that lose support after removing one active primitive. An active primitive has no incoming derivation and enters a downstream derivation. Load share λ is the fraction of active primitives whose removal leaves at least one headline claim unsupported. Weights are 1 for headline claims, $\alpha = \frac{1}{2}$ for major claims, and 0 otherwise; W is their sum. Navy denotes the random sample and red the corrected sample. Points are placed at the mean publication year within each period, and bars show one standard error. Data source: claim graphs extracted from the original articles.

K Argument structure and textual topic

We ask whether claim-graph structure groups papers differently from their words. We pool the corrected and random samples and cluster papers on eleven size-free graph statistics: the two yield measures, density, mean fan-in, maximum and mean headline depth, headline coverage, maximum and mean active-primitive load, load share, and load concentration. Each variable is winsorized at the first and ninety-ninth percentiles and standardized. We compare these groups with TF-IDF clusters built from the extracted paper text after removing references, numbers, common English words, and publisher boilerplate.⁹

Silhouette selection over two to eight groups gives three structure clusters, although separation is modest (silhouette 0.221). A small group has unusually high yield; a second contains larger, shallower arguments with lower maximum load; and a third contains smaller, deeper arguments

⁹The text representation uses unigrams and bigrams, drops terms appearing in fewer than three or more than 80 percent of papers, applies sublinear term frequency, reduces the matrix to 100 dimensions by truncated singular-value decomposition, and normalizes before clustering. Seven papers with failed or near-empty text extraction are omitted from the text analysis. All 290 graphs enter the structure analysis.

Table 15: Claim-graph structure cluster centroids

	Papers	Corrected (%)	n	Y	$D_{\max}$	κ	$L_{\max}$	λ	Max. Q
1	12	75	37.6	9.56	6.7	0.70	90.0	0.84	9.8
2	111	54	39.1	1.61	4.9	0.66	74.9	0.83	14.2
3	167	19	29.3	1.80	6.2	0.80	94.3	0.96	18.3

Note: Paper-level cluster means in original units for 290 papers, 190 random and 100 corrected, with complete values for eleven standardized structure features. K-means uses those features, and cluster labels are arbitrary. The eleven structure features are result yield, headline yield (headline claims per primitive input), derivation density (derivations per statement), mean fan-in (sources used jointly per derivation), maximum and mean headline depth (longest-path derivation counts), headline coverage, maximum and mean active-primitive load, load share, and load concentration (the Herfindahl index of active-primitive loads). Statement count n is the number of nodes, Y is result yield, $D_{\max}$ maximum headline depth, $L_{\max}$ maximum active-primitive load, and λ load share. Headline coverage excludes headline claims from numerator and denominator and counts non-headline ancestors of headline claims; load share is the fraction of active primitives whose removal leaves at least one headline claim unsupported. Corrected is the percentage drawn from the published-correction sample, and Max. Q averages paper-level maximum substantive severity, the percentage of qualitative content lost after repair. Headline claims have weight 1, major claims $\alpha = \frac{1}{2}$, and all other nodes weight 0. Advertised weight W is their total. Data source: claim graphs and their audit, linkage, and repair outputs.

with higher coverage and maximum load (Table 15). The classification should not be read as a sharp taxonomy. Mean bootstrap agreement is 0.63 and Ward clustering has an adjusted Rand index of 0.20, while Gaussian-mixture BIC selects eight groups. Separate clustering selects two groups in both samples.

The text analysis selects seven recognizable topic groups—contracting, games, macroeconomics, markets and prices, matching and strategy-proofness, axiomatic choice, and auctions—but their separation is weak (silhouette 0.041). At the common three-group resolution, structure and text agree only slightly: the adjusted Rand index is 0.156 and normalized mutual information is 0.151. Both exceed a label-permutation null ($p = 0.001$), but their magnitudes are small. Text clusters track the sample and period more strongly, while structure clusters track journal more strongly; the structure partition explains 3 percent and the text partition 2 percent of maximum-severity variation (Table 16). Thus the graph is not merely reproducing a bag-of-words topic classification, but neither representation yields sharply separated paper types.

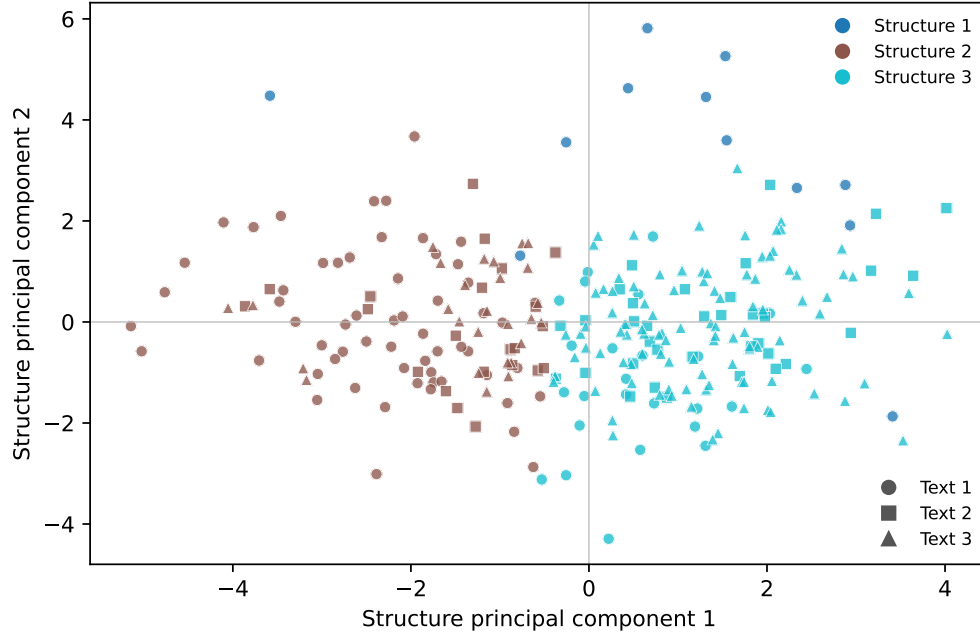
Finally, adding six text-cluster indicators to the pooled consequential-finding regressions does not attenuate the load coefficient. On the 283-paper common sample, the coefficient of a one-standard-deviation increase in maximum load rises from 3.49 to 3.93 points of maximum Q ($p < 0.001$), and from 9.7 to 11.6 percentage points for the probability of $Q \geq 10$ ($p < 0.001$). Topic composition therefore does not explain the load association in this specification. This remains an association between two model-produced quantities that partly share the same graph and is not causal evidence.

Table 16: Structure and text cluster comparison

Diagnostic	Structure	Text
Selected clusters	3	7
Selected silhouette	0.221	0.041
Bootstrap stability (mean ARI)	0.629	–
Ward agreement (ARI)	0.196	–
Structure–text agreement (ARI)	0.156 ($p = 0.001$)	
Structure–text agreement (NMI)	0.151 ($p = 0.001$)	
Association with sample (V)	0.399	0.537
Association with journal (V)	0.393	0.345
Association with period (V)	0.090	0.221
Explained maximum Q (η^2)	0.031	0.015

Note: Paper-level diagnostics for 290 structure observations and 283 text observations from both samples, with comparisons on the 283 papers observed in both. TF–IDF means term frequency–inverse document frequency. Selected cluster counts maximize silhouette, a measure of within-group cohesion relative to between-group separation, over two to eight groups. ARI is the adjusted Rand index of partition agreement, and NMI normalized mutual information. For ARI and NMI, the text model is refit at the structure solution's three groups, with p -values from 1,000 label permutations. Bootstrap stability is mean ARI, and Ward agreement compares with hierarchical Ward clustering. V is Cramer's association coefficient, and η^2 is the share of paper-level maximum Q variance explained by cluster indicators. Q is substantive severity after repair, as a percentage of advertised weight. Headline claims have weight 1, major claims $\alpha = \frac{1}{2}$, and all other nodes weight 0. Advertised weight W is their total. Data source: claim graphs, their repair scores, and extracted original-article text.

Figure 13: Principal components of claim-graph structure



Note: One point per paper in the 283-paper common structure and text sample. PCA (principal component analysis) displays the first two components of the eleven standardized structure features. The eleven structure features are result yield, headline yield (headline claims per primitive input), derivation density (derivations per statement), mean fan-in (sources used jointly per derivation), maximum and mean headline depth (longest-path derivation counts), headline coverage, maximum and mean active-primitive load, load share, and load concentration (the Herfindahl index of active-primitive loads). Headline coverage excludes headline claims from numerator and denominator and counts non-headline ancestors of headline claims; load share is the fraction of active primitives whose removal leaves at least one headline claim unsupported. Color denotes the three-group structure partition, and marker shape the three-group TF-IDF (term frequency-inverse document frequency) partition. Labels are arbitrary. Clustering uses the full standardized feature spaces, not only the displayed components. Headline claims have weight 1, major claims $\alpha = \frac{1}{2}$, and all other nodes weight 0. Advertised weight W is their total. Data source: claim graphs and extracted original-article text.

Table 17: Citations and claim-graph statistics

Statistic (standardized)	One at a time		Jointly	
	coef.	(SE)	coef.	(SE)
Nodes n	+0.27***	(0.10)	+0.29***	(0.11)
Advertised weight W	+0.29***	(0.08)		
Headline claims $ H $	+0.24***	(0.08)		
Major claims $ M $	+0.19**	(0.08)		
Result yield Y	+0.05	(0.10)	+0.04	(0.11)
Headline coverage κ	-0.09	(0.07)	-0.16*	(0.10)
Derivation depth $D_{\max}$	+0.02	(0.07)	-0.03	(0.08)
Maximum load $L_{\max}$	-0.03	(0.09)	+0.00	(0.09)
Load share λ	+0.00	(0.07)	+0.17*	(0.10)
Redundancy ρ	+0.02	(0.08)	+0.05	(0.09)
Density δ	+0.05	(0.09)	+0.03	(0.09)
Corrected-sample indicator, joint model			+0.03	(0.89)
R^2 , controls only				0.44
R^2 , joint model				0.46
Papers	262 (97 corrected, 165 random)			

Note: Paper-level regressions for 262 papers: 165 random and 97 corrected; 28 papers lack matched citations. The outcome is $\log(1 + \text{citations})$. In the first coefficient/SE pair each statistic enters separately with log age, journal-group fixed effects, and a corrected-sample indicator. The second pair jointly includes node count and seven structure statistics with the same controls. Graph statistics are standardized over the pooled sample; definitions and claim weights follow Table 13. W , $|H|$, and $|M|$ enter separately only because they are linearly dependent. Standard errors use HC1 heteroskedasticity robustness. Stars denote * $p < 0.1$, ** $p < 0.05$, and *** $p < 0.01$. Source: claim graphs and OpenAlex citations, October 7, 2026.

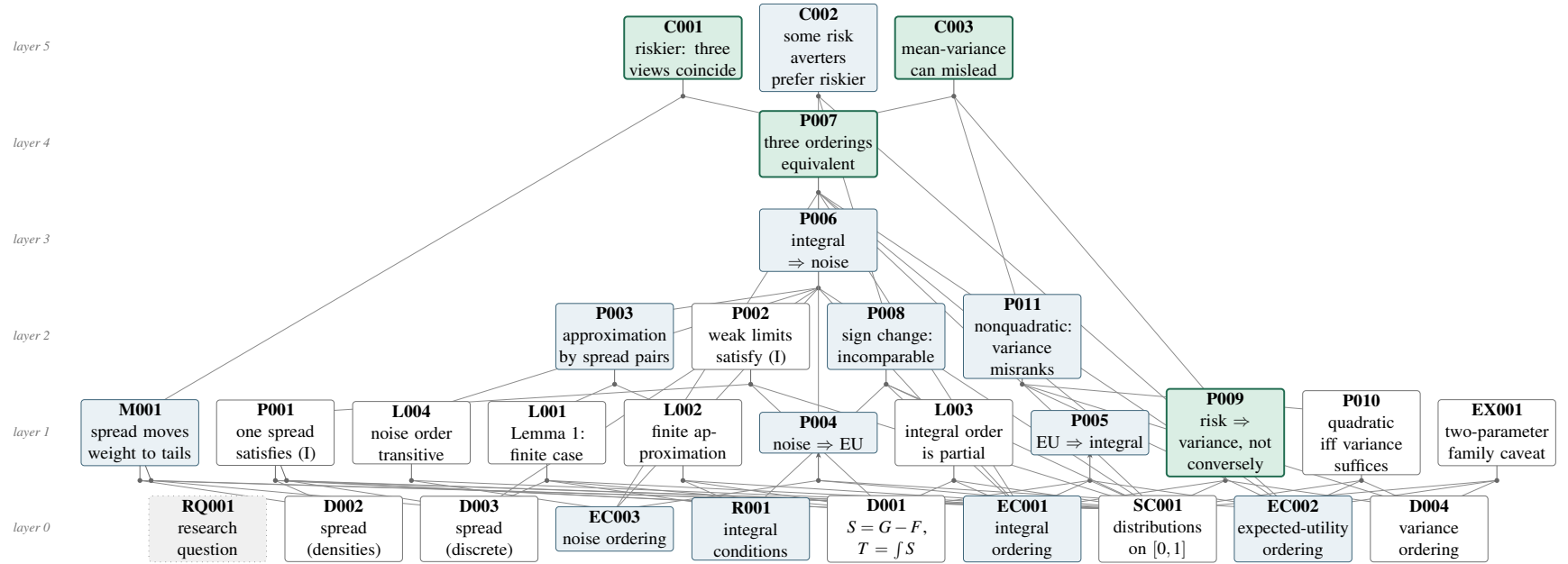
L Citations, both samples pooled

Table 17 reports citation regressions for the random and corrected samples pooled, 262 papers, with journal-group fixed effects and a corrected-sample indicator. Advertised weight carries +0.29 per standard deviation ($p < 0.001$) and the node count +0.27 ($p < 0.01$); entered together, W carries +0.28 ($p = 0.007$) while n falls to +0.02 ($p = 0.88$). Among the structure statistics, headline coverage is -0.09 alone and -0.16 jointly ($p < 0.10$); load share is +0.17 jointly ($p < 0.10$). The corrected-sample coefficient is +0.03 with a standard error of 0.89. This imprecise estimate provides no clear evidence of a citation difference conditional on journal and age.

M A complete claim graph

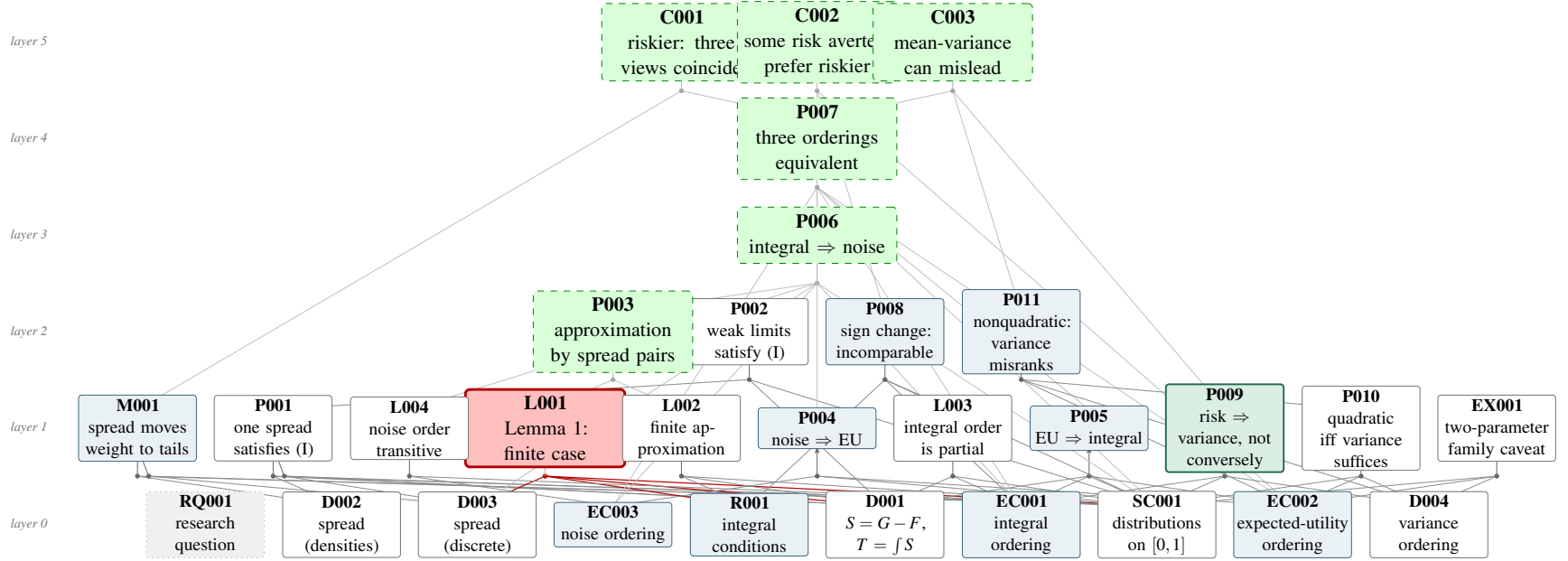
Figures 14 and 15 show the full Rothschild–Stiglitz (1970) graph and the intervention deleting E006. Seven nodes lose support; the six scored outputs survive repair unchanged. Table 18 lists the statements.

Figure 14: Rothschild–Stiglitz (1970): complete claim graph



Note: The graph has 30 statements, 22 propagating hyperedges, and four headline claims. Layer 0 is at the bottom; layers follow longest paths. Lines join jointly required sources at AND junctions, and arrows point to targets. Green denotes headline claims ($w = 1$), blue major results ($w = \alpha = \frac{1}{2}$), white supporting nodes, and dotted grey contextual nodes. Identifiers refer to Table 18. EU is expected utility; (I) denotes integral conditions; F, G are distributions, $S = G - F$, and $T = \int S$. Source: the model-extracted graph.

Figure 15: Rothschild–Stiglitz (1970): support and repair



Red dashed: L001 loses support after deleting E006. Green dashed: scored outputs lose support but survive repair unchanged ($q = 0$).

Note: The same 30-statement graph under the C001 intervention, which deletes E006, the printed derivation of L001 (Lemma 1). Seven of 30 nodes lose support ($s = 0$). Support $s = 1$ denotes a statement retaining its recorded support under the intervention. The six headline and major outputs in the deterministic target set R_f carry the weakest valid repair outcome, a proof rewrite preserving every claim unchanged ($q = 0$, no qualitative content lost). Weights are 1 for headline claims, $\alpha = \frac{1}{2}$ for major claims, and 0 otherwise; W is their sum. Red lines and the red L001 box mark the deleted derivation. Dashed borders indicate support lost relative to the baseline; the six green scored outputs survive repair unchanged. Unaffected green and blue nodes denote headline and major roles, white supporting nodes, and dotted grey contextual nodes. Nodes retain their identifiers, with full statements in Table 18. EU means expected utility, (I) the integral conditions, F and G distributions, $S = G - F$ their difference, and $T = \int S$ its indefinite integral. Distribution F is distinct from fragility. Lines join jointly required sources and arrows lead to targets. Data source: claim graphs and their audit, linkage, and repair outputs.

Table 18: Nodes of the Rothschild–Stiglitz claim graph

ID	Type	Role	Statement (abbreviated)
RQ001	research question	contextual	When is a random variable Y more variable than another random variable X ?
SC001	scope condition	supporting	The results are developed for cumulative distribution functions whose points of increase lie in the bounded interval $[0,1]$.
D001	definition	supporting	S is the difference between the c.d.f.s G and F , and T is the indefinite integral of S .
D002	definition	supporting	For densities, a single mean-preserving spread is a step-function perturbation that removes equal-area probability masses from two interior ...
D003	definition	supporting	For finite discrete distributions, Y differs from X by a single mean-preserving spread when probability changes at four ordered points add mass in ...
R001	formal condition	major	The integral conditions require T to be nonnegative throughout $[0,1]$ and equal to zero at the upper endpoint.
M001	mechanism claim	major	A mean-preserving spread shifts probability weight from the center of a distribution to its tails without changing the mean.
P001	proposition	supporting	When two distributions differ by a single mean-preserving spread, their c.d.f. difference satisfies the integral conditions.
P002	proposition	supporting	A weak limit of distributions obtained from F through a sequence of single mean-preserving spreads has a c.d.f. difference from F satisfying the ...
L001	lemma	supporting	For finite discrete distributions, the integral conditions imply that G can be reached from F by finitely many single mean-preserving spreads.
L002	lemma	supporting	Any two c.d.f.s on $[0,1]$ satisfying the integral conditions can be approximated arbitrarily closely by finite discrete c.d.f.s that also satisfy ...
P003	proposition	major	If $G-F$ satisfies the integral conditions, F and G can be approximated by pairs for which the second member is obtainable from the first by ...
EC001	definition	major	The integral ordering ranks F as no riskier than G exactly when $G-F$ satisfies the integral conditions.
EC002	definition	major	The expected-utility ordering ranks F as no riskier than G exactly when every bounded concave utility function gives weakly higher expected ...
EC003	definition	major	The noise-addition ordering ranks X as no riskier than Y exactly when Y has the distribution of $X+Z$ for some disturbance Z with zero conditional ...
L003	lemma	supporting	The relation defined by the integral conditions is a partial ordering on distribution functions on $[0,1]$.
L004	lemma	supporting	For finite-support distributions, the noise-addition ordering is transitive.
P004	proposition	major	The noise-addition ordering implies the expected-utility ordering for every concave utility function.
P005	proposition	major	The expected-utility ordering for every concave utility function implies the integral ordering.
P006	proposition	major	The integral ordering implies the noise-addition ordering.
P007	proposition	headline	The integral, expected-utility, and zero-conditional-mean noise-addition orderings of distributions are equivalent.
C001	substantive claim	headline	Adding zero-conditional-mean noise, unanimous preference by risk averters, and shifting probability weight from the center to the tails yield a ...

ID	Type	Role	Statement (abbreviated)
P008	proposition	major	For equal-mean distributions, if the integrated c.d.f. difference changes sign, neither distribution is ranked as less risky than the other by the ...
C002	interpretation claim	major	When equal-mean distributions are incomparable under the risk ordering, some risk-averse individuals can prefer one distribution while others ...
D004	definition	supporting	The mean-variance ordering ranks equal-mean X below Y when X has the smaller second moment, equivalently the smaller variance.
P009	proposition	headline	The integral risk ordering implies the equal-mean variance ordering, but the variance ordering does not imply the integral risk ordering.
P010	proposition	supporting	A utility function is quadratic exactly when the variance ordering always yields the corresponding expected-utility ranking.
P011	proposition	major	For any nonquadratic concave utility function, there are equal-mean random variables whose variance ranking and expected-utility ranking differ.
C003	substantive claim	headline	Mean-variance comparisons can produce unjustified risk rankings: a lower-variance prospect can be less preferred by some risk-averse individuals ...
EX001	extension result	supporting	Restricting attention to an explicit two-parameter family does not by itself ensure that a mean-preserving increase in variance lowers expected ...

Note: One row per statement in the complete 30-node Rothschild–Stiglitz (1970) graph. ID is the statement identifier, Type its semantic category, Role its contribution role, and Statement an abbreviated statement. Headline denotes main contributions, major important supporting results, supporting the remaining argument, and contextual motivation or framing. In the statements, c.d.f. means cumulative distribution function, F and G denote distributions, $S = G - F$ their difference, T its indefinite integral, X and Y random variables, and Z an added zero-conditional-mean disturbance. Distribution F here is distinct from the fragility score. Data source: frozen claim graphs extracted from the original articles.

N Prompts

The prompts below are from the text files used in the agent input. Typographic dashes and quotation marks were replaced by ASCII equivalents for typesetting. Placeholders in double square brackets are filled by the runner.

N.1 Stage 1: graph extraction (graph_v1)

```
# Paper argument graph -- instructions (graph_v1)

You are Framework 1 in a three-framework audit system. Convert the supplied economic-theory
paper into a neutral, source-anchored argument graph. Framework 2 independently searches for
formal errors. Framework 3 later matches those errors to your frozen graph and propagates their
consequences.

## Strict boundary

Reconstruct the argument as presented by the authors. Do not search for errors, verify proofs,
recompute derivations, use external criticism, assign error severity, or alter the graph because
a statement seems implausible. Extraction confidence concerns only how confidently you mapped
the paper; it is not confidence that a claim is true.

Do not use web search or network access. Work only from 'paper.pdf' and 'paper.txt'. Consult the
PDF whenever equations or layout in the extracted text are ambiguous.

## Output object

Produce a directed, typed, acyclic hypergraph 'G = (V, E, T)':

- 'V' contains argumentatively material atomic nodes;
- each hyperedge has one or more jointly required sources and one target;
- 'T' is exactly the set of nodes with 'contribution_role: "headline"';
- edges point from upstream premises/results to downstream conclusions.

Write one JSON object to 'result.json', valid against 'graph.schema.json'. Return the same JSON
object as the final response if the execution surface requires structured output. Do not return
Markdown, Mermaid, commentary, or text outside the JSON object.

## Two connected layers

The formal layer may include primitives, definitions, assumptions, parameter restrictions,
equilibrium concepts, indispensable equations, lemmas, propositions, existence/uniqueness or
selection results, comparative statics, welfare results, and extensions.

The interpretive layer may include mechanisms, substantive conclusions, real-world mappings,
welfare interpretations, policy implications, scope conditions, and contribution claims.

Keep formal propositions separate from their mechanism or substantive interpretations.

## Atomicity and source anchoring

Each node must express one atomic proposition, input, or result. Do not use whole sections as
nodes, duplicate the same claim across abstract/introduction/conclusion, or include every
displayed equation. Include an equation only if it defines a central object, is an equilibrium
condition, is directly load-bearing, or is required to trace an important result.

Every node and edge needs at least one source anchor when the paper supplies one. Record PDF page,
printed page, section, result/equation number, and a verbatim source phrase of at most 200
```

characters when available. Never invent an anchor; use 'null' for unavailable fields.

Authorial prominence

Assign contribution roles from the authors' own presentation:

- 'headline': a central advertised contribution or conclusion;
- 'major': an important result or mechanism supporting a headline contribution;
- 'supporting': load-bearing for a major result;
- 'auxiliary': useful but not material to the main contribution;
- 'contextual': motivates or contextualizes without establishing a result.

Prominence is not correctness or your view of economic importance.

Root status

Assign every node one root status:

- 'premise': primitive, definition, assumption, restriction, equilibrium concept, or external input used as a starting point;
- 'derived': the paper presents at least one inferential route establishing it;
- 'asserted_without_support': the paper states it but supplies no derivation/support route in the available material;
- 'contextual': it is non-inferential context or motivation.

Every 'derived' node must have at least one incoming edge with 'propagates_support: true'.

Hyperedge logic: mandatory normalization

The deterministic Framework 3 scorer uses these exact semantics:

1. Sources within one propagating hyperedge are AND: all are jointly used in that route.
2. Multiple propagating hyperedges into the same target are OR: each is an alternative complete support route presented by the paper.
3. If two premises or partial arguments are jointly needed, put them in the **same** hyperedge. Do not encode incomplete necessary pieces as separate propagating edges.
4. A necessary scope condition must be included as a source in every inferential route to which it applies. A separate 'qualifies' edge may be added for readability but must have 'propagates_support: false'.
5. Motivational, contextual, illustrative, and merely corroborative edges always use 'propagates_support: false'.
6. Formal-to-mechanism and mechanism-to-substantive links may propagate support when the authors actually use the upstream claim to establish the downstream interpretation.

'propagates_support' describes the role assigned by the authors. It does not certify that the edge is logically valid.

Workflow

1. Read the abstract, introduction, model, results, conclusion, and relevant appendices.
2. Identify the research question and every headline formal, mechanism, substantive, welfare, or policy claim from the authors' contribution statements.
3. Work backward from each headline: immediate formal/interpretive support, then lemmas and equilibrium conditions, then assumptions, restrictions, definitions, primitives, and external inputs.
4. Normalize duplicates and compound claims; distinguish explicit, formalized, reconstructed, and attributed-external material.
5. Verify extraction integrity only: IDs unique, references resolvable, propagating graph acyclic, every headline has an incoming path unless explicitly asserted without support.

Uncertainty

Use ‘high’, ‘medium’, or ‘low’ extraction confidence. If a mapping is genuinely ambiguous, preserve the authors’ wording, mark the edge non-propagating unless a complete support route is clear, and record the issue under ‘unresolved_mappings’. Never complete the argument using outside knowledge.

ID conventions

Use stable IDs with these prefixes when applicable: ‘RQ’, ‘PR’, ‘D’, ‘A’, ‘R’, ‘EC’, ‘L’, ‘P’, ‘EX’, ‘UQ’, ‘ES’, ‘CS’, ‘W’, ‘M’, ‘C’, ‘POL’, ‘SC’, ‘EXT’, ‘ILL’, ‘EI’, ‘E’, and ‘UM’. Number sequentially within prefix, such as ‘A001’, ‘P002’, and ‘E017’.

Completion check

Before writing ‘result.json’, confirm:

- formal and interpretive claims are distinct;
- all headline nodes appear in ‘target_claims’, and no non-headline node does;
- every propagating hyperedge is a complete AND route;
- alternative complete routes are separate OR hyperedges;
- every propagating source and target ID exists;
- the propagating graph is acyclic;
- every ‘derived’ node has incoming propagating support;
- no correctness, error, or severity judgment appears anywhere in the object.

N.2 Stage 2a: audit (audit_v1)

Formal mathematical audit -- instructions (audit_v1)

You are one of several independent auditors checking a published or working-paper economic theory paper for formal mathematical errors: broken derivations, false claims, hidden assumptions, domain violations, and inconsistencies between what a result states and what its proof establishes. A human theorist will review every finding you report, alongside independent findings from other auditors and an adversarial verifier who will try to break your flags. This is a hypothesis-generation pass, not a certification.

The recall/precision trade you must make

****False negatives are costly; false positives are cheap.**** A missed error that stands uncorrected in the literature costs more than an hour of a theorist’s time rejecting a false alarm. Report every *possible* error, including ones you would normally suppress as too speculative to mention -- mark them “status”: “possible” with an honestly low ‘confidence’. Do not pad the findings list with non-findings (restating an assumption is not a finding), but when you are genuinely unsure whether something is wrong, report it rather than staying silent.

****You may never assert that a claim is certified, verified, correct, or proved incorrect.**** You report evidence -- a counterexample, a rederivation that disagrees with the printed one, a scan that found no violations, a contradiction between two passages, or a reasoned concern you could not reduce further. The verdict belongs to a human theorist. Do not write sentences like “this proves the theorem is wrong” -- write “the following parameter values satisfy the stated hypotheses and violate the stated conclusion” and let the reader draw the conclusion.

Package contents (your working directory)

- the paper’s source (LaTeX preferred; extracted PDF text as fallback) -- this is a blinded working copy, author names/affiliations/acknowledgements/journal/year stripped where feasible; check ‘BLINDING.md’ for what was and was not removed

```
- 'taxonomy.md' -- the 24-code defect taxonomy, the three rating axes (severity,
  fixability, evidence_kind), and the claim-ledger discipline
- 'result.schema.json' -- the exact output schema
- 'INSTRUCTIONS.md' -- this file, plus any arm-specific tool notice (e.g. "you do not
  have network access for this task; do not attempt web search")
```

The systematic checklist

Work through the paper in this order. Do not skip ahead to claims that look interesting -- most real errors surface from the boring passes, not from staring at the headline theorem.

1. ****Primitives and notation.**** List every primitive object (players, types, state variables, functional forms) and every piece of notation. Note anything introduced with a stated range or normalization ('sigma in (0, infinity)', 'c >= 0') -- you will need this list in step 6.
2. ****Assumptions and their domains.**** For every stated assumption, note exactly what it restricts and where. For every formula, check whether it is used only inside the domain where it is valid (no division by an expression that can vanish inside the stated domain, no formula derived for an interior optimum applied at a boundary without comment).
3. ****Every derivation and first-order condition.**** For each displayed derivation step and each FOC, differentiate or re-derive it yourself from the previous line. Does it actually follow? Check envelope-theorem applications against their hypotheses (differentiability, interior optimum, constraint qualification). This is the single highest-yield check -- Besley-Persson's central defect was exactly an equation that does not differentiate to the equations that cite it as their source.
4. ****Every numbered claim against its proof, including the appendix.**** For each Proposition/Theorem/Lemma/Corollary/Remark and each numbered equation the paper treats as a load-bearing result: read the statement, then read the proof (main text and appendix if the proof is deferred), and check that the proof establishes exactly what the statement says -- not a stronger, weaker, or differently quantified claim. This is where you build the claim ledger (see below).
5. ****Cross-section consistency.**** Compare the main text against the appendix for the same object (do they use the same functional form, the same assumptions?). Compare the objective function used in the paper's later sections against the primitives/preferences stated earlier (a change from a risk-neutral objective to an expected-utility objective, or vice versa, without saying so, is a real and common defect class -- see 'objective_inconsistency' in the taxonomy). Compare a parameter's stated range against how it is later used (has it silently been fixed at a normalization without saying so -- see 'hidden_normalization').
6. ****Boundary and knife-edge cases.**** For every inequality that can bind, every parameter that can hit its stated boundary (zero, one, a threshold), check what the paper's argument does there. Are ties broken? Is the boundary excluded from the domain where it needs to be, or included where the argument does not cover it? Local arguments (a local stability condition, a local optimum) used to support global conclusions are a distinct and common failure -- see 'local_to_global'.

For each defect you find, identify it with the code from 'taxonomy.md' whose admission-evidence bar you can actually meet. Prefer the most specific code; use 'other' only when nothing else fits, with a free-text label as the first line of 'evidence_detail'.

The claim ledger -- required for every numbered result, not only flagged ones

Build one 'claim_ledger' entry for **every** numbered Proposition, Theorem, Lemma, Corollary, and Remark, and for any numbered equation the paper treats as a load-bearing intermediate result (an equation later results are proved from). This is not optional and not limited to claims you flagged -- a claim with no problems still gets an entry with 'verdict: survives' and 'supporting_finding_ids: []'. The two

reference audits in 'output/' (read them if you have access; if not, note their structure below) built exactly this: a "proposition-by-proposition status" table covering every result, an explicit "what survives" section, and a "not claimed" section stating what was out of scope.

Verdict rules:

- 'survives' -- the proof establishes the claim exactly as stated, no qualification needed. 'supporting_finding_ids' may be empty. This is the ONLY verdict that may have empty support -- every other verdict below requires at least one finding.
- 'survives_with_qualification' -- **do not use this as a default bucket for discomfort.** It requires a SPECIFIC qualification that you can show is (a) genuinely absent from the paper as written (quote the passage that should have stated it and does not), (b) material -- the claim's later use elsewhere in the paper actually depends on it, not just a stylistic nicety, and (c) tied to a 'finding_id' in 'supporting_finding_ids' that names the missing piece. If you cannot supply all three, the claim is fine as written and the verdict is 'survives', even if something about it makes you uneasy. A claim that is correct but that you would have proved differently is 'survives', not 'survives_with_qualification'.
- 'unproved_as_written' -- you found a specific gap in the proof (cite the exact step, via a 'finding_id') but no counterexample; the claim might still be true, but the printed argument does not establish it. Requires at least one 'supporting_finding_ids' entry.
- 'false_as_stated' -- you have a counterexample or an independent rederivation that contradicts the printed statement. Requires at least one 'supporting_finding_ids' entry with 'evidence_kind' of 'counterexample' or 'rederivation' -- the output will be rejected otherwise. Do not use this verdict for a claim you merely suspect is false; use 'unproved_as_written' and a 'reasoning_only' finding instead.

****Boundary case carve-out.**** If your ONLY falsifying evidence is a 'boundary_case'-coded finding -- the claim fails only at a knife-edge point ($K=0$, $\sigma=1$, a tie) that the authors plausibly never intended to be read literally as included -- that is 'survives_with_qualification' (state the domain restriction that fixes it), not 'false_as_stated'. Reserve 'false_as_stated' for a failure in the INTERIOR of the stated domain, or for a boundary failure you can show the authors clearly did intend to cover (they use the boundary point elsewhere, or the boundary is load-bearing for a later result). If you have both an interior counterexample and a boundary one, cite both -- the interior one is what earns 'false_as_stated'. The output will be rejected if every falsifying finding you cite is coded 'boundary_case'.

For each entry, quote the claim as printed ('statement_quote', verbatim), give the 'verdict_basis' (why this verdict, in prose, naming the specific finding(s) it rests on), and list 'supporting_finding_ids'. Before writing a 'survives_with_qualification' or 'false_as_stated' verdict, re-read the three-part test above and check you can actually satisfy it for THIS claim -- a paper with one real defect should not produce a ledger where most claims carry a qualification.

What survives / what is out of scope

After the claim ledger, use the 'scope' field to state plainly what you did and did not check (e.g. "I did not re-derive the numerical calibration in Section 7" or "the online appendix was not provided in this package"). If a genuine blocker prevented you from completing the audit (source truncated, a section illegible, a referenced appendix missing), set 'blocked_reason' and still report whatever you completed -- do not fail silently.

Output contract

Write a single JSON object to 'result.json' in the working directory, valid against 'result.schema.json':

```

“‘json
{
  "paper_id": "<from the package>",
  "claim_ledger": [ /* one entry per numbered result, see above */ ],
  "findings": [ /* every finding, flagged or reasoning_only */ ],
  "scope": "<what you checked and did not check>",
  "blocked_reason": null
}
“‘

```

Every finding needs a ‘quote_anchor’: a verbatim substring (8-300 characters) copied exactly from the source at the error’s location -- not a paraphrase, not a LaTeX re-typesetting, the literal characters as they appear in the source file. When your finding concerns a specific claim, set ‘claim_id’ to that claim’s id in the ledger; paper-level findings (e.g. a notational inconsistency spanning the whole paper) use ‘claim_id: null’. If a finding’s consequences reach beyond the claim it is attached to, list every affected claim’s id in ‘propagation’.

When ‘result.json’ is written and schema-valid, you are done; reply with the single word DONE.

N.3 Stage 2a: addendum concatenated to audit_v1

```
# Claim-graph integration addendum (audit_graph_addendum_v1)
```

This addendum places the unchanged ‘audit_v1’ task inside a three-framework system:

1. Framework 1 independently reconstructs the paper’s authorial claim graph.
2. ****You are Framework 2.**** You independently detect and characterize possible formal errors.
3. Framework 3 later matches your findings to the frozen graph and computes downstream damage.

Independence boundary

You are not given Framework 1’s graph. Do not infer graph node or edge IDs, calculate graph centrality, assign a numerical graph-severity score, or tailor findings to what you think the graph contains. Follow every audit instruction, taxonomy rule, claim-ledger rule, and output field in ‘audit_v1’ exactly as written.

The existing categorical ‘severity’ field is your audit-stage prior under the taxonomy. It is preserved for comparison but will not be used as an input to the deterministic graph score.

Direct-effects rule for ‘propagation’

The ‘propagation’ field remains required by the existing schema, but include only claim-ledger claims that the finding ****directly**** defects or directly qualifies. Do not fill it with every claim that might be downstream. Framework 3 owns downstream propagation after graph matching. When the finding is paper-level or the direct claim cannot be identified, use an empty array.

This rule does not change the claim ledger: apply all original ‘audit_v1’ verdict and evidence requirements. It only prevents speculative downstream paths from being counted as additional directly defective claims.

Runner compatibility

Write the original ‘audit_v1’ JSON object to ‘result.json’. If the execution surface requires a structured final response, return the same JSON object there as well; otherwise reply ‘DONE’ as specified by ‘audit_v1’.

N.4 Error taxonomy referenced by audit_v1

```
# Theory error taxonomy v1 -- standalone integration copy

Every finding carries three independent ratings.

## Severity

- 'fatal': the central claim/model fails as printed and no notational reading rescues it.
- 'major': an important claim survives only after substantive correction and downstream results inherit it.
- 'minor': a local defect changes only a narrow scope, boundary, or secondary statement.
- 'cosmetic': an unambiguous typo or notation slip with no bearing on any argument.

## Fixability

- 'trivial_fix': a sign, symbol, index, or word change with no ripple effects.
- 'local_fix': a corrected line/equation leaves the surrounding argument untouched.
- 'restatement_needed': add a hypothesis, narrow a domain, or state a selection rule; proof technique survives.

- 'substantial_rework': the argument must be redone although the mechanism may survive.
- 'no_known_fix': no correction preserving the claim was found.

## Evidence kind

- 'counterexample': explicit admissible values violate a stated conclusion.
- 'rederivation': an independent derivation yields a different expression, shown explicitly.
- 'numerical_scan': a documented scan finds violations or no violations; it does not prove a universal claim.
- 'internal_contradiction': two quoted passages cannot both be true as written.
- 'reasoning_only': a specific concern not reduced further; always 'status: possible' with honest confidence.

## Codes and admission evidence

| Code | Definition and minimum evidence |
|---|---|
| 'derivation_error' | A step does not follow; show source and independently rederived expressions. |
| 'sign_index_error' | Wrong sign/subscript/index; show printed, corrected, and violated definition. |
| 'differentiation_error' | Derivative/FOC mismatch; show function, printed derivative, and correct derivative. |
| 'unjustified_step' | Name the exact nonstandard missing step and why it is not immediate. |
| 'local_to_global' | Quote the local hypothesis and global conclusion and identify the gap. |
| 'quantifier_error' | Show printed quantifier and the quantifier actually established. |
| 'circular_argument' | Give the dependency chain that returns to the claim being proved. |
| 'missing_assumption' | Quote the proof step using a hypothesis absent from the statement. |
| 'hidden_normalization' | Quote introduction of a free object and the later silent special value. |
| 'stronger_than_stated' | Compare the stated claim with the strongest result the proof supports. |
| 'domain_violation' | Give the true domain and the point where the formula is used outside it. |
| 'boundary_case' | Give the boundary/knife-edge point and what fails or becomes nonunique. |
| 'objective_inconsistency' | Quote the two inconsistent objectives and where they diverge. |
| 'dimensional_error' | Identify the incompatible units, timing, stock/flow, or scaling. |
| 'existence_uniqueness' | Give the failing parameter region or the specific missing argument. |
| 'selection_unstated' | Give the multiplicity/indifference point and absence of a selection rule. |
| 'comparative_static_scope' | Compare stated scope to proof scope; identify coordinate/joint distinction. |
| 'envelope_foc_misuse' | Name the application point and the unmet regularity condition. |
| 'statement_proof_mismatch' | Quote statement and proof passages that establish different objects. |
| 'overclaim' | Quote the formal result and stronger prose interpretation. |
| 'false_as_stated' | Requires a counterexample or rederivation, not reasoning alone. |
| 'appendix_maintext_mismatch' | Quote incompatible definitions/assumptions in main text and appendix. |
| 'reference_error' | Show the printed cross-reference and its correct target. |
| 'math_typo' | Exact local symbol error; normally 'cosmetic' unless propagation is shown. |
| 'other' | Meet the closest evidentiary bar and name the free-text type first. |
```

```
## Claim-ledger rules
```

Every numbered Proposition, Theorem, Lemma, Corollary, Remark, and load-bearing equation gets one verdict: 'survives', 'survives_with_qualification', 'unproved_as_written', or 'false_as_stated'. Every non-'survives' verdict cites at least one 'supporting_finding_ids' item. 'false_as_stated' requires 'counterexample' or 'rederivation' evidence and cannot rest solely on a 'boundary_case' finding. A missing exposition step that is standard and valid is not a finding.

N.5 Stage 2b: neutral discovery (neutral-discovery-direct-v1.4)

```
# neutral-discovery-direct-v1.4 (prompt blocks; [[...]] are runner-filled placeholders)
```

Fresh published-paper neutral-discovery iteration.

You are an independent formal-error discovery model for economic and econometric theory. Work neutrally and reproducibly. Output candidate, rejected, and unresolved hypotheses - not a final audit. Assign no severity, importance, impact, confidence, or paper-quality judgment.

This is independent iteration [[ITERATION_NUMBER]], with run ID [[RUN_ID]], under protocol neutral-discovery-direct-v1.4. The Python runner - not you - records the batch, exact prompt, source inventory and hashes, provider, model, reasoning effort, runtime events, and validation metadata. Do not recreate those administrative records.

Your scope is exactly this clean folder. Do not inspect outside it, browse, fetch external sources, install packages, read user settings, or seek prior runs. input/ is a verified runtime snapshot of the selected canonical Papers directory in the read-only clean room; treat every PDF as immutable. Write only beneath output/ and create no PDFs.

The registered source set is:
[[INPUT_LIST]]

For each source, extracted_text/ contains deterministic pdftotext -layout output for navigation. Extraction may damage equations, symbols, columns, or page order. Establish no mathematical claim from extracted text alone: check every reported formula and quote against the rendered PDF. Render relevant pages to PNG beneath output/evidence/ when needed.

Stage 1 - Before evaluating possible issues, inspect the rendered title page and record paper_identity in formal_map.json with title, authors as an array of individual names, source, location, and rendered_source_check. Use the exact registered primary-PDF path for source. rendered_source_check must contain checked=true, a safe evidence_file beneath evidence/, and a nonempty note. Transcribe these fields from the rendered primary PDF; do not infer or silently correct them. For a blinded paper, preserve a visibly redacted title as a redaction description, set authors to an empty array, add a nonempty authors_note stating that the byline is redacted, and do not reconstruct identity from residual running heads, affiliations, citations, file metadata, or other hints. Then construct the formal map. Inventory primitives and domains; timing and information; assumptions and quantifiers; definitions; equilibrium or solution concepts; numbered claims; proof dependencies; load-bearing equations; formal appendices; identification conditions; estimators; and asymptotic statements. Finish and save output/formal_map.json before beginning Stage 2.

Stage 2 - Neutral discovery. For every possible issue: precisely locate the statement; transcribe the relevant formula from the rendered source; reconstruct the strongest reasonable intended argument; identify the first step that may fail; provide an explicit derivation, failed condition, contradiction, or admissible counterexample; check boundary cases and alternative interpretations; state the weakest possible repair; and classify the record only as candidate, rejected, or unresolved.

For economic theory, test domains, signs, quantifiers, boundary cases, hidden regularity conditions, existence, uniqueness, envelope arguments, first-order conditions, fixed points, measure-theoretic steps, dynamic programs, deviations, incentive constraints, comparative statics, and welfare arguments. For econometric theory, test identification, rank and support, conditioning, moment restrictions, measurability,

uniformity, stochastic equicontinuity, limit interchanges, rates, weak identification, nuisance parameters, generated regressors, bootstrap validity, variance formulas, misspecification, boundary parameters, high-dimensional assumptions, and whether the stated asymptotics apply to the estimator actually defined.

Distinguish a genuine logical failure from terse exposition, unconventional notation, an extraction artifact, or a repairable omitted detail. State uncertainty and the strongest alternative interpretation. Do not promote a concern unless the explicit reasoning supports it.

Use exact symbolic or numerical checks when useful. If a package is unavailable, record that and use a transparent alternative without installing it. Numerical searches are diagnostic unless they exhaust the relevant finite domain. Preserve code, inputs, package versions, deterministic seeds, raw outputs, and each check's scope beneath `output/checks/`. The runner preserves the process event stream.

Write only these model-authored artifacts beneath `output/`: `formal_map.json`, `candidates.json`, `rejected_candidates.json`, `summary.txt`, `evidence/`, and `checks/`. Do not create run manifests, source inventories, prompt copies, administrative command logs, PDF output, Markdown files, or files outside `output/`. The runner handles provenance, events, diagnostics, hashes, and packaging.

`candidates.json` must be an object containing `run_id` and records. Each record must contain: `candidate_id`; `status`; `title`; `sources`; `location`; `rendered_source_check`; `claim_being_tested`; `intended_argument`; `first_potentially_invalid_step`; `reproducible_evidence`; `alternative_interpretation`; `weakest_repair`; `directly_affected_claims`; `open_checks`; `evidence_files`; and `check_files`. In `evidence_files` and `check_files` use paths relative to the output root, such as `evidence/page.png` and `checks/test.py`, without an `output/` prefix. Use only `candidate`, `rejected`, or `unresolved` as status values. Do not add severity, importance, confidence, impact, or paper-quality fields.

`formal_map.json` must be a JSON object containing `run_id`, `built_before_discovery` set to true, `paper_identity`, `primitives`, `domains`, `timing_information`, `assumptions_quantifiers`, `definitions`, `solution_concepts`, `numbered_claims`, `proof_dependencies`, `load_bearing_equations`, `identification_estimation_asymptotics`, and `formal_appendices`. `rejected_candidates.json` must be a JSON object containing `run_id` and records. `summary.txt` must be plain UTF-8 text with a neutral coverage summary and no severity or importance assessment.

Formatting contract for the runner-generated review: in every model-authored string, write each mathematical expression as LaTeX delimited by `\( ... \)` for inline mathematics or `\[ ... \]` for display mathematics. Do not use dollar-sign delimiters. In JSON files, escape backslashes as required by valid JSON so the parsed string contains those delimiters. Keep ordinary prose outside mathematical delimiters. The Python runner safely escapes prose, renders the delimited mathematics into `review.tex`, and compiles `review.pdf`.

Preserve explicit premises, derivations, assumptions, source locations, executable checks, alternative interpretations, and reasons for retaining or rejecting a hypothesis. Do not expose private chain-of-thought. A third party must be able to reproduce each calculation from visible evidence.

Inspect every supplied formal section and appendix. In `summary.txt` mark analytical coverage complete only if all local formal material has been inventoried, candidate formulas have been checked against rendered pages, claimed executable checks have run successfully, rejected hypotheses have been retained, and limitations are truthful. Otherwise mark coverage partial and state exactly what remains. Completion does not convert candidates into verified findings.

Begin by inventorying the clean folder, verifying paper identity on the rendered title page, and constructing the formal map. Do not read or rely on any ambient Markdown instruction file.

N.6 Stage 3A: linkage (link_v1)

```
# Error-graph linking -- instructions (link_v1)
```

You are Framework 3A in a three-framework system. Framework 1 produced the frozen authorial argument graph in `'graph.json'`. Framework 2 independently produced the frozen formal audit in `'audit_result.json'`. Match each audit finding to the smallest graph intervention that represents

its direct structural consequence. A deterministic program--not you--will propagate the intervention and calculate numerical severity.

Do not search for new errors, re-audit the paper, alter either input, change an audit verdict, calculate centrality, predict the eventual score, or assign percentile classes. Do not use web search or network access. You may consult 'paper.pdf' and 'paper.txt' only to resolve anchors.

No-drop contract

Return exactly one mapping for every 'finding_id' in 'audit_result.json', including possible, cosmetic, and paper-level findings. Preserve each ID exactly and never invent a finding ID.

Match statuses

- 'matched': a unique minimal graph intervention can be identified;
- 'no_structural_effect': the finding is real or possible but no graph support relation changes, as with an unambiguous typo whose surrounding argument already uses the intended expression;
- 'ambiguous': two or more materially different graph matches remain plausible;
- 'unmatched': the graph lacks the relevant object or the finding cannot be located.

Use 'candidate_object_ids' to record plausible IDs for ambiguous cases. Unmatched and ambiguous findings are not automatically low severity; they remain unscored pending review.

Intervention operators

Choose the smallest nonredundant set of targets:

- 'delete_edge': the finding invalidates a particular presented derivation or interpretive link;
- 'delete_node': the node's own statement is false as stated or cannot serve as a premise;
- 'qualify_edge': the derivation works only under a narrower condition or domain;
- 'qualify_node': the node's statement survives only with a narrower condition or domain.

Typical mappings:

- invalid proof step, conclusion not shown false -> delete the immediate inferential edge;
- direct counterexample to a proposition -> delete that proposition node;
- missing restriction or boundary carve-out -> qualify the node or edge;
- overclaim from valid formal result -> delete or qualify only the formal-to-interpretive edge;
- local typo with no change in the argument -> no structural effect.

Do not delete a node and all of its outgoing edges: deleting the node already propagates through those edges. Do not attach the same intervention to downstream nodes merely because they may be affected; the scorer computes those consequences.

Consequence kind

Keep epistemic consequences distinct:

- 'false': the audit supplies a counterexample/rederivation, or an internal contradiction that directly proves the mapped graph object impossible;
- 'unsupported': a presented route fails but the target has not been shown false;
- 'qualified': the claim/route survives only on a narrower scope;
- 'reversed': the audit supports a conclusion with the opposite sign/direction;
- 'no_structural_effect': no support relation changes;
- 'uncertain': the candidate finding or mapping is not strong enough for another label.

Do not upgrade the audit's evidence. In particular, 'reasoning_only' cannot become 'false' here.

Scope loss

For a qualification, copy the exact lost parameter/domain region into 'scope_region_lost'. Set 'scope_loss_fraction' only when the paper, audit, or a clearly bounded stated domain makes the

fraction defensible. It is the fraction of the claim's stated domain on which support is lost, between 0 and 1. Otherwise use 'null'; the scorer will return a severity interval rather than inventing precision.

Direct claim IDs

'directly_affected_claim_ids' uses Framework 2 claim-ledger IDs, not graph IDs. Include only claims directly defected/qualified by the finding. Do not add downstream claim-ledger entries.

Output

Write one JSON object to 'result.json', valid against 'linkage.schema.json'. Return the same JSON object as the final response if structured output is required. Do not return Markdown or text outside the object. Every mapping needs a concise rationale grounded in the finding and graph anchors, plus 'mapping_confidence' from 0 to 100.

N.7 Stage 3B: weakest valid repair (repair_v1)

Weakest-valid-repair analysis -- instructions (repair_v1)

You are Framework 3B in a frozen theory-error workflow. Framework 1 produced the authorial argument graph in 'graph.json'; Framework 2 produced the independent error ledger in 'audit_result.json'; Framework 3A produced the error-to-graph mapping in 'linkage.json'. A deterministic program produced 'repair_targets.json', which lists the exact headline and major outputs that must be reassessed for every finding.

Your job is to construct the weakest valid mathematical repair for each mapped finding and then rederive every required output. Do not search for new errors, modify any frozen input, change a finding or mapping, calculate a numerical severity score, or use network access.

No-drop and coverage contracts

Return exactly one analysis for every 'finding_id' in 'linkage.json'. Preserve every ID and do not invent IDs. For each finding, 'output_assessments' must contain exactly the output IDs listed for that finding in 'repair_targets.json': no omission, duplication, or substitution. The graph fixes which claims must be checked; you determine what happens to their qualitative statements after repair.

Weakest valid repair

Use the smallest correction consistent with the paper's stated primitives, objective, equilibrium concept, and mathematical identities. Prefer, in order:

1. correcting a local formula;
2. rewriting the affected proof while preserving the model;
3. adding or changing an assumption only when independently justified by the paper;
4. changing the model only when no weaker repair is valid.

Do not add an assumption merely to recover the authors' preferred conclusion. Do not treat deleting the faulty object as the repair. State the corrected equation, condition, proof step, or assumption precisely enough for another theorist to reproduce the rederivation.

Use 'repair_burden' as follows:

- 'none': no graph-relevant repair is needed;
- 'local_formula': a local equation/index/sign correction suffices;
- 'proof_rewrite': the primitives survive but the proof or derivation must be materially redone;
- 'assumption_change': a justified restriction or assumption must change;
- 'model_change': primitives, objective, equilibrium concept, or mechanism must change;

```

- 'unresolved': no defensible weakest repair or downstream rederivation was completed;
- 'not_applicable': the mapping is ambiguous or unmatched.

## Analysis status

- 'resolved': a weakest valid repair and all required downstream outputs are resolved;
- 'partially_resolved': a repair is identified but at least one required output remains unresolved;
- 'unresolved': no defensible repair or downstream conclusion can be established;
- 'not_needed': the linker classified the finding as 'no_structural_effect';
- 'not_assessable': the linker classified the finding as 'ambiguous' or 'unmatched'.

For 'not_needed' use burden 'none', no repair text, and no output assessments. For
'not_assessable' use burden 'not_applicable', no repair text, and no output assessments.

## Repaired output status

Judge the exact qualitative statement stored for each required graph output:

- 'survives_unchanged': the mathematical and qualitative content survives;
- 'survives_quantitatively_changed': magnitudes, thresholds, or formulas change, but the stated
  sign, direction, existence, ordering, mechanism, or other qualitative conclusion survives;
- 'survives_with_scope_qualification': the qualitative statement survives only on a narrower
  parameter or state domain;
- 'qualitatively_reversed': its sign, direction, ranking, or mechanism reverses;
- 'false_after_repair': the stated qualitative conclusion fails after the valid repair;
- 'unresolved_after_repair': the repaired analysis does not determine the claim.

'affected_scope_fraction' is the fraction of the claim's stated domain whose qualitative content is
changed or lost. Use 0 for 'survives_unchanged'. Use a number in [0,1] only when the paper supplies
a defensible bounded domain, calibration, prior, or exhaustive finite state set. Otherwise use
'null' and describe the exact unknown or lost region. For a global reversal or false claim, use 1.
Quantitative changes can use 1 when they occur across the entire domain even though the qualitative
statement survives.

## Epistemic discipline

Separate "the printed proof no longer establishes this" from "the repaired model makes this
false." Trace the corrected object through the graph and paper rather than inferring outcomes from
network position. If a derivation cannot be completed, use an unresolved status and explain the
missing step. Set 'adjudication_status' to 'model_provisional'; only a human reviewer may change it
to 'human_adjudicated'.

## Output

Write one JSON object to 'result.json', valid against 'repair.schema.json'. Do not return Markdown
or text outside the object. Each analysis needs a concise but reproducible overall explanation,
and each output assessment needs a derivation summary and confidence from 0 to 100.

```

N.8 Validation: comparison against the corrigendum

```

# Known-error comparison

You are evaluating a blind paper audit against a later published correction.
This is an evaluation stage: the correction is now visible.

Read 'correction.pdf' (authoritative), 'correction.txt', 'audit_run_1.json',
'audit_run_2.json', 'graph.json', 'case.json', and 'comparison_schema.json'.

Write one complete 'result.json' matching the schema. Do not modify any input.

```

Rules:

1. Extract every distinct substantive error documented by the correction. Do not treat stylistic edits as separate errors.
2. Match each documented error independently against each blind audit run.
 - 'exact': the audit identifies the same target and central failure mechanism.
 - 'partial': the audit identifies the same target but misses or materially misstates part of the mechanism.
 - 'miss': neither exact nor partial.
3. Set aggregate 'detected=true' if at least one run is exact or partial, and set aggregate 'match_quality' to the best run result (exact > partial > miss).
4. Use one error type from: 'false_claim', 'missing_assumption', 'proof_gap', 'algebra_error', 'sign_direction_error', 'domain_error', 'definition_reference_error', 'other'. Code the audit's type from the actual mechanism it describes, not merely its title.
5. Map the documented error to the smallest relevant graph nodes/edges. Set 'graph_claim_relevant' true only when the error can affect a formal support route to a headline claim. Set it false for purely local presentation errors.
6. 'graph_severity' is the fraction of headline claims reachable downstream from the affected graph objects, between 0 and 1. Use null if the error is absent from the graph or reachability is genuinely indeterminate.
7. Inventory every candidate from both audit runs in 'audit_findings'. Use IDs 'run1:<candidate_id>' and 'run2:<candidate_id>'. An unmatched finding is 'additional_unadjudicated', not a false positive, because the correction is not an exhaustive audit of the original paper.
8. Leave 'negative_controls' empty. This corpus contains known errors but no verified clean controls.
9. Be concise and evidence-based. Include correction page anchors and graph IDs.

After writing 'result.json', reply only 'DONE'.

N.9 Validation: finding-to-graph mapping

Map every audit finding to the claim graph

Read 'comparison.json', 'audit_run_1.json', 'audit_run_2.json', 'graph.json', and 'mapping_schema.json'. Write 'result.json' matching the schema.

For every record in 'comparison.json.audit_findings':

1. Locate the corresponding candidate in the named audit run.
2. Map it to the smallest graph node IDs and/or edge IDs that represent the challenged claim or support step.
3. Set 'graph_claim_relevant=true' only if failure of those objects can affect a formal support route to at least one headline claim. Otherwise set it false.
4. For a local presentation issue or a finding absent from the graph, use an empty 'graph_object_ids' list and set relevance false.
5. Do not judge whether the finding is correct and do not use the corrigendum. This task is structural mapping only.
6. Return every finding exactly once, preserving its 'finding_id' verbatim.

After writing 'result.json', reply only 'DONE'.