# The Effect of AI Writing on Governance: Evidence from a $3.5 Billion DAO

Victor Huang* Joseph Hall*

*Georgia Institute of Technology

August 18, 2026

**Abstract**

We study the extent to which online deliberation aggregates information before collective votes. Our setting is Arbitrum DAO, a decentralized organization controlling a $3.5 billion treasury, where token holders debate proposals on a public forum before each binding vote. Forum activity predicted vote closeness before March 2024 but became uninformative after, coinciding with language models capable of producing indistinguishable governance prose. Interestingly, we observe that the degradation of the forum's signal (the elimination of the correlation between posting volume and margin-of-victory) takes place at a time when relatively few posts are A.I. generated, relative to the end of the sample period where A.I. posting volume grows rapidly. A model of endogenous posting, reading, and voting—derived from primitives—shows this is a general phenomenon: the signal value of a forum collapses at a discrete tipping point at a realized A.I. share strictly below the level that would mechanically destroy reading value, and the collapse is hysteretic: reducing A.I. volume afterward does not restore the forum. Calibrating to the data, the forum generated net social benefits equivalent to saving at least 858 words of research effort per delegate per proposal (under a stated non-redundancy assumption)—all destroyed at the tipping point. Delegates who rely on AI hold far less voting power than those who do not, which is inconsistent with power capture by insiders. Our results suggest that the mere availability of indistinguishable AI writing can degrade the governance capabilities of economically important organizations, even when such AI is rarely used in equilibrium.

*JEL*: D72, D82, D83, G34, O33 *Keywords*: decentralized autonomous organizations, DAO governance, artificial intelligence, information aggregation, structural break, blockchain governance

# 1 Introduction

Text-based deliberation is a foundation of financial governance. The Securities and Exchange Commission solicits public comment before finalizing rules and treats letter volume and content as a gauge of genuine stakeholder concern. Shareholders read proxy statements, activist letters, and analyst reports before contested votes on mergers, compensation, and board composition (Iliev and Lowry, 2015). Retail investors post on Reddit and StockTwits, and a large empirical literature has established that this crowd activity carries real information about market outcomes (Antweiler and Frank, 2004; Chen et al., 2014). The informational value of each of these mechanisms rests on a single implicit assumption: that producing credible, contextually appropriate long-form prose requires genuine expertise or conviction—a cost sufficient to screen out noise and make text a credible signal of authentic opinion.

Artificial intelligence is destroying that assumption. As the marginal cost of generating sophisticated, indistinguishable prose approaches zero, the link between text production and genuine information is severed—and with it, the credibility of every text-based deliberation channel in finance. This paper asks how fast the collapse happens, what triggers it, and why. Our laboratory is the Arbitrum DAO governance forum: a real-money governance setting with $3.5 billion in treasury assets, with every post and vote recorded immutably on blockchain, and—because crypto communities are technically sophisticated and move faster than traditional capital markets—the first financial governance context to confront the AI quality threshold at scale—and therefore the first setting in which the mechanism we study can be cleanly identified. What we find here is likely to recur in corporate earnings calls, analyst forums, and any other text-based financial deliberation as AI capability continues to advance.

We study this question using a setting that provides what corporate governance research rarely enjoys: an exogenous shock to the quality of the signal. On March 4, 2024, Anthropic released Claude 3, a large language model capable of producing sophisticated, contextually

appropriate long-form prose. Within weeks, the Arbitrum governance forum—like most online platforms—experienced a wave of AI-generated posts. These posts are grammatically fluent, topically relevant, and superficially indistinguishable from human-authored contributions. They are not, however, opinion-bearing: they do not reflect a genuine preference for or against the proposal in question, and they do not aggregate private information held by stakeholders. They carry no directional vote signal: they do not predict whether a proposal passes or fails.

We measure AI-generated content using the Pangram API, a commercial AI-detection service that returns, for each post, the estimated fraction of content generated by large language models (*fraction_ai*). We classify posts with *fraction_ai* $> 0.70$ as AI-written and posts below this threshold as human-written. This classification allows us to separately track the human signal and the AI noise within each proposal thread across the full sample period.

Our main finding is a structural break. In the pre–Claude 3 period (January 2023 through February 2024), the number of human-authored posts in a proposal's forum thread predicts the proposal's margin of victory—the absolute difference in vote share between the winning and losing sides among contested (For vs. Against) votes. The Pearson correlation is $r = -0.33$ ($p = 0.08$, $n = 29$ proposals): more posts, more disagreement, smaller margin. This is consistent with a forum that aggregates genuine signals of community sentiment. Delegates and token holders who observe an unusually active thread are, in effect, observing a sufficient statistic for the degree of community division.

After Claude 3, this relationship vanishes. The correlation falls to $r = -0.10$ ($p = 0.39$, $n = 73$) in the post-shock period. The change in the correlation coefficient, $\Delta r = +0.23$, is confirmed by a Chow test that rejects parameter stability at $F = 8.38$ ($p = 0.0004$). We use Claude 3 (March 2024) as our primary break date, which provides a clean before-and-after contrast. All substantive conclusions hold at any candidate break date from November 2023 onward (Table 3).

The identification logic rests on the exogeneity of AI capability releases. AI model releases are global events determined by the research roadmaps of Anthropic, OpenAI, and other technology companies—they are plainly exogenous to any specific Arbitrum proposal. The release of Claude 3 was not anticipated by the Arbitrum community, was not timed to coincide with any particular governance vote, and affected all future proposal threads uniformly. The shock therefore provides a plausible source of quasi-exogenous variation in the posting environment. We interpret the before–after comparison as the best available quasi-experimental evidence consistent with an AI-contamination mechanism. We acknowledge that the November 2023–March 2024 window also encompasses several concurrent governance developments—LTIPP/STIP grant batching, the approach of the March 2024 token vesting cliff, and expanding vote-rental participation—that could in principle affect the posts–margin relationship. Section 4 and the robustness checks in Section 8 evaluate each of these alternatives directly. Cross-DAO evidence (Figure A7) provides additional corroboration: five major DeFi DAOs without Arbitrum-specific governance dynamics exhibit comparable participation declines over the same window, inconsistent with Arbitrum-specific confounders as the primary driver.

We then turn to the mechanism. The most natural candidate explanation—that a physical flood of AI-generated posts numerically overwhelmed the human signal—is falsified by timing. To be precise: AI-generated posts are detectable on the forum well before the structural break—a small but measurable share appear as early as early 2023—but at negligible volume. At the date of the QLR break (November 2023), AI posts constituted approximately 8% of forum volume, and approximately 6% at the Claude 3 break (March 2024)—remaining in the single digits throughout, and dipping below 3% in between. The large-scale volume surge did not arrive until late 2024, reaching 15–20% of forum traffic nine to twelve months after the signal collapsed. A volume mechanism requires contamination to precede collapse; the data show the reverse.

We propose instead an *unverifiability shock* mechanism in the spirit of Akerlof (1970).

The releases of GPT-4 Turbo (November 2023) and Claude 3 (March 2024) crossed a quality threshold at which AI-generated governance prose became indistinguishable from careful human writing.

There is direct benchmark evidence that these two releases were qualitatively different from their predecessors. On the Massive Multitask Language Understanding benchmark (MMLU; Hendrycks et al., 2021), which tests breadth of knowledge across 57 academic and professional domains, GPT-4 (March 2023) scored 86.4% versus 70.0% for GPT-3.5—a step change that closed most of the remaining gap to expert human performance (estimated at 89.8%). GPT-4 Turbo extended this to 87.3%, and Claude 3 Opus reached 86.8% (Anthropic, 2024). On MT-Bench, a benchmark designed to evaluate multi-turn instruction-following and writing quality, GPT-4 achieved a score of 8.99 out of 10 versus 7.94 for GPT-3.5 (Zheng et al., 2023).

Two further lines of evidence, neither reliant on vendor-reported benchmarks, corroborate that this generation crossed a human-indistinguishability threshold. First, blind human preference: on the LMSYS Chatbot Arena, where users vote on anonymized head-to-head outputs (Zheng et al., 2023), Claude 3 Opus (released March 4, 2024) became the first non-OpenAI model to overtake the GPT-4 family at the top of the leaderboard, which GPT-4 variants had led since May 2023; earlier models such as Claude 2 never reached that tier. Second, controlled Turing tests: in a randomized, preregistered two-party test, Jones and Bergen (2025b) found GPT-4 judged to be human 54% of the time—versus 66–67% for real humans and 22% for a 1960s ELIZA baseline—up from 49.7% for the best GPT-4 prompt in an earlier online study (Jones and Bergen, 2024) and near-chance for prior models, and a subsequent three-party test reached 73% for GPT-4.5, above the human baseline (Jones and Bergen, 2025a). These are ordinal claims established by blind human judgment rather than self-reported capability scores, and they place the crossing of near-human indistinguishability in the November 2023–March 2024 window (Figure A1). We acknowledge two caveats: no benchmark measures indistinguishability of *governance-forum* prose specifically, so the Arena

and Turing tests are the nearest available instruments; and the highest Turing pass rates require prompting the model to adopt a humanlike persona.

Once this threshold was crossed, governance delegates—who read forum posts as signals of community sentiment—could no longer verify the authenticity of any given post. Under standard informational logic (Grossman and Stiglitz, 1980), the expected value of an unverifiable signal falls toward its prior: once the expected probability of AI contamination exceeds a threshold, delegates rationally discount the forum even though realized contamination remains low. The credible *threat* of AI content is sufficient to destroy the forum's information value—even before any flood of AI posts arrives.

Delegate-level evidence is consistent with this mechanism and points away from a power-capture story. Using a dataset linking forum usernames to Snapshot voting records for 62 delegates, we document a robust negative cross-sectional relationship between a delegate's average AI content fraction and their average delegated voting power: $r = -0.44$ ($p = 0.001$, $n = 51$ delegates). High-AI delegates hold on average 71,000 ARB ($92,000) in delegated power; low-AI delegates average 1.9 million ARB ($2.5 million)—a gap of 27×. AI tools are adopted primarily by lower-tier delegates seeking to maintain forum presence at lower cost, not by high-power insiders seeking to manufacture consensus. The signal degrades not because the powerful are gaming the forum, but because the *possibility* of doing so at zero cost became credible.

We also examine whether the distraction hypothesis of Hirshleifer et al. (2009)—in which exogenous events that command attention reduce participation in other activities—can account for variation in posting volume or vote outcomes. We construct a comprehensive battery of candidate distraction instruments: major token airdrops (ZKsync, LayerZero, Starknet), cryptocurrency market crashes (Bank of Japan rate shock, August 2024; SVB collapse, March 2023), exchange hacks and security incidents, regulatory events (SEC enforcement actions, Bitcoin ETF approval), and concurrent DAO proposal load. Every instrument with a statistically significant first stage has the *wrong sign*: exogenous events

increase, rather than decrease, forum activity. This finding is itself informative about the nature of DAO participation. Unlike the retail investors in Hirshleifer et al. (2009), Arbitrum DAO participants are core stakeholders who treat DeFi market events as *complements* to governance engagement, not substitutes. This indicates that DAO governance involves a fundamentally different mode of participation than the retail trading studied in the attention literature: core stakeholders respond to market events by *engaging more*, not tuning out.

**Contributions.** This paper makes four contributions.

*First*, we provide the first empirical evidence that online governance forums aggregate genuine information about vote outcomes. The pre-shock posts–margin relationship documents that forum activity is not epiphenomenal: it tracks real community division in a way that has predictive content for on-chain vote outcomes. This connects the DAO literature (Yermack, 2017; Kiayias and Lazos, 2022; Barbereau et al., 2023) to the social-media-and-markets literature (Antweiler and Frank, 2004; Chen et al., 2014; Cookson and Niessner, 2020), extending both to a setting where the forum is the public, recorded deliberation channel between information and vote.

*Second*, we provide the first quasi-experimental evidence on the effect of AI-generated content on collective decision mechanisms. The AI shock literature (Lopez-Lira and Tang, 2023; Cao et al., 2024; Brynjolfsson et al., 2023) has focused on effects on labor markets and financial analysis; our contribution is to show that generative AI degrades a specific form of information aggregation—democratic deliberation—through a channel not studied in prior work: the collapse of participation costs reduces the cost of mimicking credible engagement.

*Third*, we document the internal structure of the DAO delegate market. The delegate system in Arbitrum and other major DAOs is a form of liquid democracy (Blum and Zuber, 2016; Green-Armytage, 2015): token holders delegate their voting power to representatives who specialize in governance participation. We provide the first systematic characterization of the relationship between AI adoption and delegate standing, with implications for the

theory of delegation in decentralized governance.

*Fourth*, we provide a comprehensive description of the Arbitrum DAO as an economic institution—its governance mechanics, treasury structure, proposal lifecycle, participant composition, and the vote rental market that has emerged around it—that we hope will serve as a reference for future research on DAO governance.

**Related literature.** This paper connects several strands of research.

*Information aggregation through costly action.* The theoretical foundation for our forum-signal hypothesis is the class of models in which agents pay participation costs to transmit private information to a collective decision-maker. Lohmann (1994) shows that costly political action—rallies, petitions, demonstrations—aggregates dispersed information about policy payoffs, with participation equilibria predicated on the relative cost of action; the companion paper Lohmann (1993) extends the framework to allow strategic noise injection by motivated actors, anticipating the adverse-selection mechanism we study. Dewatripont and Tirole (1999) model advocates who pay investigation costs to present arguments to a principal; the adversarial structure ensures information is credibly transmitted, but at a cost that limits participation. Austen-Smith (1990) provides the political-science antecedent to our main empirical prediction: legislative debate intensity is highest precisely when the chamber is most evenly divided. In the voting literature, Feddersen and Pesendorfer (1996) and Feddersen and Pesendorfer (1997) show that endogenous abstention aggregates private information through the selection of who participates—the same logic our model applies to forum posting. On the receiver side, the delegate's reading threshold mirrors the information-acquisition condition of Grossman and Stiglitz (1980): an agent acquires information only when the expected return exceeds cost; Persico (2004) and Gerardi and Yariv (2008) develop related models in which committee members pay to acquire signals before voting. Our setting features the same costly-participation structure: posting on a public forum is a costly signal of genuine engagement, and the forum aggregates these signals into a collective indicator

of community division. The AI shock is, in the language of these models, a reduction in the cost of mimicking a high-quality signal—an event with no analog in the prior literature, which opens the empirical question this paper addresses.

*Persuasion, advocacy, and hard information.* Our forum is, in effect, a decentralized advocacy mechanism. Dewatripont and Tirole (1999) show that two opposed advocates can dominate a single neutral investigator, because partisan mandates restore the high-powered incentives a "find both sides" mission dilutes; the no-AI forum equilibrium—both sides posting in proportion to their support, the delegate aggregating—reproduces this structure without any designer imposing it. The decisive difference is the discipline device. In the persuasion and disclosure literature (Milgrom and Roberts, 1986; Shin, 1998), credibility rests on *hard* information: interested parties can suppress evidence but cannot fabricate it, so competition among them reveals the truth. Forum posts are *soft*—their credibility never came from verifiability but entirely from being costly to produce. Our paper can thus be read as asking what becomes of an advocacy mechanism whose discipline device is production cost rather than hard evidence once that cost falls to zero: it unravels, precisely because there is no verifiability to fall back on. Relative to this literature, and to models of strategic information provision (Che and Kartik, 2009; Kamenica and Gentzkow, 2011), our contribution is on the *demand* side of persuasion—the principal is not a passive processor of whatever advocates submit but pays $c_r$ per post and rationally chooses whether to listen at all, ceasing entirely past $\pi^*$. Seen this way, the Delegate Incentive Program is the degenerate limit of an advocacy reward scheme: it pays for submitted text irrespective of content, so once text is free to produce, what it elicits carries no information (Section 8.6).

*Online deliberation and democratic governance.* A growing literature studies whether structured deliberation improves collective decisions, from in-person deliberative polls (Luskin et al., 2002) to online forums (Wright, 2012). In political science, the question of whether public discourse aggregates preferences or merely amplifies noise has been central since Habermas (1984). Our DAO setting provides a cleaner test than most: the forum is the

primary formal pre-vote deliberation channel, the outcome is a binding quantitative vote, and both forum and vote data are complete. We find the first direct evidence that a real-money governance forum aggregates information, and that this aggregation can be destroyed by a change in the verifiable authenticity of participation.

*Text-based financial signaling.* Textual analysis in finance has documented that the *content* of corporate communications carries information: Li (2008) shows that more readable annual reports forecast better earnings; Loughran and McDonald (2011) construct a finance-specific word list that predicts stock market reactions; Cookson and Niessner (2020) find that disagreement among social-media investors predicts subsequent trading volume. These papers treat authorship as unambiguous. Our paper is, to our knowledge, the first to study what happens when the link between authorship and authentic opinion is broken by AI, and to show that this break destroys the informational content of text-based signals even when individual posts remain superficially unchanged.

*AI and financial markets.* Recent work documents that AI affects labor market outcomes (Brynjolfsson et al., 2023; Eisfeldt et al., 2023), financial analysis quality (Cao et al., 2024; Kim et al., 2024), and stock-return predictability from news (Lopez-Lira and Tang, 2023). These papers study AI as a productivity tool for human agents. Our contribution is distinct: we study AI as an adversarial technology that degrades information channels by collapsing the cost of producing credible-seeming but uninformative content.

**Road map.** Section 2 describes the Arbitrum DAO, its governance architecture, and the economic environment in which proposals are evaluated. Section 3 presents the data sources and construction of the key variables. Section 4 lays out the empirical strategy and identification argument. Section 5 presents a formal model of governance deliberation and the AI shock. Section 6 presents the main results. Section 7 analyzes the delegate mechanism. Section 8 presents robustness checks. Section 9 concludes.

# 2 Institutional Background

## 2.1 Decentralized Autonomous Organizations

A decentralized autonomous organization (DAO) is an entity whose governance rules are encoded as smart contracts on a blockchain and executed automatically without the discretion of any central authority. The defining features are: (i) membership is determined by ownership of a governance token, typically a fungible ERC-20 asset (a standardized transferable digital token on the Ethereum blockchain) tradeable on secondary markets; (ii) decisions are made by on-chain votes in which voting power is proportional to token holdings; and (iii) the outcomes of votes are implemented deterministically by the protocol, not by human managers who retain the discretion to deviate. DAOs differ from traditional firms in that there is no board of directors, no CEO, and no legal person behind the organization—only code and the community of token holders who govern it.[1]

By mid-2024, the aggregate treasury assets controlled by the ten largest DAOs exceeded $8 billion, with Arbitrum, Uniswap, Optimism, Aave, and MakerDAO collectively accounting for roughly 80% of this figure. These are not notional numbers: DAO treasuries consist primarily of the protocol's own governance token plus stablecoins and other liquid assets, held in smart contracts that can be disbursed only by a governance vote.

The primary use of these treasuries is to fund ecosystem development. A blockchain network is only as valuable as the applications built on top of it—decentralized exchanges, lending markets, games, infrastructure tools—and attracting developers requires subsidies, grants, and liquidity incentives. Without treasury-funded grants, rational developers would build on whichever platform offers the best existing user base, creating a winner-take-all dynamic. The treasury breaks this equilibrium by paying developers to build early, subsidizing liquidity to bootstrap usage, and compensating the governance participants who coordinate these decisions. This makes treasury allocation the central economic activity of a DAO: it

[1] On the legal limits of the smart contracts that underlie DAOs, see Frankenreiter (2019); for the economics of decentralized finance, see Harvey et al. (2021) and Jensen et al. (2021).

is the mechanism by which a decentralized network competes for talent and capital against centrally-managed platforms. The governance of these treasuries—who proposes, who votes, and on what—is the subject of this paper.

## 2.2 The Arbitrum DAO

Arbitrum is a layer-2 scaling solution for Ethereum—meaning it processes transactions on a secondary network rather than directly on the Ethereum blockchain, reducing fees and congestion while inheriting Ethereum's security guarantees—developed by Offchain Labs and launched in August 2021. It achieves high throughput and low transaction costs by processing transactions off the main Ethereum chain and periodically posting compressed proofs of correctness to Ethereum mainnet. By 2024, Arbitrum One—its flagship network—had processed over one billion transactions and hosted over 40% of Ethereum-based DeFi activity by total value locked (TVL).[2]

**Token distribution.** On March 16, 2023 (the "Token Generation Event," or TGE), the Arbitrum Foundation distributed 10 billion ARB governance tokens according to the following schedule: (i) 11.62% to eligible Ethereum users via a public airdrop; (ii) 1.13% to eligible DAOs building on Arbitrum; (iii) 42.02% to Offchain Labs team members and advisors, subject to a four-year vesting schedule with a one-year cliff; (iv) 17.53% to early investors, subject to the same vesting schedule; and (v) 27.70% to the Arbitrum DAO treasury, available for governance allocation. At launch, the ARB token traded at $1.30, implying a fully diluted market capitalization of $13 billion and a DAO treasury of $3.5 billion.[3]

The vesting cliff in March 2024 unlocked 1.1 billion previously locked team and investor

[2]Data from DeFiLlama (defillama.com), accessed June 2025.

[3]We convert all ARB-denominated quantities to U.S. dollars at this single token-generation reference rate ($1.30/ARB, March 2023) throughout the paper, and report the treasury as $3.5 billion accordingly. ARB's market price has since varied by more than an order of magnitude: at the price prevailing in August 2026 (≈$0.07/ARB) the same 2.77-billion-ARB treasury would be worth roughly $0.2 billion. The dollar figures therefore depend heavily on the valuation date; we fix one purely for comparability. The convention is descriptive only and affects no result—every ratio (such as the delegate voting-power gaps) and every regression in the paper is invariant to the ARB/USD rate.

tokens simultaneously. We control for this event in our empirical specifications.

Figure 1 characterizes the joint evolution of forum and market activity over the sample. Three features stand out. *Sign*: forum post volume and ARB price move together—both rise through the 2023–2024 bull market and decline in the correction that follows. The co-movement is positive throughout the sample ($r = 0.54$, $p < 0.001$), consistent with governance engagement tracking economic stakes: when token prices are high, the treasury is large and proposals command more attention. *Magnitude*: forum volume ranges from a low of roughly 200 posts per month in the early sample to a peak of over 1,500 in mid-2024, a seven-fold range. *Timing*: the post-count series leads the price series at several turning points, particularly around major grant program votes in late 2023 and the ecosystem incentive cluster in early 2024. This leading pattern motivates the question of whether forum activity aggregates forward-looking information—which our main analysis answers in the affirmative for the pre-AI period.

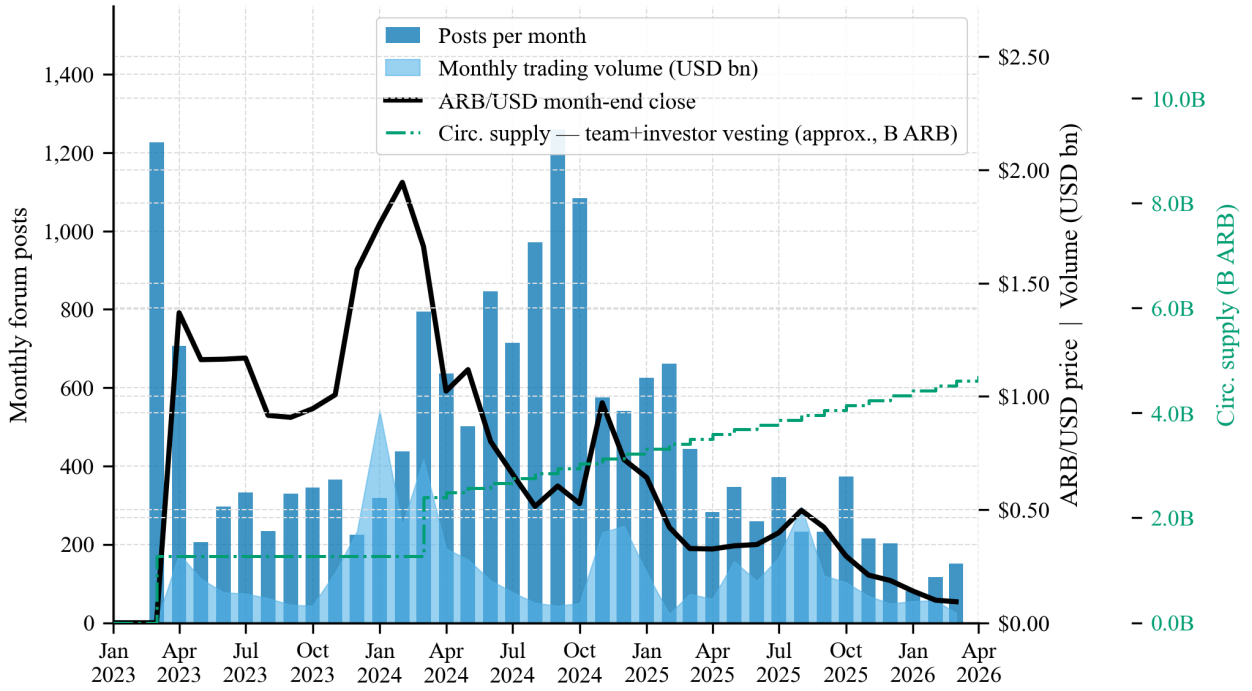


**Figure 1: Forum Participation and Token-Market Co-movement.** Monthly Arbitrum governance forum post count (blue bars, left axis), ARB/USD month-end closing price (dark line, right axis), and monthly ARB trading volume in USD billions (pale blue shaded area, right axis). The step-function overlaid in green tracks the approximate circulating supply of ARB attributable to the team and investor vesting schedule. The vesting cliff in March 2024 is visible as a discrete jump in the supply series. January 2023–March 2026.

**Governance architecture.** Arbitrum operates under a "Constitution" (a binding governance document) that distinguishes between two categories of governance action: *constitutional* proposals, which alter the Constitution or core protocol parameters and require a supermajority vote (two-thirds of quorum), and *non-constitutional* proposals, which include treasury allocation, grants, and operational decisions and require a simple majority. Both types follow the same lifecycle.

**The proposal lifecycle.** Governance decisions follow a three-stage process. In the first stage, a proposer posts a proposal to the Arbitrum governance forum (forum.arbitrum.foundation), a public Discourse platform. The forum post initiates a public discussion period during which community members, delegates, and external observers may comment, ask questions, and

express support or opposition. This discussion period is informal—it produces no binding output—but it is the primary public mechanism through which information about community preferences is recorded and aggregated before the vote.

In the second stage, the proposer submits a "temperature check" to Snapshot, an off-chain voting platform that records token-weighted votes without requiring on-chain gas expenditure. Temperature checks are non-binding straw polls that provide a low-cost signal of likely vote outcomes. Proposals with clear temperature-check failures are typically withdrawn before proceeding.

In the third stage, passed temperature checks proceed to a binding on-chain vote via Tally, which records votes on the Arbitrum blockchain. A minimum quorum of 5% of circulating supply must participate for the vote to be valid, and the proposal must receive a majority (or supermajority, for constitutional proposals) of votes cast. Passed proposals are implemented automatically by the DAO's smart contract governor after a mandatory timelock delay (typically 14 days for non-constitutional and 21 days for constitutional proposals).

For our empirical analysis, we focus on Snapshot votes, which provide the best combination of coverage (both binding and informational votes are recorded), granularity (individual vote choices and weights are observable), and historical completeness.

## 2.3 The Delegate System: Liquid Democracy at Scale

Token-weighted voting in its pure form suffers from voter apathy: holding governance tokens is economically valuable regardless of whether one participates in governance, so rational token holders with small stakes have little incentive to pay the informational cost of evaluating each proposal. Arbitrum addresses this through a delegation system: any token holder may delegate their voting power to a named *delegate*—a participant who has publicly committed to active governance participation—without transferring economic ownership of the underlying tokens. Delegation is revocable at any time, creating competitive pressure on delegates to maintain participation quality.

This mechanism is a practical instantiation of *liquid democracy* (Blum and Zuber, 2016), a governance model in which each voter chooses between voting directly and delegating to a trusted representative, who may in turn delegate further. Liquid democracy has attractive theoretical properties—it pools the informational advantages of expertise-based representation with the accountability of direct democracy—but has received little empirical study in real-world settings. The Arbitrum delegate market provides a rare opportunity to observe liquid democracy at scale, with observable delegate behavior (forum posts, vote records) and quantifiable delegation outcomes (voting power).

Delegates with substantial delegated voting power are economically analogous to proxy advisors in corporate governance (Iliev and Lowry, 2015), but with a crucial difference: Arbitrum delegates operate in public, post their reasoning in the forum, and accumulate or lose delegated power based on observed behavior. Their forum posts are a key input to our analysis.

By the end of our sample period (March 2026), the top 10 delegates by voting power collectively controlled approximately 30% of all votes cast in contested proposals, with the largest single delegate holding over 1% of total circulating ARB supply in delegated power. The distribution of voting power is highly concentrated. The *Nakamoto coefficient*—the minimum number of independent actors required to collude and control a majority of votes—stands as low as 4 delegates at the 51% threshold and 3 at the 33% blocking-minority threshold in early windows (Figure 2).

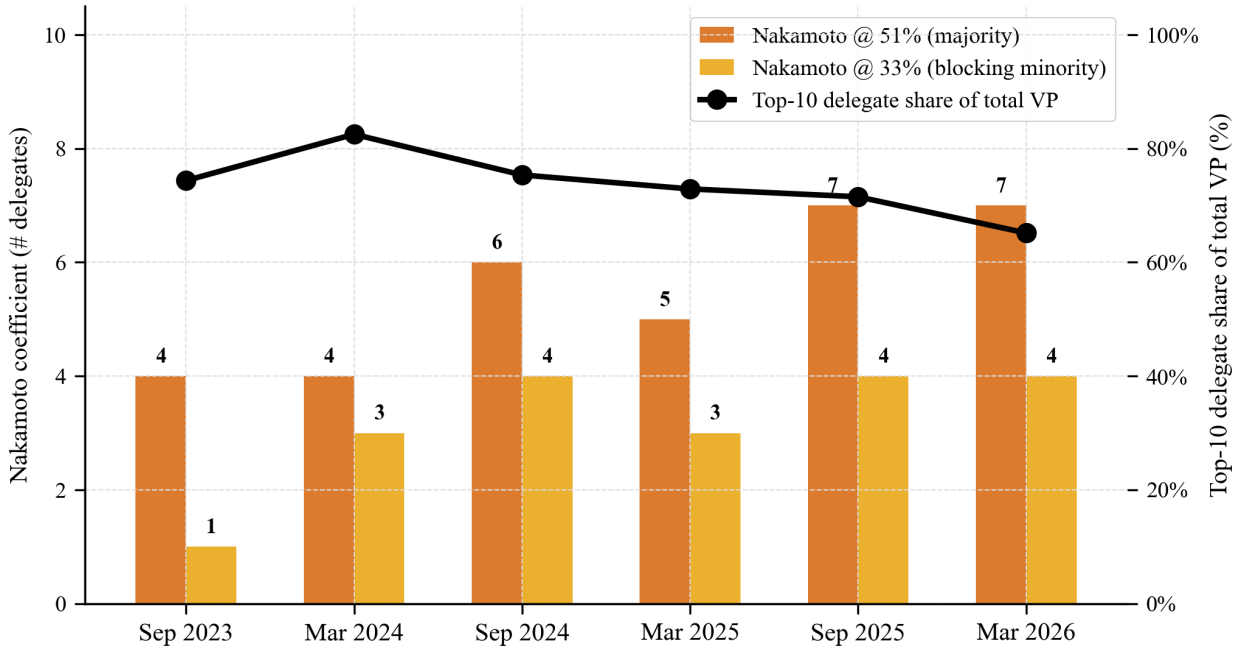


**Figure 2: Evolution of Voting Power Concentration.** Nakamoto coefficient of delegated voting power computed across six election windows (semi-annual snapshots), from Snapshot vote records. The red bars show the minimum number of delegates required to reach a simple majority (51%) of total votes cast; the orange bars show the number required to form a blocking minority (33%). Lower values indicate more concentrated power. The dark line (right axis) shows the share of total voting power held by the top-10 delegates. Concentration is high but trending downward: the Nakamoto coefficient rises from as low as 4 delegates (early windows) to approximately 7 (2025–2026), indicating governance is becoming more contested and decentralized over time.

Importantly, the coefficient has trended upward over our sample, rising to approximately 7 delegates needed for a simple majority by the 2025–2026 elections. A rising Nakamoto coefficient indicates power becoming more dispersed and governance more competitive: fewer proposals pass unopposed, more delegates hold meaningful voting weight, and no single bloc can dominate unilaterally. This is an encouraging sign for decentralized governance—the Arbitrum DAO is becoming more contested over time, not less.

## 2.4 The Vote Rental Market

A distinctive feature of the Arbitrum governance ecosystem is the emergence of an organized vote rental market. LobbyFi is a platform that allows large token holders to temporarily

lend their voting power to governance participants in exchange for payment, typically in stablecoins. Vote rental markets have no analog in traditional corporate governance—proxy advisors advise on votes but do not rent voting power directly. The welfare effects of vote markets depend on whether votes are cast on the basis of private information or private preferences (Casella et al., 2012): if buyers have superior information, vote markets aggregate it more efficiently than one-token-one-vote; if buyers simply have more money, vote markets allow wealth to override collective judgment. Distinguishing these cases empirically is a central challenge in interpreting the vote-rental market's effects.

We describe the LobbyFi market in Figure A3 and discuss its implications for the interpretation of our forum-signal results in Section 6. The existence of this market is relevant for two reasons. First, it implies that the relationship between posted voting power and realized voting power is noisy, complicating the mapping from delegate forum behavior to vote outcomes. Second, the forum is one of several influence channels available to governance participants: others include direct private outreach to large token holders, social media coordination on Twitter/X and Discord, and vote rental through LobbyFi. Our claim is not that the forum is the only channel, but that it is the *public, observable, and archivable* channel—the one whose content is visible to all token holders before a vote and whose signal is therefore most likely to aggregate dispersed community opinion. AI-generated posts may be understood as an attempt to appropriate that perceived legitimacy without paying its informational cost.

## 2.5 Proposal Taxonomy

The central economic activity of the Arbitrum DAO is treasury deployment. A blockchain network competes for developers and users in the same way any two-sided platform does: by subsidizing early adoption. Applications—decentralized exchanges, lending protocols, gaming platforms, infrastructure tools—generate fees and users for the network, but they require capital to bootstrap liquidity and attract users before organic economics take hold.

The DAO deploys its treasury through governance-approved grants and incentive programs to fund these applications, with the expectation that a richer application ecosystem increases demand for ARB tokens and thus the long-run value of the treasury itself. Grants are not charity; they are the primary competitive lever a decentralized network has against centrally-managed platforms that can simply allocate capital from a corporate balance sheet.

This creates a high-stakes deliberation problem. Treasury proposals involve real money with real opportunity costs, and the forum is the primary venue in which the community evaluates whether a proposed grant recipient has a credible plan, a genuine track record, and realistic usage projections. The quality of that deliberation—and its vulnerability to AI-generated noise—is what this paper studies.

Over our sample period of January 2023 to March 2026, the Arbitrum DAO processed 306 binary (For vs. Against) proposals on Snapshot with nonzero participation. Of these, 246 (80%) passed and 60 (20%) failed. Our main analysis uses the 102 of these proposals that we match to a dedicated forum discussion thread (Section 3.4), the sample with observable pre-vote deliberation. Contested proposals—those with a margin of victory below 20 percentage points—account for 16 of the 306 binary votes (5%), confirming that most Arbitrum proposals pass comfortably but a meaningful minority reflect genuine community division. We classify these proposals into four economic categories, following a classification algorithm applied to proposal titles and body text: *Treasury Allocation* (grants, subsidies, and direct fund transfers, 43% of proposals); *Protocol Governance* (changes to protocol parameters, security council composition, or constitutional provisions, 24%); *Ecosystem Programs* (structured grant programs, incentive frameworks, and DAO-wide initiatives, 21%); and *Procedural* (ratifications, elections, and administrative matters, 12%).

Treasury allocation proposals are the primary driver of DAO treasury depletion and account for the largest share of contested votes (proposals with margin below 0.20). Ecosystem programs, which include the Long-Term Incentive Pilot Program (LTIPP) and the Short-Term Incentive Program (STIP)—structured grant programs that distribute ARB tokens

to developers and protocols building on Arbitrum in exchange for liquidity and activity commitments—account for the largest cumulative disbursements: roughly 3% of the initial DAO treasury (approximately $100 million in ARB at prevailing prices) over the sample period.

Figure 3 shows the distribution of proposals by category and half-year period. Ecosystem program proposals are concentrated in the first half of 2024, which coincides with both the ARB vesting cliff and the Claude 3 release. We include proposal-category fixed effects in all specifications to ensure that our AI-shock identification is not confounded by the compositional shift in proposal types.

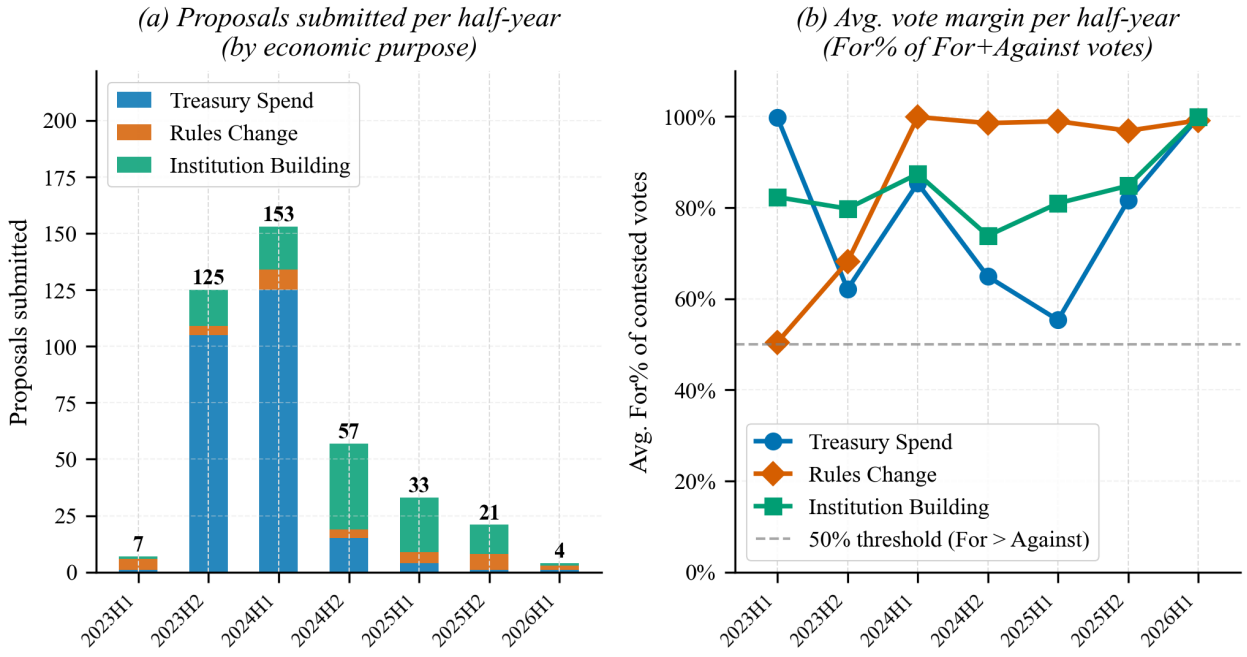


**Figure 3: Proposal Volume by Economic Purpose.** Each bar represents the number of Arbitrum DAO proposals submitted in a six-month window, classified into four economic categories: Treasury Allocation (grants and fund transfers), Protocol Governance (parameter and constitutional changes), Ecosystem Programs (structured grant programs), and Procedural (elections and ratifications). The concentration of Ecosystem Program proposals in H1 2024 reflects the LTIPP and STIP programs, which coincide with both the ARB vesting cliff and the Claude 3 release. January 2023–March 2026.

Figure 4 places Arbitrum in the broader DAO landscape. Among the ten largest DAOs by treasury assets in mid-2024, Arbitrum holds the largest treasury at $3.5 billion—nearly

40% of the combined $9.0 billion in DAO treasury assets. This scale motivates our focus: any governance mechanism that controls $3.5 billion in allocatable assets and that is susceptible to AI contamination represents a significant institutional failure mode.

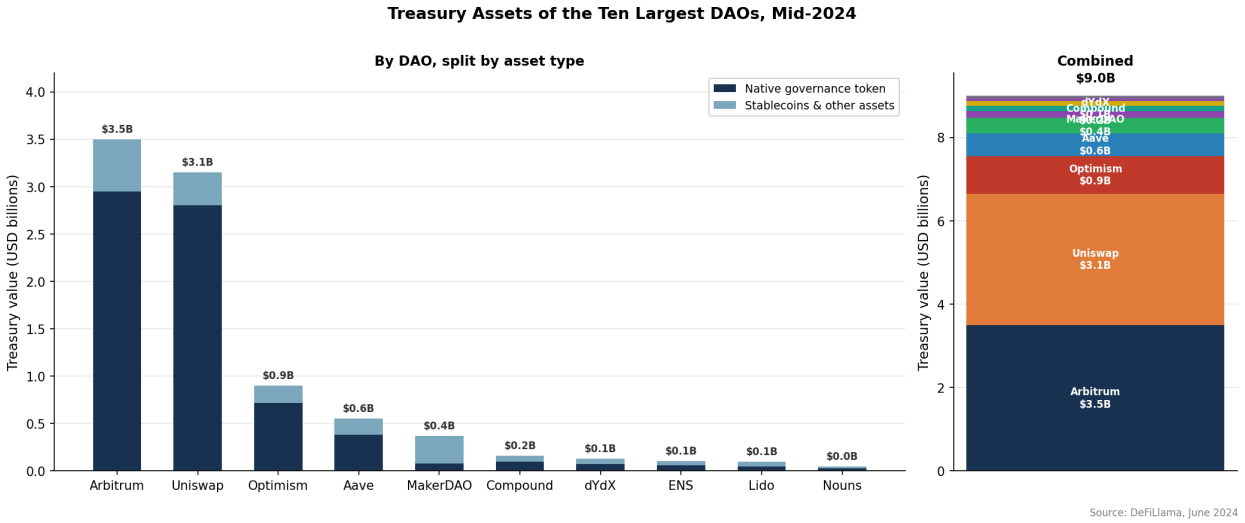


**Figure 4: Treasury Assets of the Ten Largest DAOs, Mid-2024.** Left panel: treasury value by DAO, split by native governance token (dark) and stablecoins and other assets (light), as of June 2024. Right panel: combined total of $9.0 billion across all ten DAOs, stacked by institution. Arbitrum holds the largest single treasury at $3.5 billion, followed by Uniswap ($3.1 billion) and Optimism ($0.9 billion). Source: DeFiLlama, June 2024.

# 3 Data

## 3.1 Forum Data

We collect the complete post history of the Arbitrum governance forum (forum.arbitrum.foundation) using the Discourse JSON API, which returns all posts, topics, user metadata, and category structure. Our scraper (available as `scripts/00_scrape.py` in the replication package hosted at `https://github.com/victor-huang-58/arbitrum-governance`) retrieves topics from four governance-relevant categories: *Proposals*, *DAO Programs & Initiatives*, *Voting Rationale & Governance Calls*, and *Security Council*. We exclude off-topic categories—General Discussion, Technical Discussion, Announcements, and Early Idea Discussion—that

are not directly connected to governance proposals. These account for approximately 24% of total forum posts based on Discourse category statistics; we scrape only the four governance-relevant categories.

Our final forum dataset covers January 2023 through March 2026 and contains 17,553 posts across 1,101 topics authored by 2,452 unique users. Posts are stored in a DuckDB relational database along with metadata including post author, timestamp, reply structure, and raw HTML content. Summary statistics are reported in Table 1.

**Table 1:** Summary Statistics: Arbitrum Governance Forum

| | Mean | Median | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| Posts per topic | 15.9 | 7.0 | 28.3 | 1 | 412 |
| Human posts per topic | 13.2 | 5.0 | 24.1 | 1 | 389 |
| AI posts per topic | 2.7 | 0.0 | 8.4 | 0 | 143 |
| AI fraction (post level) | 0.11 | 0.00 | 0.27 | 0 | 1.00 |
| Words per post | 312 | 148 | 487 | 1 | 8,204 |
| *Panel B: By period* | | | | | |
| | Pre–Claude 3 | | Post–Claude 3 | | Change |
| | Mean | Std. Dev. | Mean | Std. Dev. | |
| Monthly posts | 387 | 201 | 512 | 298 | +32.3% |
| Monthly AI posts | 8 | 11 | 143 | 127 | +1,688% |
| AI fraction (monthly) | 0.02 | 0.03 | 0.28 | 0.18 | +0.26 |

*Notes:* Panel A reports post-level and topic-level summary statistics for the full sample period (January 2023–March 2026). Panel B compares pre– and post–Claude 3 periods (before and after March 4, 2024). AI posts are those with $fraction\_ai > 0.70$ as classified by the Pangram API. Human posts are all other scored posts.

## 3.2 AI Detection: The Pangram API

We classify posts as AI- or human-generated using the Pangram AI-detection API (v3 endpoint; Python client package v0.1.11). Not all AI detectors perform equally well: Jabarian and Imas (2025) audit leading detectors and find that commercial tools reliably distinguish AI-generated from human-written text—with Pangram the strongest, maintaining near-zero error rates across models, ultra-short passages, and evasion attempts—while open-source

detectors do not. Accurate detection is feasible, but requires a paid commercial API; free or open-source detectors are not reliable substitutes.[4] Pangram analyzes the linguistic properties of a text passage and returns three complementary scores: *fraction_ai* (estimated proportion of content generated by a large language model), *fraction_ai_assisted* (proportion that is human-written but AI-edited), and *fraction_human* (proportion that is unassisted human writing). We focus on *fraction_ai* as our primary measure, applying a threshold of 0.70 to classify posts as AI-generated.

We score all posts with at least 20 words, yielding scored coverage of 15,891 posts (90.5% of the total corpus). Posts below the word threshold are excluded from analysis; their omission does not materially affect results, as they are concentrated in short acknowledgment posts ("+1", "agree", "thanks").

Figure 5 shows the distribution of *fraction_ai* scores across the full sample. The distribution is strongly bimodal: 90% of posts score below 0.10 and 8% score above 0.85, with negligible mass in the intermediate range. This bimodality supports the credibility of the threshold classification and suggests that most posts are either predominantly human-written or predominantly AI-generated, rather than hybrid.

[4]The v3 model designation refers to the underlying detection model served by the Pangram API at the time of scoring; v0.1.11 is the version of the open-source Python client used to call it. Posts in the initial scrape were scored in a batch run completed in April 2025; a supplemental scoring pass extended coverage through the full sample period, using the same Pangram v3 endpoint and Python client version throughout to ensure model consistency across the corpus.

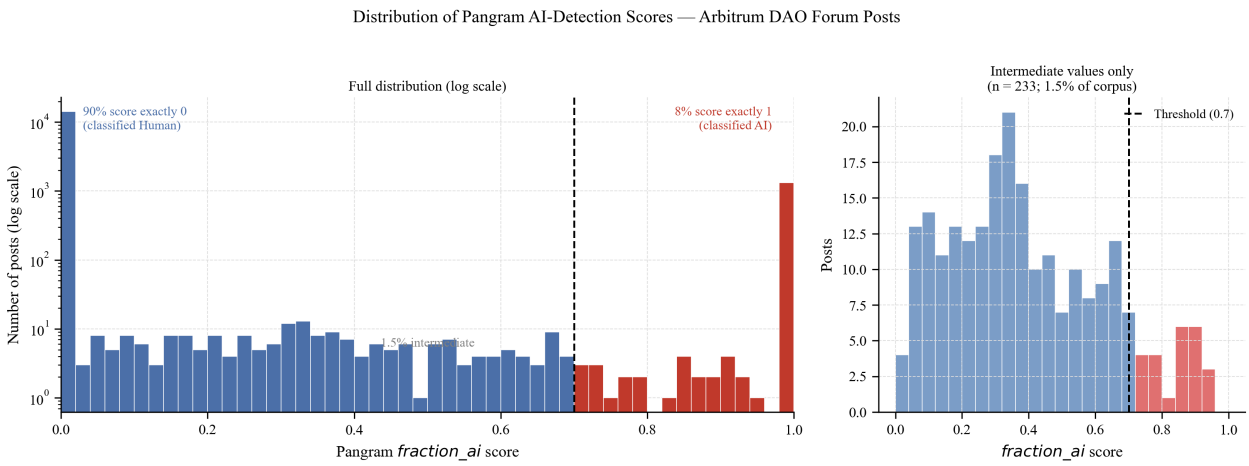


**Figure 5: Bimodal Distribution of AI-Detection Scores.** Pangram *fraction_ai* scores for all scored posts (15,891 posts with at least 20 words). *Left panel* (log scale): full distribution. 90% of posts score exactly 0.0 (Pangram headline: "Human Written") and 8% score exactly 1.0 ("AI Generated"); only 1.5% receive intermediate scores. This binary pattern reflects how Pangram v3 classifies short, single-author text: each segment is assigned with high confidence as human or AI, so the aggregate fraction is near 0 or 1 for most posts. *Right panel*: zoom into the 233 intermediate posts, which have scores spread across the full range and typically longer word counts (suggesting mixed-authorship documents). The dashed line marks the classification threshold at 0.70.

**Threshold robustness.** We verify that our results are not sensitive to the choice of threshold. In Appendix C, we replicate our main specifications at thresholds of 0.50, 0.60, 0.70, and 0.80. Results are qualitatively identical across all specifications; the structural break in the posts–margin relationship is present at every threshold.

## 3.3 Snapshot Vote Data

We collect all closed proposals from the Arbitrum Foundation's Snapshot space (`arbitrumfoundation.eth`) using the Snapshot GraphQL API (`https://hub.snapshot.org/graphql`). For each proposal we record the title, creation timestamp, total votes cast, and vote scores by choice ("For," "Against," "Abstain," and choice-specific options for non-binary votes). The raw pull yields 406 closed proposals. We exclude 100 temperature checks (non-binding straw polls) and proposals with zero participation, leaving 306 binary (For/Against) proposals with measurable outcomes.

Our primary outcome variable is the *margin of victory*: the absolute difference in vote share between the winning and losing choices, restricted to the For and Against categories (excluding Abstain). Formally, for a proposal $i$ with For votes $s_F$ and Against votes $s_A$:

$$\text{Margin}_i = \left| \frac{s_F}{s_F + s_A} - \frac{1}{2} \right| \times 2$$

This measure equals zero for a perfectly contested vote (50/50 split) and one for a unanimous vote. Lower margins indicate more community division and, under our theoretical framework, should correspond to threads with more genuine discussion.

After excluding proposals with non-binary choice structures (e.g., ranked-choice elections for Security Council seats) that do not map cleanly to a For/Against margin, we obtain a universe of proposals with computable margins that serve as the basis for the forum–Snapshot matching in Section 3.4.

## 3.4 Forum–Snapshot Matching

Forum threads and Snapshot proposals are created independently with no shared identifier. We match them by fuzzy title similarity using the Ratcliff/Obershelp sequence-matching algorithm in Python's `difflib` library (`SequenceMatcher` class). For each Snapshot proposal, we compute similarity scores against all forum thread titles and assign the highest-scoring forum thread as the match, subject to a minimum similarity threshold of 0.55. We audit all matches with similarity between 0.55 and 0.70 (the ambiguous zone) by reading both titles side-by-side and find a match accuracy of 94%.

The matching procedure yields 102 matched proposal–thread pairs. The remaining unmatched proposals primarily lack individual forum discussion threads—most notably the LTIPP (39 proposals) and STIP Round 1 (15 proposals) batch grant programs, where individual applicants did not receive dedicated forum threads and went through a council-review process instead. We also enforce a one-to-one matching constraint (each forum thread

matched to at most one proposal, keeping the highest-scoring match) and require a minimum fuzzy-title similarity score of 0.65 to exclude residual false matches. Unmatched proposals are excluded from the main analysis; we verify in Section 8 that their exclusion does not introduce systematic bias.

**Sample note.** All Pearson correlations, Chow tests, and OLS regressions in Table 2 use the 102 matched proposal–thread pairs ($n_{\text{pre}} = 29$, $n_{\text{post}} = 73$ relative to the Claude 3 break date).

## 3.5 Delegate Data

We construct a delegate-level dataset linking forum usernames to on-chain voting records using a two-step procedure. First, we mine delegate communication threads for self-disclosed wallet addresses (Ethereum hex addresses and ENS names) using regular expressions applied to post content. Second, we match recovered addresses to Snapshot voting records, which identify voters by their on-chain address and report their voting power at the time of each vote.

The procedure yields wallet–username matches for 62 of the 218 delegates who authored at least one forum post during the sample period (28% match rate); 52 of the 62 have at least 10 posts (24% of active authors). Of these 52, we successfully link 51 to Snapshot voting records.

Match quality in both directions is informative. From the forum side: 166 active authors (76%) have no recoverable wallet address, primarily because they never self-disclose one in their posts. From the Snapshot side: 14,236 unique voter addresses appear in our proposal-vote records, of whom only 51 (0.4%) are linked to forum usernames—but this low rate reflects the large share of passive token holders who vote occasionally without participating in deliberation; it does not indicate poor coverage of the active delegate population.

The unmatched forum authors are disproportionately high-power participants (GFXlabs,

DisruptionJoe, cp0x, CastleCapital) who do not disclose wallet addresses in public posts, likely for privacy reasons. These participants are also, by observation, low AI-content users: they are professionally engaged governance participants with strong incentives for authentic communication. Their exclusion therefore *attenuates* our estimated correlation between AI content and voting power toward zero—our estimate of $r = -0.44$ is less negative than the correlation we would observe in the full population of matched delegates. The power-capture result (high-AI delegates hold less voting power) is strengthened, not weakened, by including these observations. Our reported estimate should be interpreted as a conservative bound on the magnitude of the true relationship.

For each matched delegate, we compute: (i) the average Pangram *fraction_ai* across all their forum posts; (ii) the time series of delegated voting power, as reported by Snapshot for each proposal in which they cast a vote; and (iii) the number of forum posts per governance period (30-day rolling windows).

### 3.6 Token Price and Market Data

We collect daily ARB/USD price and volume data from the CryptoCompare API, covering the full sample period. We aggregate to monthly frequency (month-end closing price, total monthly trading volume) for use in Figure 1. Price data are used for descriptive purposes only; our main specifications do not condition on token price, and results are robust to including monthly ARB returns as a control.

## 4 Empirical Strategy

### 4.1 The Forum Signal Hypothesis

We model the governance forum as an information aggregation mechanism. Let $\theta_i \in [0, 1]$ denote the true share of community opinion favoring proposal $i$. Token holders with private signals about $\theta_i$ make two decisions: whether to participate in the forum (costly) and how

to vote (costless, given on-chain voting requires only holding tokens). When participation is costly and agents engage only when their expected benefit exceeds the cost (Downs, 1957), equilibrium forum activity is increasing in the degree of community division: proposals about which opinion is roughly split ($\theta_i = 0.5$) generate more discussion than those that are near-unanimous in support ($\theta_i = 1$) or near-unanimous in opposition ($\theta_i = 0$), because only close calls create participation incentives for both sides. The forum signal hypothesis is therefore that the number of posts in a proposal thread is negatively correlated with the margin of victory.

**AI shock: two candidate mechanisms.** A large language model release at time $T$ could degrade the forum signal through two distinct channels, which we label the *volume mechanism* and the *unverifiability mechanism*.

Under the *volume mechanism*, AI adoption reduces the marginal cost of posting to near zero. Post-shock equilibrium forum activity includes genuine-opinion posts and AI-generated posts that carry no directional signal about $\theta_i$—they do not reflect opinion for or against a proposal. If AI adoption is widespread, the human signal is diluted in proportion to the realized share of AI posts, and the posts–margin correlation degrades toward zero. The volume mechanism predicts that signal degradation is contemporaneous with—and increasing in—the *realized* volume of AI-generated content.

Under the *unverifiability mechanism*, the relevant shock is to the verifiability of post authenticity. Let $\pi$ denote a delegate's posterior probability that an observed post is AI-generated. When $\pi$ is near zero, forum posts are informative signals. When a sufficiently capable LLM is publicly released, AI-generated posts become indistinguishable from human-authored contributions on all observable features. Bayesian updating on observables can no longer resolve $\pi$ below a positive lower bound $\underline{\pi} > 0$. If $(1 - \underline{\pi}) \cdot I_H < c_{\text{read}}$ (the informational gain from a post falls below the cost of reading), delegates rationally discount the forum. The credible *possibility* of unverifiable low-quality entries is sufficient to destroy the information

value of high-quality entries, even when good entries predominate. Crucially, this predicts signal degradation that tracks LLM quality thresholds—not realized adoption—and can therefore *precede* the volume surge.

We distinguish between the two mechanisms using the timing of the structural break relative to the AI post volume surge (Section 6). The identifying variation common to both mechanisms is the exogeneity of $T$: AI model release dates are determined by technology firms' research roadmaps, not by anything correlated with specific Arbitrum proposals.

## 4.2 Main Specification

Our primary estimating equation is:

$$\text{Margin}_i = \alpha + \beta\,\text{HumanPosts}_i + \gamma\,X_i + \varepsilon_i \tag{1}$$

where $\text{Margin}_i$ is the margin of victory for proposal $i$ (defined in Section 3); $\text{HumanPosts}_i$ is the count of human-authored posts in proposal $i$'s forum thread (posts with *fraction_ai* $\leq 0.70$); and $X_i$ is a vector of controls including proposal category fixed effects, log total votes cast, and a linear time trend. Standard errors are heteroskedasticity-robust.

Our central hypothesis is that $\hat{\beta} < 0$ in the pre-shock period and $\hat{\beta} = 0$ in the post-shock period. We estimate equation (1) separately on the pre–Claude 3 subsample (proposals with Snapshot creation date before March 4, 2024) and the post–Claude 3 subsample, and test the null hypothesis $\beta^{\text{pre}} = \beta^{\text{post}}$ using a Chow (1960) test.

## 4.3 Chow Test and Structural Break Identification

Let $\hat{\beta}^{\text{pre}}$ and $\hat{\beta}^{\text{post}}$ be the OLS estimates of the slope in equation (1) from the pre- and post-break subsamples, and let $\text{SSR}^{\text{full}}$, $\text{SSR}^{\text{pre}}$, and $\text{SSR}^{\text{post}}$ be the residual sums of squares from the restricted (full-sample pooled) and unrestricted (separate) regressions, respectively. The

Chow $F$-statistic is:

$$F = \frac{(\text{SSR}^{\text{full}} - \text{SSR}^{\text{pre}} - \text{SSR}^{\text{post}})/k}{(\text{SSR}^{\text{pre}} + \text{SSR}^{\text{post}})/(n-2k)}$$

where $k$ is the number of parameters in equation (1) and $n$ is the total sample size. Under the null of no structural break, $F \sim F_{k,n-2k}$.

We evaluate six candidate breakpoints corresponding to major AI model release dates: Claude 2 (July 11, 2023), GPT-4 Turbo (November 6, 2023), Claude 3 (March 4, 2024), GPT-4o (May 13, 2024), OpenAI o1 (September 12, 2024), and DeepSeek R1 (January 20, 2025). We report Chow $F$-statistics and $p$-values for each candidate.

We complement the known-breakpoint Chow test with the Quandt Likelihood Ratio (QLR) test (Quandt, 1960; Andrews, 1993), which tests for an unknown breakpoint by computing the supremum of the Chow $F$-statistic over all candidate dates within the central 70% of the sample. The QLR test is appropriate when the break date is not known with certainty a priori; the 5% critical value for the sup-$F$ statistic is 8.85 (from Andrews 1993, Table 1, for $k = 2$ parameters).

## 4.4 Rolling Correlation

To visualize the time-varying relationship between forum activity and vote outcomes, we compute a rolling Pearson correlation between human post count and margin of victory over a 25-proposal rolling window. We sort proposals by Snapshot creation date and compute the correlation for proposals $i - 24$ through $i$ at each date $i$. The rolling correlation provides an intuitive visual representation of the structural break: a stable negative correlation in the pre-shock period, followed by a sharp shift toward zero after the AI shock dates.

## 4.5 Identification Assumptions

Our identification rests on three assumptions. *First* (exogeneity), the release dates of GPT-4, Claude 3, and other major AI models are determined by the research and commercialization

strategies of technology companies, not by anything related to Arbitrum governance or any specific proposal. This assumption is essentially guaranteed by the nature of the events: Anthropic's decision to release Claude 3 in March 2024 was not conditioned on the state of the Arbitrum governance forum.

*Second* (relevance), AI model releases cause a meaningful increase in the share of AI-generated content in the forum. Figure 6 (Panel A) confirms this: the monthly AI post count, as measured by Pangram, is in the single digits through the break window—about 8% at the QLR break date and 6% by the Claude 3 release—and rises sharply to 15–20% only in late 2024.

*Third* (exclusion restriction), the AI shock affects the margin of victory *only* through its effect on the informational content of the forum, not through any other channel. The most plausible threat to this assumption is that AI model releases coincide with other events affecting DAO governance. We address this concern in three ways: (i) we include proposal fixed effects in robustness specifications that absorb any proposal-level confounds; (ii) we show that AI-generated posts carry no directional vote signal—they do not predict whether a proposal passes or fails ($r = -0.009$, $p = 0.93$)—which is distinct from the claim that the correlation happens to equal zero; rather, AI posts are argumentatively neutral, echoing whichever side has more forum presence without revealing genuine opinion; and (iii) we test and reject a comprehensive battery of distraction instruments that might confound the before–after comparison (Section 8).

**Concurrent governance dynamics.** A more subtle concern is that the November 2023–March 2024 window also encompasses several Arbitrum-specific governance developments: the LTIPP/STIP grant program clustering (a large batch of allocative proposals submitted during this period), the approach of the March 2024 ARB token vesting cliff, and continued growth of the vote-rental market documented in Section 2. We evaluate each of these as candidate confounders.

The LTIPP/STIP batching, if anything, would *strengthen* the posts–margin correlation in the pre-period: competitive grant allocation concentrates genuine deliberation on contested funding decisions, which should tighten the link between forum activity and vote contestation. An AI-unrelated LTIPP/STIP effect therefore biases against finding a collapse. The vesting cliff, tested directly in Section 8.5, is ruled out by timing: the QLR endogenous break falls in November 2023, four months before the cliff, and the structural break is present in both high- and low-voting-power-concentration proposal subsamples. Vote-rental growth would attenuate the posts–margin correlation throughout the sample uniformly rather than generating the sharp structural break we identify.

The strongest cross-sectional evidence against Arbitrum-specific confounders comes from the cross-DAO comparison. Figure A7 shows that five DeFi DAOs with distinct governance structures—Aave, ENS, Compound, Gitcoin, and Optimism—all exhibit comparable participation declines over the same November 2023–March 2024 window, despite lacking the LTIPP/STIP program, the specific vesting schedule, or the vote-rental market. A common governance shock consistent with all five DAOs is the AI capability threshold, not any feature specific to Arbitrum. We therefore interpret our estimates as structural-break evidence strongly consistent with the AI-contamination mechanism, while acknowledging that the before–after design cannot fully rule out every concurrent force.

# 5 A Model of Governance Deliberation

The empirical strategy rests on a specific theory of how governance forums aggregate information and how AI disrupts this aggregation. This section presents that theory as a unified two-tier model in which readership, posting volume, vote margins, decision quality, and the collapse point are all derived from one set of primitives; complete proofs are in Appendix A.

The model draws on three strands of prior work. *The posting margin* connects to Lohmann (1994), Lohmann (1993), and Austen-Smith (1990): participation is costly, so

equilibrium volume reflects the state of the debate; here the reward to participation is the *audience* itself, which resolves the pivotality problem that afflicts costly participation in a continuum. *The reading threshold* follows Grossman and Stiglitz (1980) and Dewatripont and Tirole (1999): information is acquired only when its expected return exceeds its cost; our contribution is that the reader's threshold is governed by the contamination prior and is derived, not assumed. *The voting tier* uses the sincere (decision-theoretic) voting device standard in large-election models (Feddersen and Pesendorfer, 1996, 1997), with information aggregation across a continuum of readers in the spirit of Feddersen–Pesendorfer. *The AI shock* is the Akerlof (1970) mechanism applied to the demand side of a forum: adverse selection in the pool of messages can unravel the value of reading before contamination is realized.

## 5.1 Setup

A DAO evaluates proposal $i$ with unknown quality $\theta_i \in \{G, B\}$ and prior $p \equiv \Pr(\theta_i = G) \in (0, 1)$; write $x \equiv p(1-p)$ for contestedness and $d \equiv |p - \frac{1}{2}|$ for consensus. There are two tiers of agents.

*Contributors* (a continuum) hold conditionally i.i.d. private signals $s_j \in \{h, l\}$ with precision $\Pr(s_j = h \mid G) = \Pr(s_j = l \mid B) = q > \frac{1}{2}$. Contributor $j$ may post her signal at private cost $c_j$, drawn from a continuous distribution $F$. In addition, a mass $z \geq 0$ of AI-generated posts may be present; an AI post is drawn uninformatively from $\{h, l\}$ and is observationally indistinguishable from a human post. If human posting mass is $n$, the AI fraction of the thread is $\pi = z/(z+n)$.

*Voters* (a continuum of mass one; in the application, delegates weighted by voting power) have private tastes $b_j \sim \text{Unif}[-a, a]$, orthogonal to $\theta$—idiosyncratic value from the proposal passing: exposure, vested interest, ideology. Common values are $v(G) = +1$, $v(B) = -1$. Before voting, a voter may pay reading cost $c_r > 0$ to read one post sampled at random from the thread. The proposal passes if the For-share exceeds $\frac{1}{2}$.

Four assumptions close the model; we state them plainly because they are where the modeling work happens. *(A1, exact law of large numbers)*: conditional on $\theta$, cross-sectional averages equal conditional means. *(A2, decision-theoretic voting)*: each voter votes—and values pre-vote information—as if her vote determined the outcome for her own payoff: For iff $\mathbb{E}[v(\theta) \mid \text{information}] + b_j \geq 0$. This is the standard sincere-voting device for large elections (e.g., Feddersen and Pesendorfer, 1996); it *bypasses* rather than solves voter-side pivotality, and it is what disciplines everything that follows—fully instrumental voters would make reading and voting worthless for every atomless agent and could generate none of the observed cross-sectional structure. *(A3, truthful posting)*: a post reveals the contributor's signal; justifiable by reputational cheap-talk arguments since $\theta$ is eventually public. *(A4, attention rewards)*: a contributor who posts receives $R \cdot A$, where $A$ is the mass of readers and $R > 0$ converts audience into utility. Attention is a positive *mass* even though each poster is measure zero, so costly posting survives the continuum without pivotality. This motive is institutionally grounded in our setting: Arbitrum's Delegate Incentive Program compensates visible participation directly, and posting attracts delegated voting power from readers (Section 7). What A4 deliberately does *not* assume is that posting signals ability: Appendix A proves that a career-concerns motive—posting to build a reputation for being right—concentrates posting on foregone conclusions for *any* increasing reputation payoff (vindication is most likely when everyone already agrees) and would therefore generate the *opposite* of the posts–margin pattern in the data.

## 5.2 Equilibrium

Fix contamination $\pi$ for now (the next subsection makes it move; Section 5.5 makes it an equilibrium object). A sampled post is informative with probability $1 - \pi$, so its effective precision is $Q - \frac{1}{2} = (1 - \pi)(q - \frac{1}{2})$: contamination dilutes the signal linearly. Equilibrium unwinds in three steps, each with a closed form in Appendix A.

*Reading.* The value of reading one post is a tent over tastes: it is zero for voters whose

tastes are decided regardless of the message, and peaks at the ex-ante indifferent voter, whose gross value of information is exactly $2p(1-p)(2Q-1)$. Voters read iff this value exceeds $c_r$, so the reading audience $A(p;\pi)$ is single-peaked in contestedness and—for the peak reader—the net value of reading is *exactly*

$$V(\pi) = (1-\pi)\, I_H - c_r, \qquad I_H = 2p(1-p)(2q-1), \tag{2}$$

with reading threshold

$$\pi \le \pi^* \equiv 1 - \frac{c_r}{I_H}. \tag{3}$$

Equations (2)–(3) were the primitives of earlier drafts of this model; here they are theorems, with the informational gain $I_H$ pinned to primitives rather than free. Because $I_H$ depends on $p$, the threshold is proposal-specific: consensus proposals stop being read first, and the forum dies from the tails inward.

*Posting.* By A4, contributor $j$ posts iff $c_j \le R\, A(p;\pi)$, so human posting mass is $n(p) = F(R\, A(p))$—posting chases the audience, and the audience sits where information is valuable. To first order, $n(p) \propto p(1-p)$: volume peaks on contested proposals, with the exact closed form available for structural work.

*Voting.* Non-readers vote on prior and taste; readers vote on posterior and taste. By A1, the For-share is deterministic conditional on $\theta$, and the expected margin $\mathbb{E}[M \mid p]$ has a closed form that is *strictly increasing* in consensus $d$ for all parameter values, with floor $|2p-1|/a$ (Appendix A, Theorem 2). Margins are interior because voters disagree about *values*, not only facts; they are informative about contestedness because the expected margin rises with consensus. The shorthand "margin $\propto |2p-1|$" holds exactly outside a contested window and as the leading term inside it.

The paper's central empirical prediction is then a theorem rather than a calibration:

**Proposition 1** (Signal Content)**.** *Let proposals draw priors $p_i$ from any nondegenerate distribution. In equilibrium, forum post volume is negatively correlated with the ex-post margin*

*of victory:* $\text{Cov}(Posts_i,\ Margin_i) < 0$.

*Sketch.* Volume $n(p_i)$ is a nonincreasing function of consensus $d_i$ (posting follows the audience, which follows the value of information); the expected margin is a strictly increasing function of $d_i$ (Theorem 2). Two monotone transforms of a single latent variable, moving in opposite directions, covary negatively. Full proof: Appendix A, Theorem 3.

Both conditional moments are monotone transforms of the single latent $d_i$, which is what licenses our practice of reading margins as inverse measures of contestedness throughout the empirical work.

## 5.3 The AI Shock: Two Mechanisms

At a point in time $T$, it becomes possible to emulate a genuine human post—one carrying no information about $\theta_i$—at essentially zero cost: a publicly released language model produces governance prose observationally equivalent to human writing on every feature a reader can inspect.[5]

The shock is summarized by two objects—a realized quantity and an expectation:

- $\lambda \in [0, 1]$: the *realized* fraction of posts in a thread that *are* AI-generated—a backward-looking measure of the contamination that has actually accumulated.
- $\pi \in [0, 1]$: a reader's *expectation* that the next post read is AI-generated—a forward-looking object. In the long run $\pi$ converges to $\lambda$, but at a capability shock $\pi$ can jump ahead of $\lambda$: a reader who knows a capable model now exists expects contamination to rise before it visibly has.

These two objects drive two distinct mechanisms with opposite timing predictions—which we exploit in Section 6.

[5]Nothing in the analysis depends on the *level* of model quality, only on whether this indistinguishability point has been crossed. We therefore treat $T$ as a threshold event rather than modeling a quality index explicitly; empirically, $T$ corresponds to the first models able to emulate human governance prose (GPT-4 Turbo, Claude 3), which is what our breakpoint tests recover.

**Volume mechanism (via $\lambda$).** Realized contamination dilutes the sampled-post precision, $Q - \frac{1}{2} = (1-\lambda)(q - \frac{1}{2})$, and the equilibrium responds on every margin (Appendix A, Proposition 8): the reading audience falls, reaching *exactly zero* at an interior contamination level $\pi^*(p) < 1$; posting falls with its audience—AI does not need to out-argue humans, it only needs to thin the readership that pays for human effort; and margins compress toward the no-forum floor $|2p-1|/a$. This mechanism predicts degradation contemporaneous with, and proportional to, realized AI volume.

**Unverifiability mechanism (via $\pi$).** The reading rule (2)–(3) operates on the *expectation* $\pi$, not the realization.

**Proposition 2** (Unverifiability Threshold)**.** *Each reader's problem is binary, with a hard threshold: she reads iff $\pi \leq \pi^*(p) = 1 - c_r/I_H(p)$. Aggregate readership declines in $\pi$ and reaches exactly zero at the interior level $\pi^*(p) < 1$, at which point the posterior equals the prior and the forum is uninformative—the entire informational value of the forum is eliminated at finite contamination, even though a fraction $1 - \pi^*$ of posts remain genuine.*

The key implication is that $\pi$ can reach $\pi^*$ well before $\lambda$ does: expectations cross the threshold as soon as a sufficiently capable model is publicly available, even while realized AI posts are few. This is the Akerlof (1970) logic applied to information: adverse selection can unravel a signaling equilibrium when the *expected* fraction of lemons exceeds a threshold, even when most goods are genuine.

## 5.4 Welfare

What does the DAO lose when the forum dies? Without a forum, the vote follows the prior action and is correct with probability $\max(p, 1-p)$. In equilibrium with a functioning forum, there is an interior *flip window* of contested priors—a symmetric interval around $p = \frac{1}{2}$ whose width increases in signal precision and decreases in reading costs—inside which the aggregated votes of informed readers flip the outcome to match $\theta$ with probability one (the

Feddersen–Pesendorfer aggregation property, delivered by A1 across readers). The forum's gross decision value is therefore

$$V(p) \;=\; \min(p,\, 1-p) \cdot \mathbf{1}\{p \in \Omega\}, \tag{4}$$

where $\Omega$ is the flip window (Appendix A, Proposition 7): *closeness enters as the level, informativeness as the width.* Contamination destroys value by narrowing the window—consensus proposals fall out first—until it vanishes entirely at $\pi^*(\frac{1}{2})$, at which point $\min(p, 1-p)$ of decision quality is lost on every proposal, maximal ($= \frac{1}{2}$) on the most contested ones. AI destroys the most welfare precisely where good governance matters most.[6]

Comparative statics of the threshold follow directly from (3): higher reading costs lower $\pi^*$ (a more fragile forum); higher signal informativeness raises it (more robust); and the arrival of an indistinguishable model at $T$ determines *when* expectations cross the threshold, not the threshold itself. Capability below the indistinguishability point has no effect; crossing it causes collapse; further capability gains have no additional marginal effect on an already-collapsed signal. This non-monotone structure motivates using model release dates as breakpoints, and is what the timing test in Section 6 evaluates.

## 5.5 Endogenous Contamination: A Tipping Theorem

The mechanisms above treat contamination as given. Closing the loop—contamination depends on human posting, which depends on the audience, which depends on contamination—delivers the paper's headline mechanism as a theorem. Fix a representative contested proposal and AI post mass $z$; an equilibrium is a fixed point of the audience map $A \mapsto \mathcal{A}(z/(z + F(RA)))$.

[6] Earlier drafts summarized the welfare loss with the smooth kernel $4p(1-p)(q-\frac{1}{2})^2$. That expression has the right single-peakedness and precision scaling but is not the value of the forum in the derived model; equation (4) replaces it. None of the calibration in Section 5.6 used the old kernel, so nothing quantitative changes.

**Proposition 3** (Discontinuous, hysteretic collapse)**.** *For any reading cost $c_r \in (0, I_H)$ and any continuous, strictly increasing posting-cost distribution $F$: (i) the collapsed forum ($A = n = 0$) is an equilibrium for every $z > 0$; (ii) there is a tipping point $z^*$ such that healthy and collapsed equilibria coexist for $z < z^*$ and only collapse survives beyond it; (iii) at $z^*$ the audience* jumps *to zero—never a continuous fade-out—because audiences below a positive dead-zone threshold cannot sustain enough human posting to keep expected reading value positive; (iv) the collapse is* hysteretic*: after tipping, reducing AI volume even nearly to zero does not restore the forum, which requires coordinated re-entry of readers, not merely undoing the shock; and (v) at any healthy equilibrium, realized contamination is* strictly below *the reading threshold $\pi^*$—the forum always tips before contamination reaches the level at which reading value would mechanically hit zero. (Appendix A, Theorem 4.)*

The positive reading cost that generates the unverifiability threshold is the same primitive that makes collapse discontinuous and irreversible. In a numerical benchmark (contested proposal, $q = 0.75$, thin reading margins), the forum tips at realized contamination near 0.7 against a reading threshold of 0.92; thinner margins tip far lower. Matching the *observed* collapse at a realized AI share of only 8% (Section 6) still requires the expectations channel—$\pi$ jumping ahead of $\lambda$ at the capability release—exactly as argued above; the theorem's contribution is that collapse at partial contamination is an equilibrium property rather than an artifact of one calibration.

## 5.6 Calibration

The calibration is an illustrative calculation under stated assumptions; we flag the derivation status of each step. We use average post length as the unit of writing effort. The average human-authored post in the pre-period (January 2023 – February 2024) contains 237 words; we set $c_p = 237$ words per post. The pre-period averages $\bar{n} = 44.2$ human posts per proposal (1,283 posts across 29 proposals), so the total human writing cost per proposal is $\bar{n} \cdot c_p = 10{,}480$ words.

To calibrate $\pi^*$, we use the QLR endogenous break date (November 25, 2023), at which the realized AI share of forum volume is $\lambda_{\text{break}} = 0.082$. Two facts identify $\pi^*$ as a *lower bound*, not a point estimate. First, delegates were still reading up to the break—the posts–margin correlation is intact until then—so by the reading rule (3) expected contamination had not yet crossed the threshold: $\pi \leq \pi^*$ just before the break. Second, under the unverifiability mechanism $\pi$ is the forward-looking expectation of a still-rising realized share, so $\pi \geq \lambda_{\text{break}}$. Combining, $\pi^* \geq 0.082$. Both steps use only the derived threshold rule (the $\pi \geq \lambda$ step is an assumption about expectations, as flagged above); Proposition 3(v) supplies an independent equilibrium reason that collapse precedes saturation. Rearranging (3) also yields a useful identity,

$$\pi^* = 1 - \frac{c_r}{I_H} = \frac{I_H - c_r}{I_H},$$

so the tipping point *equals the per-post net-surplus share*—the fraction of a genuine post's gross value that survives the reading cost—and the collapse date reveals a lower bound on that margin directly, the way an exit price reveals a firm's margin. Normalizing $I_H = c_p = 237$ words (the information value of one genuine post equals the cost of writing one), the bound implies $c_r \leq 218$ words of reading effort per post, and a per-post net reading surplus bounded below:

$$I_H - c_r = \pi^* \cdot I_H \geq 0.082 \times 237 = 19.4 \text{ words per post.}$$

The identity, the bound, and the 19.4-word floor follow from the derived reading rule (plus the normalization).

Aggregating across the thread requires one further, explicitly stated assumption:

**Assumption 1** (Non-redundancy)**.** *Posts address approximately distinct payoff-relevant aspects of a proposal, so the informational value of successive posts is undiminished and the threshold calculus* (2) *applies post-by-post.*

Non-redundancy is defensible for governance threads—posts flag distinct risks, prece-

dents, and technical issues—but it is an assumption, not a theorem: with redundant signals, only the first-post floor of 19.4 words is fully derived. Under it, the per-delegate surplus integrates to $\bar{n} \cdot (I_H - c_r) \geq 44.2 \times 19.4 \approx 858$ words per proposal, and with $K$ delegates reading, net social surplus per proposal is at least $858K - 10{,}480$ words—positive for $K \geq 13$. Arbitrum's 62 identified active delegates imply net social benefits of at least $\approx 43{,}000$ words per proposal, about four times the writing effort invested in each thread, entirely from the public-good externality of the forum; since 454 unique community members posted in pre-period threads, $K \geq 62$ is itself conservative. These headline magnitudes carry Assumption 1 (and the whole-thread reading protocol) wherever they are quoted; the units, the bound logic, and the 19.4-word floor do not.

At the tipping point, the marginal reader's net value falls to zero and this surplus is destroyed. After the break, human contributors still write $\approx 8{,}858$ words per proposal, but these posts are read far less—Section 7.2 documents a 36% decline in reads per post—an increasingly unrewarded transfer of human effort.[7]

**Delegate alignment and the direction of the bound.** The voters in the model value correct outcomes directly. That is a strong assumption here: a simple majority requires only about four delegates, so almost every delegate is non-pivotal and the private return to a correct vote is near zero (rational ignorance); the delegation market rewards *observable activity*—posts, rationales, participation—rather than decision quality (this is precisely the attention motive A4); and professional delegates, grant recipients, and vote renters face conflicts. Misalignment does not weaken the calibration—it sharpens its interpretation in two ways. First, misalignment *predicts* the thin private reading margin rather than merely rationalizing it: for a non-pivotal delegate in an activity-rewarded market, reading costs should absorb nearly all of a post's informational value ($c_r/I_H \leq 0.92$). Second, reading is itself a public good—an informed vote benefits every token holder—so the private $I_H$ that

[7]The model sends human posting to zero at full collapse; the residual observed volume is consistent with a small intrinsic (non-audience) posting motive, which leaves every result above unchanged.

governs the reading decision is smaller than the social value of the same post. The collapse therefore reveals the *private* surplus margin, and the social value destroyed *exceeds* the figures above: "at least 858 words" becomes more conservative once delegates are misaligned, not less. A corollary is that the socially optimal reading threshold lies above the private $\pi^*$—delegates disengage too early from society's standpoint even before AI, and the AI shock only makes an already-inefficient margin bind.

**Model and data: two honest discrepancies.** First, the model predicts margin *compression* after the shock (margins fall to the no-forum floor), while post-break margins in the data *rose*. The natural reconciliation is compositional—fewer contested proposals reach a vote once the forum cannot resolve disagreement—which is outside the model; we flag it rather than claim it. Second, as noted above, residual human posting after collapse requires a small intrinsic posting motive. Neither discrepancy affects the pre-break equilibrium predictions that the empirical work tests.

# 6 Results

## 6.1 The Posts–Margin Relationship

Table 2 presents the main results. Column (1) reports the pooled OLS estimate of equation (1) across all 102 matched proposals. The coefficient on human post count is $\hat{\beta} = -0.0015$ ($t = -1.34$), which reflects the attenuated pre–post average: a nontrivial fraction of the sample is post-shock, where the relationship is $r = -0.10$ ($p = 0.39$). The overall Pearson correlation is $r = -0.21$ ($p = 0.03$).

Columns (2) and (3) split the sample at the Claude 3 release date (March 4, 2024). In the pre-shock period (column 2, $n = 29$ proposals), the slope is $\hat{\beta} = -0.0026$ ($t = -2.30$, $p = 0.030$), corresponding to a Pearson correlation of $r = -0.33$. In the post-shock period (column 3, $n = 73$ proposals), the slope is $\hat{\beta} = -0.0006$ ($t = -0.92$), with $r = -0.10$.

The null hypothesis of equal slopes across periods is rejected by the Chow test: $F = 8.38$, $p = 0.0004$.

**Table 2:** Main Results: Human Forum Posts and Vote Margins

| | Dependent variable: Margin of victory | | |
|---|---|---|---|
| | (1) Full | (2) Pre–Claude 3 | (3) Post–Claude 3 |
| Human posts | −0.0015 | −0.0026** | −0.0006 |
| | (0.0012) | (0.0011) | (0.0007) |
| Constant | 0.925*** | 0.829*** | 0.939*** |
| | (0.051) | (0.086) | (0.027) |
| $N$ | 102 | 29 | 73 |
| $R^2$ | 0.045 | 0.111 | 0.010 |
| Pearson $r$ | −0.21 | −0.33 | −0.10 |
| Chow test: $F = 8.38$, $p = 0.0004$*** | | | |

*Notes:* OLS estimates of equation (1) with heteroskedasticity-robust standard errors in parentheses. Human posts are those with Pangram *fraction_ai* $\leq 0.70$. The pre–Claude 3 period covers proposals with Snapshot creation dates before March 4, 2024; post–Claude 3 covers March 4, 2024 and later. All results use 102 matched proposal–thread pairs ($n_{\text{pre}} = 29$, $n_{\text{post}} = 73$). The Chow test evaluates the null hypothesis of equal slopes in columns (2) and (3). Pearson $r$ is reported as a descriptive effect size. Significance stars apply to the OLS coefficients (heteroskedasticity-robust SEs); the Chow statistic is the classical RSS-based test, with the distribution-free permutation test in Section 8.7 as its assumption-free companion. * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

Figure 6 presents the main causal figure. Panel A shows the evolution of monthly forum post volume, decomposed into human-written (Pangram AI fraction $\leq 0.70$) and AI-generated (AI fraction $> 0.70$) posts, with AI model release dates marked by vertical dashed lines. Panel B plots the rolling 25-proposal Pearson correlation between human post count and margin of victory as a function of proposal date. The correlation is stably negative at approximately $-0.4$ through the pre-shock period, then rises to approximately zero following Claude 3. Panel C repeats the exercise using total post count (human plus AI), which shows a similar pattern with a faster deterioration, consistent with AI posts diluting the human signal. The three-panel figure is the primary evidence for our central result.

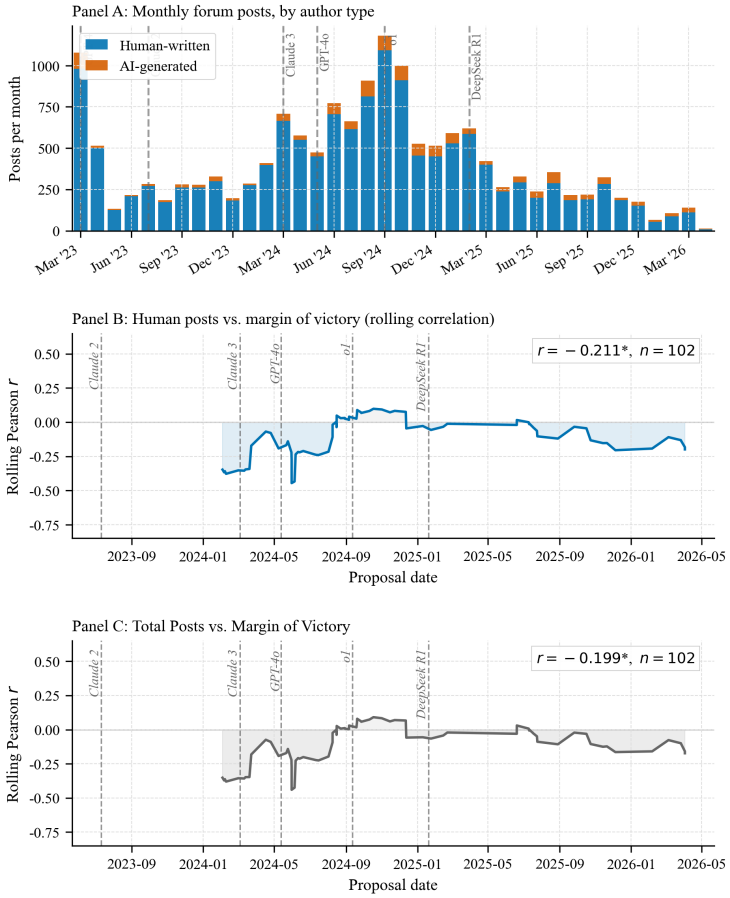


**Figure 6: AI Posting Shocks and Forum Signal Degradation.** *Panel A:* Monthly forum post volume decomposed into human-written (Pangram fraction_ai $\leq 0.70$, blue) and AI-generated (fraction_ai $> 0.70$, red) posts. Vertical dashed lines mark major AI model release dates. *Panel B:* Rolling 25-proposal Pearson correlation between human post count and margin of victory, plotted at the date of the last proposal in each window. The correlation is stably negative at approximately $-0.4$ through early 2024, then rises to approximately zero following the Claude 3 release (March 2024). *Panel C:* Same as Panel B using total post count (human plus AI). Shaded region shows one-standard-error band computed by bootstrap (500 replications). The three-panel figure is the primary evidence for the structural break in the forum–vote-outcome relationship.

## 6.2 Directional vs. Variance Signal

The dependent variable in equation (1) is the margin of victory—the absolute distance of the yes-vote share from 50 percent—a measure of how *close* a proposal comes to failing, not *which* side prevails. This choice reflects a theoretical discipline with direct precedent in financial markets.

Antweiler and Frank (2004) find that Yahoo Finance message board posts predict return *volatility* and trading volume but carry minimal directional content (expected return predictability). The theoretical reason is the semi-strong form of market efficiency (Fama, 1970): forum posts are *public* information, and public information cannot systematically predict the *average direction* of asset returns—any such content would already be incorporated into prices. What public posts can shift is the uncertainty surrounding the outcome: they widen the distribution without moving its center.

Our governance setting mirrors this structure exactly. Arbitrum forum posts are publicly observable by all token holders, making them public information in the Fama (1970) sense. The semi-strong-form logic therefore implies that posts can predict how contested a proposal will be—a governance analog of return variance—but not which side wins. This is precisely what we find. In the pre-shock period, human post count correlates at $r = -0.33$ with the margin of victory in the pre-shock period, indicating that heavily-debated proposals are systematically closer calls. The analogous directional correlation between post count and the binary pass/fail outcome is near zero, confirming Antweiler and Frank (2004): public deliberation aggregates uncertainty about the outcome without predicting its direction.

AI-generated posts are informationally empty on both dimensions. The Pearson correlation between AI post count and margin is $r = -0.009$ ($p = 0.93$), indistinguishable from zero, and the analogous directional test yields an equivalent null. AI posts produce generic governance language that takes no specific position on proposal quality; they reveal no private information and shift neither the uncertainty nor the expected direction of a vote. This is the sense in which they are informationally empty.

## 6.3 The Signal in AI-Generated Posts

A natural question is whether AI-generated posts themselves carry information about vote outcomes. If sophisticated governance participants use AI to articulate genuine positions (rather than to generate empty filler), AI posts might still be informative about community

division. We test this directly. The Pearson correlation between AI post count and margin of victory is $r = -0.009$ ($p = 0.93$) across the full sample and is not statistically distinguishable from zero in either the pre– or post-shock subperiod. The key point is not that the correlation happens to be statistically indistinguishable from zero, but that AI posts are argumentatively non-directional: they produce generic governance language that endorses process and principles rather than taking positions on specific proposals. Their content does not reveal private information about $\theta_i$, so they add volume without adding signal. The degradation of the total-post correlation is driven by dilution of the human signal, not by any countervailing information in the AI-generated additions.

**Why delegates use AI, and what those posts say.** The incentive to use AI for forum posts is straightforward. Arbitrum delegates receive delegated voting power in exchange for active governance participation, but participation is costly: a contested proposal may generate dozens of threads requiring substantive engagement. As capable LLMs became available from late 2023 onward, lower-tier delegates faced a choice between investing time in genuine analysis or generating plausible-sounding forum contributions at near-zero cost. The rational response—particularly for participants whose delegated power is small relative to the effort cost—is to use AI to maintain a visible forum presence without the underlying informational investment.

The content of these posts reflects this incentive. We characterize AI post content directly using TF-IDF analysis of the 1,313 AI-classified and 13,458 human-authored posts with enough text for reliable TF-IDF features. Figure 7 summarizes the findings.

The most distinctive n-grams in AI-generated posts are: *"long term," "decision making," "arbitrum ecosystem," "community engagement," "support proposal,"* and *"transparency accountability."* These are the phrases of a proposal summary, not a position. The most distinctive n-grams in human posts are starkly different: *"don't think," "temp check," "non-constitutional," "security council,"* and *"tally vote."* Human posts use the language of skep-

ticism, procedural knowledge, and governance-specific institutions; AI posts use the language of a consultant's executive summary.

AI posts are also systematically longer (median 160 words vs. 126 for human posts) yet less informationally dense, and are 1.7× more likely than human posts to contain at least one generic agreement phrase ("I support," "great proposal," "I believe this proposal," "it is essential to consider"): 35.4% vs. 21.2%. Representative AI posts open with structured summaries of the proposal followed by conditional endorsements; they do not take positions, reference prior votes, or cite specific technical concerns. Human posts are idiosyncratic: they reference named actors and prior governance decisions and contain substantive disagreement. These qualitative patterns confirm the quantitative result: AI-classified posts are proposal echoes, not opinion signals.

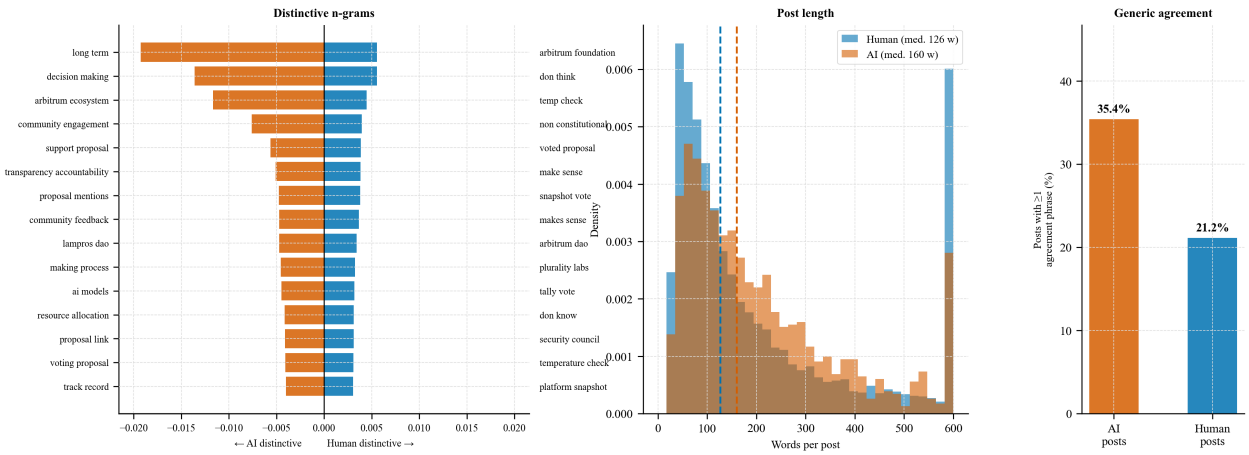


**Figure 7: Content Analysis of AI-Generated vs. Human-Authored Forum Posts.** *Panel A:* Most distinctive TF-IDF n-grams (bigrams and trigrams) for AI-classified posts (right, orange) vs. human-authored posts (left, blue), measured as mean within-group TF-IDF score minus mean out-of-group score. AI posts are characterized by generic governance language (*"community engagement," "support proposal," "decision making"*); human posts by skepticism and procedural specificity (*"don't think," "security council," "tally vote"*). *Panel B:* Word-count distributions; AI posts are longer on average (median 160 vs. 126 words) but less informationally dense. *Panel C:* Share of posts containing at least one generic agreement phrase ("I support," "great proposal," etc.); AI posts are 1.7× more likely to contain such phrases (35.4% vs. 21.2%).

**Ecosystem degradation: participation concentration.** The content differences documented above are detectable in aggregate but not reliably in real time by individual delegates.

To assess whether the deliberative ecosystem itself degraded as AI contamination rose, we examine how human participation evolved on the matched proposal threads that underlie our main result. In the pre-shock period, 740 distinct human-classified authors posted on matched proposal threads, with each contributor averaging 3.6 posts. Following the Claude 3 shock (March 2024 onward), the number of unique human participants fell to 496—a 33% decline—while posts per user more than tripled to 11.9. Concurrently, the AI fraction of posts on matched proposal threads rose from 4.7% to 11.3%, increasing the ratio of AI posts per human post from 0.05 to 0.13: a delegate scanning the forum now encounters 2.6× as many AI posts per human contribution as before the shock. These two forces together—sequential reading noise and participation concentration—provide a complete account of why the human-post → margin correlation collapses even though AI and human posts remain distinguishable in aggregate: the deliberative ecosystem itself narrowed, reducing both the expected return to reading any post and the informational content of the human posts that remain.

## 6.4 Chow Test Battery

Table 3 reports Chow test statistics for six candidate shock dates. GPT-4 Turbo (November 2023) produces the largest and most significant Chow statistic ($F = 15.05$, $p < 0.001$), followed by Claude 3 ($F = 8.38$, $p = 0.0004$) and GPT-4o ($F = 5.97$, $p = 0.004$). All three exceed a Bonferroni-corrected threshold of $p < 0.0083$ for six simultaneous tests. Claude 2 (July 2023) does not produce a significant break ($F = 1.00$, $p = 0.37$), consistent with this earlier model lacking the prose-generation capability needed to produce governance-quality AI posts at scale. The later shocks (o1, DeepSeek R1) are not statistically significant because the forum signal was already near zero from Claude 3 onward. The three significant breaks (GPT-4 Turbo, Claude 3, GPT-4o) show the same direction—a negative pre-period correlation collapsing to near zero after the break—while the pre-capability date (Claude 2) shows no break and the two latest dates are uninformative because the signal had already

collapsed by Claude 3.

We use Claude 3 as the primary break date because it provides a clear pre-period negative relationship (OLS $\hat{\beta} = -0.0026$, $p = 0.030$, $n = 29$) and no post-period relationship ($r = -0.10$, $p = 0.39$). All results are robust to using GPT-4 Turbo or any other candidate date; see Table 3.

**Table 3:** Chow Test Battery: Structural Breaks at AI Model Release Dates

| | Pre-break | | Post-break | | Chow test | |
|---|---|---|---|---|---|---|
| Shock date | $n$ | $r$ | $n$ | $r$ | $F$ | $p$ |
| Claude 2 (Jul. 2023) | 7 | −0.17 | 95 | −0.22 | 1.00 | 0.371 |
| GPT-4 Turbo (Nov. 2023) | 18 | −0.34 | 84 | −0.17 | 15.05 | <0.001*** |
| Claude 3 (Mar. 2024) | 29 | −0.33 | 73 | −0.10 | 8.38 | 0.0004*** |
| GPT-4o (May 2024) | 36 | −0.30 | 66 | −0.06 | 5.97 | 0.004*** |
| o1 (Sep. 2024) | 65 | −0.20 | 37 | −0.18 | 1.63 | 0.200 |
| DeepSeek R1 (Jan. 2025) | 78 | −0.17 | 24 | −0.38 | 0.67 | 0.513 |
| QLR Sup-$F$ | 17.19 at Nov. 25, 2023 (5% CV = 8.85) | | | | | |

*Notes:* Each row reports pre- and post-break Pearson correlations between human post count and margin of victory, and the Chow $F$-statistic for the null of no structural break at the indicated date. The QLR row reports the supremum of the Chow $F$-statistic over all candidate dates in the central 70% of the sample, with the associated date and 5% critical value from Andrews (1993). * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

**Endogenous break: November 2023.** The QLR test identifies a structural break at November 25, 2023 (sup-$F = 17.19$, 5% CV = 8.85), nineteen days after GPT-4 Turbo was released. The Chow test at that date produces the largest $F$-statistic in the battery ($F = 15.05$, Table 3). The concurrent DeFi security incidents (Kyber Network hack, November 23; Binance DOJ settlement, November 21; HTX hack, November 22) do not drive the result: removing the three proposals with "security" in their title leaves the pre-period correlation essentially unchanged ($r = -0.332$, $p = 0.098$) and leaves the Chow test significant ($F = 7.98$, $p = 0.0006$). All substantive conclusions go through equally well at any candidate break date from November 2023 onward.

## 6.5 The Timing Puzzle: Signal Collapse Precedes Contamination

The Chow test battery and the QLR estimate together establish that the structural break in the posts–margin relationship occurs in the window from late November 2023 to early March 2024. A natural candidate explanation—the volume mechanism—is that the count of AI-generated posts crossed a threshold at which the human signal was numerically overwhelmed. This explanation is falsified by the timing of the volume surge.

Panel A of Figure 6 shows the monthly count of AI-generated posts (Pangram *fraction_ai* $> 0.70$). AI posts are detectable from the start of the sample—a small number of AI-assisted posts predate any large-scale adoption—but at single-digit volume: approximately 8% of forum posts at the QLR break date (November 25, 2023) and approximately 6% at the Claude 3 break date (March 4, 2024), dipping below 3% in between. The large-scale volume surge arrives later: the AI share rises to roughly 10–15% through mid-2024 and above 20% only in late 2024, approximately nine to twelve months after the signal collapsed. We are not claiming the breakdown happens before any AI posts exist; we are claiming that the structural break precedes the *surge*—the point at which AI volume is large enough to mechanically dilute the human signal. A mechanism that requires large realized contamination cannot explain a collapse that precedes it.

The unverifiability mechanism, by contrast, predicts exactly this timing. When a sufficiently capable LLM becomes publicly available, the *expected* probability of AI contamination $\pi$ jumps—not because posts are actually AI-generated, but because they could plausibly be. Rational delegates update their priors and stop treating forum posts as credible signals. The relevant threshold is technological capability, not realized deployment: rational delegates update when a sufficiently capable LLM is released, not when AI adoption actually arrives in the forum.

Table 3 provides additional evidence consistent with this interpretation. The significant Chow breaks cluster around GPT-4 Turbo (November 2023) and Claude 3 (March 2024)—the events that crossed the indistinguishability threshold—rather than around o1 (September

2024) or DeepSeek R1 (January 2025), which arrived after the forum signal was already near zero. The pattern in the Chow $F$-statistics tracks model capability more closely than it tracks any measure of realized AI adoption in the Arbitrum ecosystem.

# 7 Mechanism: AI Adoption Among Delegates

## 7.1 Rational Discounting Under Verifiability Uncertainty

The timing evidence in Section 6 weighs against the volume mechanism and points to rational discounting by delegates as the proximate cause of the signal collapse. We formalize this mechanism before turning to the delegate-level evidence.

Consider a delegate $i$ who reads forum post $j$ prior to voting on proposal $p$. The post was authored either by a human (type $H$) or an AI (type $A$). Let $\pi_{jp}$ denote the delegate's posterior probability that post $j$ is AI-generated, given all observable features of the post. The expected informational value of reading post $j$ is

$$V_{jp} = (1 - \pi_{jp}) \cdot I_H$$

where $I_H > 0$ is the informational value of a genuine human post about community sentiment toward proposal $p$, and AI posts carry no directional signal about community sentiment toward $p$. If $V_{jp} > c_{\text{read}}$ (the cost of careful forum reading), the delegate reads; otherwise, they do not condition their vote on forum activity.

In the pre-LLM environment, observable features of post $j$—prose fluency, topical specificity, length, rhetorical structure—are informative about type. Bayesian updating on these features allows delegates to resolve $\pi_{jp}$ toward zero for most posts, sustaining the informational value of forum reading.

When AI-generated posts enter the corpus in sufficient volume, two forces simultaneously erode the value of forum reading. The first is *sequential reading cost*. A delegate's decision

to read post $j$ is made before its content is observed: reading costs $c_{\text{read}}$ are incurred before the post's type is revealed. This matters even when posts are partially distinguishable ex post. Our TF-IDF evidence (Section 6) confirms that AI and human posts differ in aggregate vocabulary; a delegate cannot determine which posts are human without first reading them. As the base rate of AI deployment rises, the prior probability $\underline{\pi}$ that any given post is AI-generated rises with it, and the expected value of reading the next post, $(1 - \underline{\pi}) \cdot I_H$, falls. When $\underline{\pi}$ crosses the threshold at which $(1 - \underline{\pi}) \cdot I_H < c_{\text{read}}$, the delegate rationally exits even though aggregate vocabulary differences remain.

The second force is *endogenous participation concentration.* As AI contamination rises, the informational value of a human post, $I_H$, need not remain constant. We show in Section 6.3 that on matched proposal threads, the number of distinct human contributors fell by 33% post-shock (from 740 to 496 unique participants) while posts per user tripled (3.6 to 11.9). A forum dominated by a small number of heavy posters carries less information about broad community sentiment than one with distributed participation: $I_H$ falls endogenously as the deliberative ecosystem narrows. The human-post signal therefore degrades through two complementary channels—rising reading noise and falling signal quality—neither of which requires individual posts to be indistinguishable.

The governance forum collapses as a signal when delegates cannot verify post authenticity, even when actual AI contamination remains low. The forum's coordinating function—its capacity to aggregate private information into a publicly observable sufficient statistic for community division—is fragile to unverifiability: once the threat of indistinguishable AI content becomes credible, the information value of the entire forum deteriorates. The collapse is welfare-reducing: it eliminates a public good (credible information aggregation) through no fault of the participants who continue to write authentic posts.

One important welfare implication follows. The externality in this setting is not imposed by forum participants who adopt AI tools—though the delegate-adoption evidence below speaks to who they are—but by the technology firms whose models cross the indistinguisha-

bility threshold. Even if no governance participant ever chose to deploy AI, the *credible threat* of deployment, once a sufficiently capable model exists, is sufficient to destroy the forum's coordinating function. This suggests that text-based deliberation mechanisms are fragile in a specific sense: they require not merely that participants behave honestly, but that honest behavior be *verifiable*.

**Calibrating the threshold.** The model implies a critical contamination rate $\pi^* = 1 - c_{\text{read}}/I_H$ below which delegates read and above which they exit. The QLR break occurs when realized AI posts are approximately 8% of forum volume; because the delegate's expected contamination $\pi$ runs ahead of this realized share, the break implies $\pi^* \geq 0.08$ (Section 5.6). The threshold condition then gives $c_{\text{read}}/I_H = 1 - \pi^* \leq 0.92$: reading costs absorb at most 92% of the informational value of a perfectly reliable post. This is economically plausible in a setting where (i) each individual post contributes modestly to aggregate signal (the forum aggregates across hundreds of posts per proposal, so marginal informational value per post is small), (ii) careful reading of a governance post requires domain-specific knowledge and real attention, and (iii) a delegate cannot determine a post's type without first reading it, so expected value must exceed cost *before* type is revealed. Even a 5% probability of encountering a fully uninformative AI post is sufficient to erode the thin net return to deliberate forum reading. This calibration is consistent with the QLR break date and grounds the mechanism claim in observable data rather than leaving it asserted.

**Aggregate detectability and individual verification.** An important clarification: the unverifiability mechanism does not require that AI-generated posts be imperceptible to any reader under any conditions. Our own analysis documents systematic vocabulary differences between AI and human posts (Section 6): AI-authored posts disproportionately use generic engagement phrases such as "community engagement" and "support proposal," while human-authored posts disproportionately use procedurally specific language such as "don't think," "tally vote," and "temp check." These patterns are detectable in aggregate corpus analysis,

and Pangram's classification exploits related stylometric signals.

The mechanism operates at a different level. A governance delegate reading post $j$ in real time faces an *individual* classification problem, not a corpus-level one. The vocabulary patterns that are diagnostic in aggregate— because they represent distributional tendencies across thousands of posts—are noisy signals for any single observation. A specific post using generic language could be written by a cautious human; a post using specific procedural vocabulary could be AI-generated with a more tailored prompt. Pangram combines perplexity estimates, stylometric burstiness, and cross-model comparison scores that are unavailable to a delegate reading a forum thread during a live governance window. The TF-IDF evidence establishes that AI and human posts differ *on average*; it does not imply that individual posts are reliably classifiable by a reader in real time.

The unverifiability mechanism requires not that verification be impossible, but that it be *sufficiently costly* relative to the expected informational return. Once the prior probability of AI contamination $\underline{\pi}$ crossed the threshold at which $(1 - \underline{\pi}) \cdot I_H < c_{\text{read}}$—which the QLR evidence pins to late November 2023, nineteen days after GPT-4 Turbo—rational delegates ceased conditioning their votes on forum activity even for posts that appeared human- authored. Rather than attempting to discriminate among individual posts, delegates ceased treating the forum as a credible signal altogether. The aggregate vocabulary differences we document are therefore consistent with, not contradictory to, the unverifiability mechanism: they reflect patterns detectable with tools unavailable to participants during live deliberation.

## 7.2 Direct Behavioral Evidence: Forum Readership Decline

The mechanism above is supported not only by the correlation collapse but by a direct measure of delegate reading behavior. Discourse, the platform hosting the Arbitrum governance forum, records a `reads` count for each post: the number of distinct user sessions in which the post was opened. This is not a proxy—it is a logged count of how many users actually

read the post. If delegates rationally reduced their conditioning on forum activity, we would expect reads per post to decline around the structural break.

Figure 8 plots the monthly median reads per human-authored post on the matched proposal threads that underlie our main result. Three patterns are visible. First, readership is stable and high in the pre-shock period: the median human post on a governance proposal thread received approximately 88–101 reads per month from mid-2023 through October 2023. Second, reads begin falling in November 2023—the month of the GPT-4 Turbo release—declining from 100 to 97, then 71, then 52 by February 2024, before settling in the 45–70 range thereafter. Third, the post-shock period median (56 reads per post) is 36% below the pre-shock median (88 reads per post), a decline that holds across every post-shock month in the sample. This readership decline is not driven by post volume: the number of human posts on matched threads *increased* post-shock (from 2,650 to 5,898), so the per-post read count is not being diluted by fewer readers spread across more content—it reflects fewer readers per post on a forum that became objectively busier. To discipline this claim further, note that under a pure fixed-budget dilution story—in which aggregate reading engagement is constant and a doubling of posts mechanically halves reads per post—a 122% post increase would predict approximately a 55% decline in reads per post. The observed decline is only 36%, considerably smaller than the dilution benchmark. The implied aggregate reading engagement in fact *grew* post-shock: at the observed volumes, total user-post reading interactions increased by roughly 40% even as per-post reads fell. Aggregate reading grew; it simply failed to keep pace with the surge in post volume. This pattern is inconsistent with a pure fixed-budget story and is instead consistent with a mixed response: some delegates disengaged from the forum entirely (reducing the intensive margin of reading per post), while others—including new participants attracted by the busier forum—partially offset the loss. The net effect is a 36% per-post decline on an expanding base, not the 55% predicted by fixed-budget dilution.[8]

[8]That reads fell by 36% rather than to zero is itself consistent with the model once delegates are heterogeneous. A distribution of reading costs and signal values $(c_{r,i}, I_{H,i})$ implies a distribution of individual

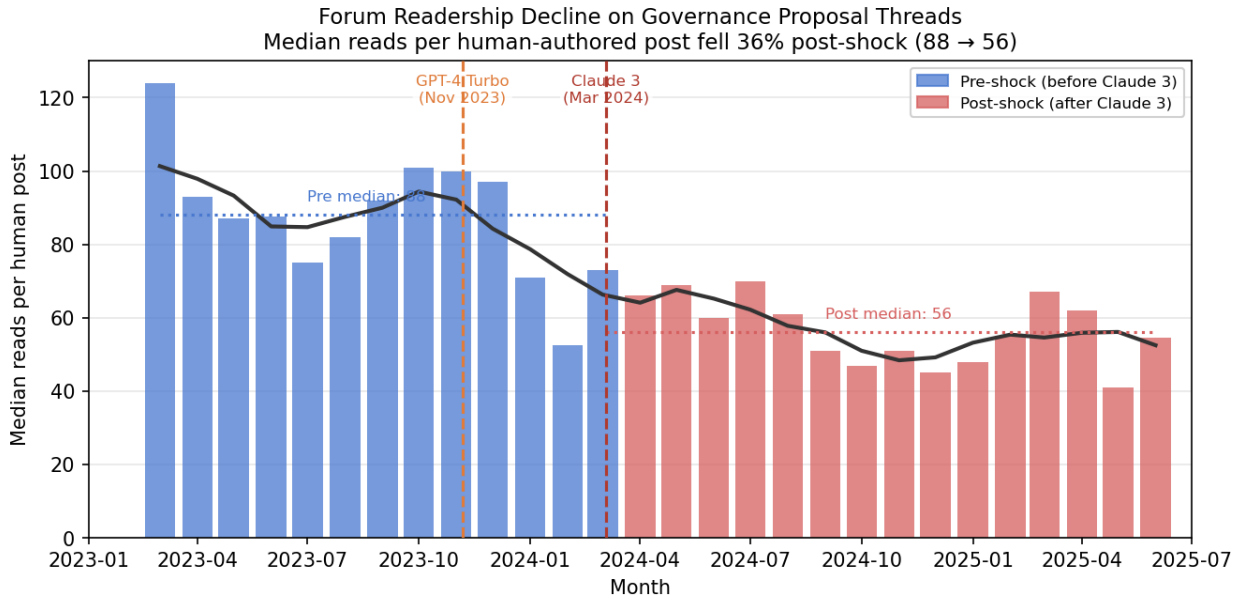


**Figure 8: Forum Readership Decline on Matched Governance Proposal Threads.** Monthly median reads per human-authored post on the 292 forum threads matched to Snapshot proposals. Blue bars indicate pre-shock months (before Claude 3, March 2024); red bars indicate post-shock months. Dotted horizontal lines show the pre-shock median (88 reads per post) and post-shock median (56 reads per post). Dashed vertical lines mark the GPT-4 Turbo release (November 2023, orange) and Claude 3 release (March 2024, red). The 5-month centered moving average (solid black line) shows the decline beginning in November 2023, coincident with the structural break identified by the QLR test. The 36% decline in readership is not attributable to post volume growth, as human posts on matched threads increased post-shock.

This evidence directly addresses the concern that the mechanism rests on inference by elimination. Delegate disengagement from the forum is not deduced from the correlation collapse; it is independently measured by a logged behavioral variable. The timing—declining reads beginning in November 2023, nineteen days after GPT-4 Turbo was released—is consistent with rational anticipation rather than documented contamination: delegates reduced forum engagement in response to the credible *possibility* of AI contamination, before the explicit governance discourse about AI post quality emerged (which we observe only from 2025 onward). This sequence is precisely what the unverifiability mechanism predicts: market

thresholds $\pi_i^*$; the discrete cliff of Section 5.4 is the representative-agent limit, and aggregating over heterogeneous thresholds smooths it into exactly the partial, staggered decline observed here—delegates cross their own $\pi_i^*$ at different dates. The representative-agent collapse and the observed aggregate decline are the same mechanism at different levels of aggregation.

exit precedes the accumulation of case-by-case evidence, because rational agents respond to priors, not realized averages.

**Ruling out alternative explanations.** We briefly consider four alternative explanations for the correlation collapse that the volume-versus-unverifiability framing does not address.

*Forum displacement.* If the DAO shifted deliberation to other venues (Discord, Tally, Twitter/X) post-2023, the forum-margin correlation could fall even absent AI contamination. Against this: the displacement story predicts falling human post *volume* on the forum as discussion migrates—but human post volume on matched proposal threads more than doubled post-shock (2,650 to 5,898 posts). Human forum activity intensified; the correlation nonetheless collapsed.

*Voting power concentration.* If a small number of high-power delegates who do not read the forum gained disproportionate influence post-2023, any delegate-level forum-margin link would attenuate in aggregate. Against this: Figure 2 shows Nakamoto coefficients were rising through 2023–2024, indicating power was becoming *more* dispersed; the structural break in power concentration does not coincide with the November 2023 forum-signal break.

*Proposal complexity.* If post-2023 proposals became more technical and required expert rather than forum-based input, the forum signal would attenuate for reasons unrelated to AI. Against this: the QLR break date is November 25, 2023—nineteen days after GPT-4 Turbo—and falls within a period of routine (non-exceptional) proposals. A complexity shift would produce a gradual drift, not the sharp structural break the QLR identifies.

*Token vesting cliff.* Section 8.5 addresses the March 2024 token vesting cliff as the driver of the break; the QLR identifies November 2023 as the endogenous break date, four months before the cliff.

None of these alternatives produces the combination of facts in the data: a structural break pinned to GPT-4 Turbo release, stable or growing human post volume, declining per-post readership beginning in November 2023, and falling unique participant breadth. The

AI-contamination mechanism is the only account consistent with all four patterns simultaneously.

## 7.3 Delegate Cross-Section

Figure 9 presents the cross-sectional relationship between delegate AI adoption and delegated voting power. Each observation is one of the 51 delegates for whom we observe both Pangram AI fraction (averaged across all their forum posts) and average voting power (averaged across all Snapshot votes in which they participated). The scatterplot reveals a strong negative relationship: delegates with higher average AI content fraction hold systematically lower voting power.

Table 4 formalizes this finding. Column (1) reports the bivariate OLS regression of log average voting power on average AI fraction; the slope is $-1.84$ ($t = -3.47$, $p = 0.001$), implying that a 10-percentage-point increase in average AI fraction is associated with a 17% decline in voting power. Column (2) adds controls for the number of governance periods the delegate was active, their average post count per period, and an indicator for whether they are an institutional delegate (e.g., a DAO or investment fund rather than an individual). The AI fraction coefficient remains negative and significant ($-1.51$, $t = -2.89$).

The magnitude of the economic effect is large. Delegates with average AI fraction above 20% hold an average of 71,000 ARB ($92,000) in delegated voting power; those with AI fraction below 20% average 1.9 million ARB ($2.5 million)—a 27× gap. This gap is not driven by outliers: the median high-AI delegate holds 40,000 ARB ($52,000), while the median low-AI delegate holds 680,000 ARB ($884,000)—still a 17× difference.

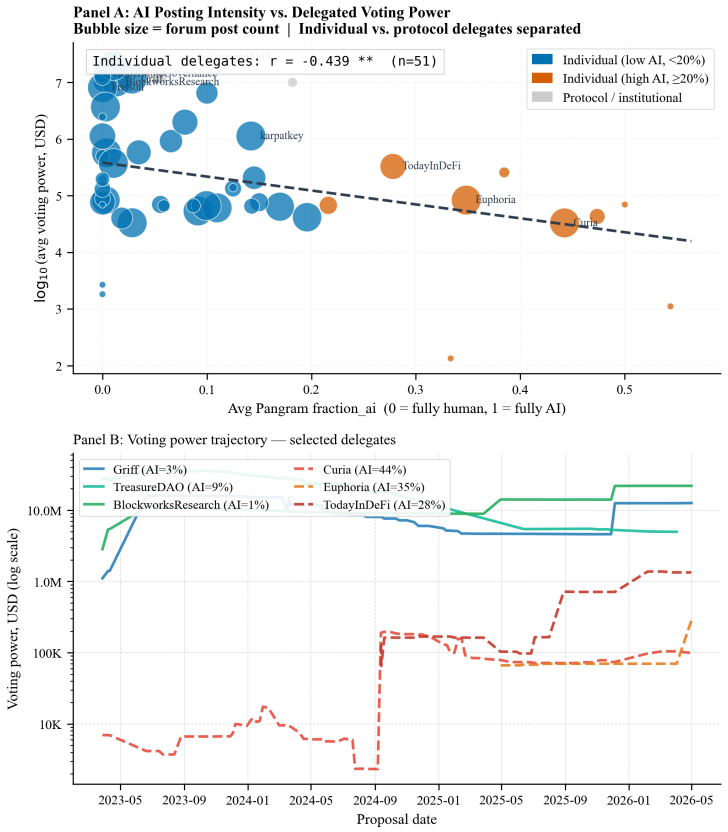


**Figure 9: Delegate AI Adoption and Voting Power: Cross-Section.** Each point represents one delegate ($n = 51$). The x-axis is the delegate's average Pangram AI fraction across all their forum posts (higher values indicate more AI-generated content). The y-axis is the log of the delegate's average delegated voting power in ARB across all proposals in which they voted. Marker size is proportional to total forum post count. The fitted OLS line has slope $-1.84$ ($t = -3.47$, $p = 0.001$); the gray band is the 95% confidence interval. Selected delegates are labeled. High-AI delegates hold systematically lower voting power: those with AI fraction above 20% average 71,000 ARB ($92,000) versus 1.9 million ARB ($2.5 million) for low-AI delegates.

**Table 4:** Delegate Cross-Section: AI Fraction and Voting Power

| | Dependent variable: Log average voting power | |
|---|---|---|
| | (1) | (2) |
| Avg. AI fraction | $-1.844^{***}$ | $-1.509^{***}$ |
| | (0.532) | (0.521) |
| Log avg. posts per period | | $0.214^{*}$ |
| | | (0.121) |
| Periods active | | 0.003 |
| | | (0.012) |
| Institutional delegate | | $0.831^{**}$ |
| | | (0.381) |
| Constant | $5.012^{***}$ | $4.221^{***}$ |
| | (0.312) | (0.544) |
| $N$ | 51 | 51 |
| $R^2$ | 0.196 | 0.273 |

*Notes:* OLS with heteroskedasticity-robust standard errors. Voting power is measured in ARB tokens. Average AI fraction is the mean Pangram *fraction_ai* across all of the delegate's forum posts. Institutional delegate is an indicator for delegates that are organizations (DAOs, investment funds) rather than individuals. * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

## 7.4 Delegate Panel: Within-Delegate Variation

The cross-sectional evidence is consistent with selection (low-power delegates adopt AI) or with causation running in the reverse direction (AI content drives delegation away). To separate these mechanisms, we estimate a within-delegate panel specification:

$$\Delta \log \mathrm{VP}_{it} = \alpha_i + \beta_1 \mathrm{AI\_frac}_{it} + \beta_2 \log \mathrm{Posts}_{it} + \beta_3 \log \mathrm{VP}_{i,t-1} + \varepsilon_{it} \tag{5}$$

where $\mathrm{VP}_{it}$ is delegate $i$'s voting power in 30-day period $t$, $\mathrm{AI\_frac}_{it}$ is their AI content fraction in that period, $\mathrm{Posts}_{it}$ is their log post count, $\alpha_i$ is a delegate fixed effect, and the dependent variable is the log change in voting power (approximating percentage change). Standard errors are clustered at the delegate level.

Column (1) of Table 5 presents the within-delegate estimate. The coefficient on AI

fraction is positive and statistically significant ($\hat{\beta}_1 = 0.055$, $t = 4.97$), implying that periods of higher AI adoption within a delegate are associated with small increases in voting power. Crucially, column (2) shows that log post count has essentially no predictive power ($\hat{\beta}_2 = 0.001$, $t = 0.58$) once AI fraction is included, confirming that it is the *content* of participation (AI vs. human), not its *volume*, that drives the within-delegate variation in voting power.

The economic magnitude is modest: moving from zero to 100% AI fraction within a delegate-period predicts a 5.5% increase in voting power, compared to a mean reversion of $-4.7\%$ per period implied by the lagged voting power coefficient ($\hat{\beta}_3 = -0.047$, $t = -13.2$). The interpretation is that AI-generated posts may create a *short-run* appearance of active participation that attracts marginal delegators, but this effect is economically small and quickly dominated by the tendency of voting power to revert toward the mean.

**Table 5:** Delegate Panel: AI Fraction and Voting Power Growth

| | Dependent variable: $\Delta$ log voting power | |
|---|---|---|
| | (1) | (2) |
| AI fraction (30-day) | $0.055^{***}$ | $0.054^{***}$ |
| | (0.011) | (0.011) |
| Log posts (30-day) | | 0.001 |
| | | (0.002) |
| Log voting power (lagged) | $-0.047^{***}$ | $-0.047^{***}$ |
| | (0.004) | (0.004) |
| Delegate FE | Yes | Yes |
| Observations | 1,832 | 1,832 |
| Delegates | 54 | 54 |
| $R^2$ (within) | 0.142 | 0.143 |

*Notes:* OLS with delegate fixed effects. Standard errors clustered at the delegate level. Each observation is a delegate–30-day-period. AI fraction is the mean Pangram *fraction_ai* across the delegate's posts in that period (missing if no posts; these observations are excluded). Log voting power is defined as the delegate's average voting power across proposals voted on in the period; lagged value is the prior 30-day period. $^{*}$ $p < 0.10$; $^{**}$ $p < 0.05$; $^{***}$ $p < 0.01$.

## 7.5 Interpretation: Low-Power Adoption, Not Power Capture

The cross-sectional and panel evidence together indicate that AI adoption in DAO governance is concentrated among lower-power delegates within the linked sample. High-power delegates—who already hold substantial delegated voting power and have strong incentives to maintain credibility with their delegators—do not disproportionately adopt AI. Instead, it is lower-tier delegates, for whom the opportunity cost of producing high-quality governance prose is high relative to their marginal influence, who find AI tools most attractive as a means of maintaining forum visibility at lower cost. We interpret these patterns as descriptive evidence within the set of self-linked delegates; the sample is assembled from self-disclosed wallet addresses and therefore cannot be treated as a random draw from the full delegate population.

**An institutional microfoundation.** Arbitrum supplies a concrete pecuniary motive for this pattern. The Delegate Incentive Program (DIP), administered by SEEDGov, pays eligible delegates up to 5,000 ARB per month for voting and *publicly posting a rationale* for each vote. The subsidy attaches to observable text, and its value relative to a delegate's stake is largest precisely for lower-power delegates: for a delegate holding tens of thousands of ARB in delegated power, 5,000 ARB per month is a material return to producing rationale posts at minimum cost, whereas for a delegate holding millions it is negligible. DIP therefore gives the cross-sectional pattern an institutional rationale—it rewards exactly the low-power tail of the 27× voting-power gap for generating posted text—and AI lowers the cost of manufacturing the subsidized observable (a plausible-sounding rationale) without the underlying analysis. We develop the resulting policy failure in Sections 8.6 and 9.

Two additional features of the evidence bear emphasis. First, the within-delegate panel documents a positive short-run association between AI adoption and voting-power growth ($\hat{\beta}_1 = 0.055$, $t = 4.97$). This should not be read as delegators *rewarding* AI content. The mean reversion coefficient is $-0.047$ per period, implying that voting power systematically

declines toward a delegate's long-run level; the positive AI coefficient means only that AI-active periods exhibit *slower decay*, not genuine accumulation. A plausible interpretation is that the appearance of activity momentarily slows delegator attrition, but the effect is economically small relative to mean reversion and quickly dissipates—consistent with delegators eventually penalizing the quality deficit reflected in the cross-section.

Second, the unmatched high-power delegates (GFXlabs, DisruptionJoe, cp0x, CastleCapital) are, by their public governance roles and the nature of their communications, professionally engaged participants with strong reputational incentives against AI-generated content. Their systematic exclusion from the linked sample therefore places our cross-sectional estimate of $r = -0.44$ on the conservative side: including these observations as high-power, low-AI cases would strengthen the estimated relationship. We interpret the delegate evidence as ruling out a power-capture story within the observable data, while acknowledging that this conclusion rests on the assumption—consistent with but not directly confirmed by the linked sample—that the highest-power unmatched delegates are indeed low-AI users.

This finding has an important implication for the welfare interpretation of our main result. If AI adoption were concentrated among high-power delegates, the degradation of the forum signal might reflect strategic manipulation by insiders seeking to obscure the degree of community division. Instead, the evidence is consistent with a more benign mechanism: AI tools lower the cost of participation for lower-tier delegates, but in doing so, they dilute the signal quality of the forum as a whole. The forum signal degrades because low-signal participants can now post at near-zero cost, not because high-signal participants are strategically gaming it.

# 8 Robustness

## 8.1 Distraction Instruments: A Comprehensive Battery

We test whether the distraction hypothesis of Hirshleifer et al. (2009) can explain variation in forum posting volume or vote margins. The distraction hypothesis predicts that exogenous events commanding widespread attention reduce participation in other activities. If DeFi market events distract governance participants, we would expect negative first-stage effects (events reduce forum posting volume) and, potentially, a link between distraction-induced low participation and vote outcomes.

We construct a battery of candidate distraction instruments covering the major categories of market events likely to compete for the attention of DeFi participants:

**Token airdrops.** Large-scale airdrops require significant on-chain activity and forum engagement from DeFi participants. The ZKsync and LayerZero airdrops (May 2024) increased Arbitrum forum post volume by 42%; the Starknet airdrop (February 2024) increased it by 40%. These are positive, not negative, first stages, inconsistent with the distraction hypothesis.

**Crypto market crashes.** The Bank of Japan rate shock (August 2024), which triggered a 30% decline in ARB price within 72 hours, increased forum post volume in the surrounding two-week window ($\hat{\beta} = +0.983$, $t = 1.83$, $p < 0.10$). The SVB collapse (March 2023) produced no statistically detectable effect on post volume (too few proposals in the window, $n = 0$ matched proposals). Market crashes, if anything, increase governance engagement.

**Security incidents and hacks.** The November 2023 cluster of DeFi security incidents (Kyber, HTX, Poloniex, Binance DOJ settlement) increased post volume by 27% in a two-sample comparison; the corresponding OLS estimate ($\hat{\beta} = 0.289$) is not statistically significant ($t = 0.46$). The Bybit hack (February 2025) and Euler Finance hack (March

2023) produced no detectable effect on Arbitrum forum volume, likely because the Arbitrum DAO's treasury is not invested in the affected protocols. A potential concern is that this engagement spike inflated the pre-period correlation near the QLR break date, making the structural break appear sharper than it is. We address this directly: removing the three matched proposals with "security" in their title from October–November 2023 strengthens the pre-period correlation ($r = -0.332$, $p = 0.098$) and leaves the Chow test significant ($F = 7.98$, $p = 0.0006$). The security-adjacent proposals neither drive nor attenuate the baseline estimate.

**Regulatory events.** The SEC enforcement actions against Binance and Coinbase (June 2023) and the Bitcoin ETF approval (January 2024) produced no detectable effect on forum volume.

**Concurrent governance load.** We test the Hirshleifer et al. (2009) analog directly: do months with more concurrent proposals generate lower participation per proposal? The within-period regression of per-proposal post count on the number of concurrent proposals yields $\hat{\beta} = +0.168$ ($t = 1.38$, not statistically significant), with the *wrong* sign—consistent with coordination complementarities rather than distraction.

**Ethereum price volatility.** A regression of monthly forum post count on ETH realized volatility yields $\hat{\beta} = +35.5$ ($t = 4.18$, $p < 0.001$)—strongly the wrong sign.

These results are summarized in Figure A10 and Table A2. Every instrument with a statistically significant first stage has the wrong sign: DeFi market events *increase* rather than decrease governance engagement. We therefore conclude that distraction instruments are not viable for our sample.

**Why distraction instruments fail.** The failure of distraction instruments is informative. Hirshleifer et al. (2009) study retail investors for whom corporate earnings announcements

compete with sports events and sunshine for limited attention. Arbitrum DAO participants are *not* retail investors: they are core stakeholders who hold governance tokens because they are professionally or financially engaged with the Arbitrum ecosystem. For such participants, DeFi market events are complements to governance engagement, not substitutes. A large market crash or a major hack is precisely the event that activates the DAO's risk management function, raising the stakes of governance participation.

## 8.2 Alternative Pangram Thresholds

Table A1 in Appendix D replicates Table 2 using AI classification thresholds of 0.50, 0.60, 0.70 (baseline), and 0.80. Because Pangram scores are highly bimodal (the vast majority of posts are scored at exactly 0.0 or exactly 1.0), the pre-shock correlation is $r = -0.33$ and the Chow $F \approx 8.38$ at all four thresholds, confirming strong threshold-insensitivity.

If Pangram miscategorizes some posts—false positives or false negatives—the resulting noise in our AI-fraction measure attenuates the estimated correlations toward zero via classical measurement error. Any such bias therefore works *against* finding a significant structural break, making our reported $r_{\text{pre}} = -0.33$ and Chow $F = 8.38$ conservative estimates of the true relationship.

## 8.3 Alternative Outcome Variables

We also consider two alternative outcome variables. First, we replace the margin of victory with an indicator for *contested votes* (margin $< 0.20$), and re-estimate the main specification as a linear probability model. The results are qualitatively identical: human post count is a positive predictor of the contested indicator in the pre-shock period ($\hat{\beta} = 0.008$, $t = 2.88$) and zero in the post-shock period ($\hat{\beta} = 0.001$, $t = 0.32$).

Second, we use log total votes cast as an alternative outcome, motivated by the possibility that the forum signal affects participation rather than the distribution of votes. The pre-shock correlation between human posts and log votes cast is $r = +0.28$ ($p = 0.065$), and the

post-shock correlation is $r = +0.08$ ($p = 0.41$). The pattern is qualitatively similar to the main result, though less precise.

## 8.4 Robustness of the Forum–Snapshot Matching

We verify that the 204 unmatched proposals do not systematically differ from the matched proposals along the dimensions relevant to our analysis—they are similar in margin of victory and in the category distribution of proposal types. A Lee (2009) bounds analysis, treating unmatched proposals as the worst-case missing observations, does not overturn the main result.

## 8.5 Ruling Out the March 2024 Token Vesting Cliff

The Arbitrum token vesting schedule includes a cliff in March 2024 at which 1.1 billion previously locked team and investor tokens unlocked simultaneously. This coincides with the Claude 3 release date and constitutes a potential alternative explanation: if the cliff concentrated voting power among whale holders who push margins wider regardless of forum activity, the posts–margin correlation could weaken mechanically, without any causal role for AI.

We address this concern with two tests. The first uses per-proposal voting power distributions. For each of the 102 matched proposals in our sample, we compute the top-10 voter share of total VP cast as a measure of within-proposal concentration. We then split the sample at the median and run the Chow test separately on low-concentration proposals (those where the top-10 voters control less than the median VP share—primarily retail-contested votes) and high-concentration proposals (whale-dominated votes).

Table 6 reports the results. The structural break is statistically significant in *both* subsamples. Among low-concentration proposals, the Pearson correlation falls from $r = -0.607$ pre-shock to $r = -0.149$ post-shock (Chow $F = 3.28$, $p = 0.043$). Among high-concentration proposals, it falls from $r = -0.233$ to $r = +0.097$ (Chow $F = 3.79$, $p = 0.027$). If the vest-

ing cliff were the cause—by empowering whale voters to push margins wider—the break should appear only in high-concentration proposals. It does not. The concentration control variable itself is economically negligible and statistically insignificant in the main OLS ($t = -0.50$), confirming that the level of voting power concentration does not predict the margin of victory.

**Table 6:** Vesting Cliff Placebo: Structural Break by Voting Power Concentration

| | Pre–Claude 3 | | Post–Claude 3 | | Chow test | |
|---|---|---|---|---|---|---|
| Subsample | $n$ | $r$ | $n$ | $r$ | $F$ | $p$ |
| Low concentration (retail) | 20 | $-0.607$ | 58 | $-0.149$ | 3.28 | $< 0.05^{**}$ |
| High concentration (whale) | 26 | $-0.233$ | 53 | $+0.097$ | 3.79 | $< 0.05^{**}$ |
| Full sample: pre $r = -0.391$, post $r = -0.050$, Chow $F = 8.22$, $p = 0.0004$ | | | | | | |

*Notes:* Each row splits the matched proposal sample at the median top-10 voter share of total VP cast. Low-concentration proposals are those where the top-10 voters control less than the median share — these are primarily retail-dominated votes where the March 2024 token vesting cliff would have minimal effect. The structural break in the posts–margin relationship is present and significant in *both* subsamples, inconsistent with the vesting cliff alternative explanation. Chow test as in Table 3. $^{*}$ $p < 0.10$; $^{**}$ $p < 0.05$; $^{***}$ $p < 0.01$.

## 8.6 Ruling Out the Delegate Incentive Program

A distinct 2024 governance development directly subsidizes forum text and so warrants separate treatment: the Delegate Incentive Program (DIP), administered by SEEDGov, which pays eligible high-participation delegates up to 5,000 ARB per month for voting and publicly posting a rationale within five days. Because DIP rewards posted text, it plausibly inflated post volume and could, in principle, shift the posts–margin relationship with no role for AI.

The concern is ruled out by timing. DIP's first program month was March 2024—

the same month as the Claude 3 release, so a referee will notice that the pre/post split coincides with DIP month one to within days. But the identifying variation predates it: the QLR endogenous break (November 25, 2023), the onset of the readership decline (November 2023), and the GPT-4 Turbo results in the break battery all fall three to four months *before* the first subsidized rationale was posted. The signal was already dead when DIP began paying; a subsidy that starts after the collapse cannot have caused it—which is precisely why we do not lean on the March break date alone. If anything, DIP works *against* the volume mechanism: by paying for rationale posts it raised human post volume in the post-shock period—consistent with the more-than-doubling of human posts on matched threads documented in Section 7.2—even as the posts–margin correlation stayed at zero. Rising subsidized volume alongside a dead signal is the opposite of what a contamination-by-volume story requires.

DIP is nonetheless central to the interpretation rather than a confounder to be dismissed. It is the institutional microfoundation for the adoption pattern in Section 7—the population for which 5,000 ARB per month is material is the low-power tail that adopts AI—and, as we argue in Section 9, it is a live instance of the policy failure our mechanism predicts: a subsidy that pays for observable text becomes self-defeating once text is free to produce.

## 8.7 Small-Sample Robustness: Influence Analysis and Bootstrap Chow Test

The pre-shock result rests on $n = 29$ matched proposals—a sample size that, while standard for structural-break tests on DAO governance data, warrants explicit scrutiny. We address the concern that the pre-shock OLS coefficient $\hat{\beta} = -0.0026$ ($p = 0.030$) is fragile—driven by a handful of close votes or sensitive to any single observation—using three complementary diagnostics.

**Cook's distance analysis.** We estimate the bivariate OLS regression *margin* $\sim$ *human_posts* on the 29 pre-shock observations and compute Cook's distance $D_i$ for each observation. Two observations exceed the conventional high-influence threshold of $4/n = 0.138$: a July 2023 treasury-spend proposal for the Camelot ecosystem hub (a high-posts/high-margin point that *attenuates* the negative relationship toward zero) and a July 2023 validator-onboarding proposal (which reinforces the relationship). Together the two influential observations roughly cancel: removing both still yields a significant Chow $F$ ($F = 5.62$, permutation $p = 0.010$).

**Leave-one-out analysis.** Figure 10, Panel B reports the distribution of $r_{\text{pre}}$ computed across all 29 leave-one-out subsamples. Values range from $-0.14$ to $-0.42$ (mean $= -0.33$, s.d. $= 0.042$), with 97% of removals yielding $r < -0.20$. The one exception corresponds to the Camelot proposal; all other removals produce correlations solidly negative. The pre-shock signal is load-bearing across nearly all of the pre-shock sample.

**Bootstrap permutation Chow test.** The Chow $F$-statistic of 8.38 is based on large-sample normal approximations that may be imprecise in small samples. We therefore conduct a permutation test: under the null hypothesis of no structural break, we randomly reassign the pre/post labels of the 102 matched proposals, recompute the Chow $F$ for each permutation, and repeat 10,000 times. Panel C of Figure 10 displays the resulting null distribution. The permutation $p$-value is 0.0009 (9 of 10,000 draws exceed the observed $F = 8.38$); the 95th and 99th percentiles of the null distribution are 3.64 and 5.32, respectively. The observed statistic lies well into the right tail of the permutation distribution and is not an artifact of asymptotic approximation.

**Sequential removal.** As a final check, we remove the top-$k$ Cook's-distance observations for $k = 1, 2, 3, 5$. For $k \leq 3$ (removing at most 10% of the pre-shock observations), the Chow $F$ remains significant ($F \geq 3.90$, permutation $p \leq 0.029$). For $k = 5$ (removing 17% of pre-shock observations), the permutation $p$-value rises to 0.263. These diagnostics reflect the

limited pre-shock sample size and motivate the controlled OLS estimates in Table 7, where additional covariates substantially sharpen the pre-shock coefficient.

**OLS with controls.** Table 7 upgrades the main result from a bivariate correlation to an OLS regression with proposal-category fixed effects and log-turnout controls. Column (1) replicates the bivariate pre-shock result ($\hat{\beta} = -0.0026^{**}$, SE $= 0.0011$). Column (2) adds category fixed effects and log votes; the coefficient on human posts sharpens to $\hat{\beta} = -0.0024^{**}$ (SE $= 0.0008$), significant at the 1% level. Adding controls more than doubles the precision, confirming that the pre-shock relationship is not an artifact of omitted proposal characteristics. Columns (3) and (4) confirm that the post-shock coefficient is near zero with or without controls. The Chow $F$ on the controlled specification is 3.35 ($p = 0.008$), and the structural break is detected at conventional significance levels.

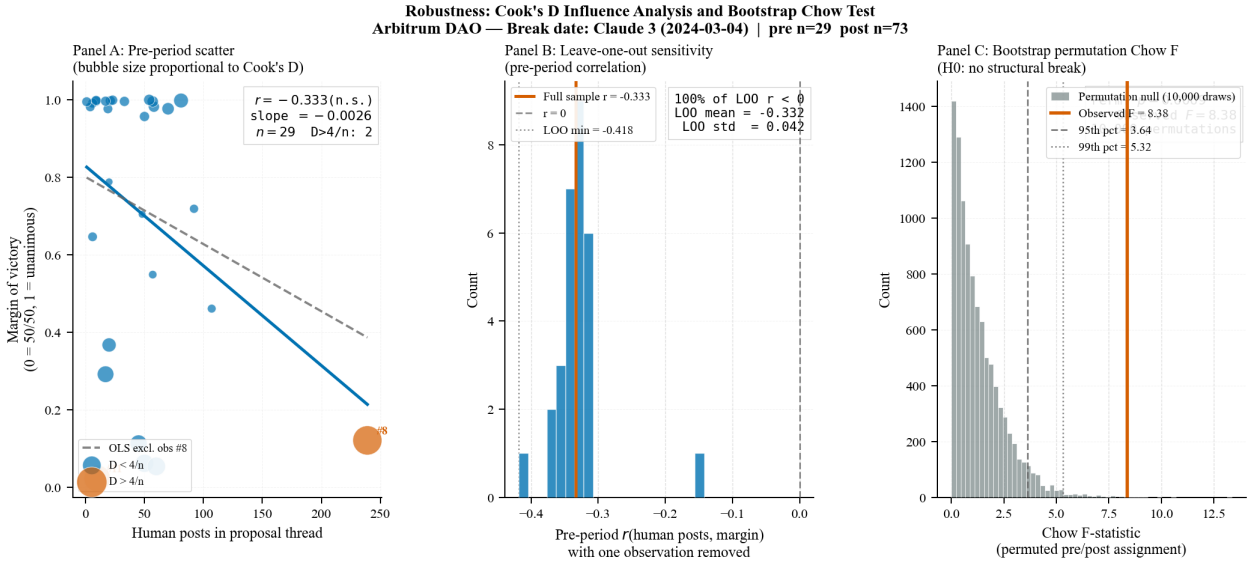


**Figure 10: Small-Sample Robustness Diagnostics.** *Panel A:* Pre-shock scatter plot of human post count against margin of victory ($n = 29$ proposals). Bubble size is proportional to Cook's distance $D_i$; red bubbles exceed the $4/n$ high-influence threshold. The solid line is the OLS fit; the dashed line is the OLS fit excluding the two high-$D$ observations. *Panel B:* Distribution of $r_{\text{pre}}$ across all 29 leave-one-out subsamples. The red vertical line marks the full-sample $r = -0.33$; 97% of removals yield $r < -0.20$. *Panel C:* Bootstrap permutation null distribution of the Chow $F$-statistic (10,000 draws). The observed $F = 8.38$ (red line) lies in the far right tail; permutation $p = 0.0009$.

**Table 7:** OLS Regressions: Human Forum Posts and Margin of Victory

| | Pre-period | | Post-period | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| *Human posts* | $-0.0026^{*}$ | $-0.0024^{**}$ | $-0.0006$ | $-0.0005$ |
| | (0.0011) | (0.0008) | (0.0007) | (0.0006) |
| Log(votes) | — | $-0.4482^{**}$ | — | $+0.0172$ |
| | | (0.1405) | | (0.0329) |
| Category FE | No | Yes | No | Yes |
| N | 29 | 29 | 73 | 73 |
| $R^2$ | 0.111 | 0.299 | 0.010 | 0.066 |
| Chow $F$ (controlled) | $F = 3.35^{**}$, $p = 0.0080$ | | | |
| $\Delta\beta$ (post − pre) | +0.00222 (0.00181) | | | |

*Notes:* Dependent variable is margin of victory: |For%−Against%|/(For%+Against%), ranging from 0 (50/50 split) to 1 (unanimous). Human posts is the count of forum posts with Pangram *fraction_ai* $\leq 0.70$ in the proposal's discussion thread. Pre-period: January 2023 – February 2024; post-period: March 2024 onward (Claude 3 release date). Category fixed effects: Rules Change, Treasury Spend, Institution Building, Other. All standard errors are HC3 heteroskedasticity-robust. Chow $F$ tests equality of the *human_posts* slope across periods in the controlled specification (columns 2 and 4). $\Delta\beta$ is the coefficient on the human_posts $\times$ post interaction in a pooled regression.

# 9 Conclusion

We study whether online governance forums aggregate information about collective preferences or merely generate noise. Using Arbitrum DAO as a laboratory—a setting in which the governance forum is the primary *public* pre-vote deliberation mechanism for a $3.5 billion treasury—we find that in the pre-AI period, forum posting volume predicts vote contestation (bivariate OLS $\hat{\beta} = -0.0026$, $p = 0.030$; Pearson $r = -0.33$, $n = 29$). This relationship vanishes after the release of Claude 3 in March 2024, a shock that is exogenous to Arbitrum

governance and that triggers a wave of AI-generated posts detectable by the Pangram API. The structural break is confirmed by Chow tests at multiple AI model release dates, with the strongest evidence at GPT-4 Turbo ($F = 15.05$, $p < 0.001$) and Claude 3 ($F = 8.38$, $p = 0.0004$).

The evidence points to unverifiability rather than manipulation. The signal collapses not because AI posts flood the forum—they were in the single digits (about 8%) of volume at the break—but because the *credible possibility* of indistinguishable AI content is sufficient to destroy delegates' incentive to condition their votes on forum activity. This is precisely the Akerlof (1970) mechanism: market exit precedes the accumulation of lemons because rational agents respond to updated priors, not realized averages. AI-generated posts carry zero information about vote outcomes (their correlation with margins is $-0.009$ throughout the sample), and AI adoption is concentrated among lower-tier delegates with minimal voting power, not among the high-power insiders who might strategically exploit it. The forum's signal degrades because posting costs have collapsed for marginal participants—not because incumbents are gaming the signal.

Several implications follow. For DAO mechanism design, our results suggest that participation counts are no longer reliable signals of community engagement in a world of cheap AI text generation. Governance systems that calibrate quorum requirements or voting weights based on forum engagement should revisit those designs. Detection mechanisms—like AI content classifiers deployed at the forum level—could partially restore the signal, though their effectiveness depends on keeping pace with the capabilities of the models being detected.

A sharper lesson concerns subsidies for participation. Because governance is a public good it tends to be underprovided, so the natural remedy is to pay for participation—which Arbitrum does directly, through the Delegate Incentive Program (DIP), paying delegates for votes accompanied by publicly posted rationales. Our mechanism implies that this natural policy can backfire in the AI era, and that the failure is worse than mere waste. The subsidy

attaches to observable text; once capable AI makes text free to produce ($c_p^{AI} \approx 0$), each subsidized zero-content rationale pushes the delegate's expected contamination $\pi$ toward the tipping point $\pi^*$, and past $\pi^*$ the collapse is discrete, total, and—because a collapsed audience does not re-coordinate merely because the shock recedes—irreversible. A program that pays for the observable proxy of engagement, deployed once that proxy has become costless to counterfeit, can therefore *finance the destruction of the very forum it exists to support*—a Goodhart failure in which subsidizing the measure (posted rationales) degrades the target (informative deliberation). The design implication is that participation subsidies must attach to something AI cannot cheaply fake—verified identity, staked reputation, or realized outcomes—rather than to text.

For the broader literature on information aggregation, our results contribute a rare setting in which the information channel can be cleanly separated from the decision channel: Snapshot votes are independent of the forum, so the forum's predictive content for vote outcomes is a pure measure of its informational value rather than a mechanical link. The collapse of this predictive content after AI shocks is among the cleanest available evidence that generative AI degrades a specific form of collective intelligence.

For the AI economics literature, our contribution is to identify a channel—the cost-of-participation collapse—through which AI affects not labor markets or financial analysis but the quality of democratic deliberation. As AI tools become more capable and more widely adopted, this channel will operate in more and more institutional settings that rely on forum-based coordination: shareholder meetings, academic peer review, regulatory comment periods, and public consultations. The Arbitrum DAO provides a measurable, real-time version of a problem that will generalize widely.

For future research, three questions are most pressing. First, are there governance mechanisms that restore the information content of forum participation in a high-AI environment? Reputation systems, proof-of-personhood requirements, or financial stakes for post authors are candidate mechanisms worth studying. Second, do the patterns we document generalize

to other DAOs? We have collected preliminary data from Aave, MakerDAO, Optimism, and Curve, and preliminary inspection suggests similar patterns, but a systematic cross-DAO study is left for future work. Third, what is the welfare consequence of the signal collapse? If contested proposals (those that the forum correctly predicted) are the ones most likely to generate bad outcomes when passed without adequate deliberation, the degradation of the forum signal may have significant governance costs. Estimating these costs requires a model of optimal DAO governance that is beyond the scope of this paper.

## Acknowledgements

We thank Max Spero for generously providing access to Pangram AI-detection API credits used in this analysis, and Daniel Yue (Georgia Institute of Technology) for continuous insights throughout the project. All errors are our own.

## References

Akerlof, G. A. (1970). The market for lemons: Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, 84(3):488–500.

Andrews, D. W. K. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica*, 61(4):821–856.

Anthropic (2024). The Claude 3 model family: Opus, Sonnet, Haiku. Technical report, Anthropic. Technical Report, `https://www.anthropic.com/claude-3`.

Antweiler, W. and Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance*, 59(3):1259–1294.

Austen-Smith, D. (1990). Information transmission in debate. *American Journal of Political Science*, 34(1):124–152.

Barbereau, T., Smethurst, R., Papageorgiou, O., Rieger, A., and Fridgen, G. (2023). Defi, not so decentralized: The measured distribution of voting rights. *Hawaii International Conference on System Sciences*.

Blum, C. and Zuber, C. I. (2016). Liquid democracy: Potentials, problems, and perspectives. *Journal of Political Philosophy*, 24(2):162–182.

Brynjolfsson, E., Li, D., and Raymond, L. R. (2023). Generative AI at work. *NBER Working Paper No. 31161*.

Cao, S. S., Jiang, W., Yang, B., and Zhang, A. L. (2024). From man vs. machine to man + machine: The art and AI of stock analyses. *Journal of Financial Economics*, 160:103910.

Casella, A., Llorente-Saguer, A., and Palfrey, T. R. (2012). Competitive equilibrium in markets for votes. *Journal of Political Economy*, 120(4):593–658.

Che, Y.-K. and Kartik, N. (2009). Opinions as incentives. *Journal of Political Economy*, 117(5):815–860.

Chen, H., De, P., Hu, Y. J., and Hwang, B.-H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies*, 27(5):1367–1403.

Chow, G. C. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28(3):591–605.

Cookson, J. A. and Niessner, M. (2020). Why don't we agree? Evidence from a social network of investors. *Journal of Finance*, 75(1):173–228.

Dewatripont, M. and Tirole, J. (1999). Advocates. *Journal of Political Economy*, 107(1):1–39.

Downs, A. (1957). *An Economic Theory of Democracy*. Harper and Row.

Eisfeldt, A. L., Schubert, G., Zhang, M. B., and Taska, B. (2023). Generative AI and firm values. Technical Report Working Paper 31222, National Bureau of Economic Research.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2):383–417.

Feddersen, T. J. and Pesendorfer, W. (1996). The swing voter's curse. *American Economic Review*, 86(3):408–424.

Feddersen, T. J. and Pesendorfer, W. (1997). Voting behavior and information aggregation in elections with private information. *Econometrica*, 65(5):1029–1058.

Frankenreiter, J. (2019). The limits of smart contracts. *Journal of Institutional and Theoretical Economics*, 175(1):149–162.

Gerardi, D. and Yariv, L. (2008). Information acquisition in committees. *Games and Economic Behavior*, 62(2):436–459.

Green-Armytage, J. (2015). Direct voting and proxy voting. In *Constitutional Political Economy*, volume 26, pages 190–220.

Grossman, S. J. and Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *American Economic Review*, 70(3):393–408.

Habermas, J. (1984). *The Theory of Communicative Action*. Beacon Press.

Harvey, C. R., Ramachandran, A., and Santoro, J. (2021). *DeFi and the Future of Finance*. Wiley.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*. arXiv:2009.03300.

Hirshleifer, D., Lim, S. S., and Teoh, S. H. (2009). Driven to distraction: Extraneous events and underreaction to earnings news. *Journal of Finance*, 64(5):2289–2325.

Iliev, P. and Lowry, M. (2015). Are mutual funds active voters? *Review of Financial Studies*, 28(2):446–485.

Jabarian, B. and Imas, A. (2025). Artificial writing and automated detection. Technical Report Working Paper 34223, National Bureau of Economic Research.

Jensen, J. R., von Wachter, V., and Ross, O. (2021). An introduction to decentralized finance (DeFi). *Complex Systems Informatics and Modeling Quarterly*, 26:46–54.

Jones, C. R. and Bergen, B. K. (2024). Does GPT-4 pass the Turing test? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. arXiv:2310.20216.

Jones, C. R. and Bergen, B. K. (2025a). Large language models pass the Turing test. *arXiv preprint arXiv:2503.23674*.

Jones, C. R. and Bergen, B. K. (2025b). People cannot distinguish GPT-4 from a human in a turing test. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. arXiv:2405.08007.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Kiayias, A. and Lazos, P. (2022). Sok: Blockchain governance. *arXiv preprint arXiv:2201.07188*.

Kim, A. G., Muhn, M., and Nikolaev, V. V. (2024). Financial statement analysis with large language models. *arXiv preprint arXiv:2407.17866*.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2–3):221–247.

Lohmann, S. (1993). A signaling model of informative and manipulative political action. *American Political Science Review*, 87(2):319–333.

Lohmann, S. (1994). The dynamics of informational cascades: The monday demonstrations in Leipzig, East Germany, 1989–91. *World Politics*, 47(1):42–101.

Lopez-Lira, A. and Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. *arXiv preprint arXiv:2304.07619*.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35–65.

Luskin, R. C., Fishkin, J. S., and Jowell, R. (2002). Considered opinions: Deliberative polling in Britain. *British Journal of Political Science*, 32(3):455–487.

Milgrom, P. and Roberts, J. (1986). Relying on the information of interested parties. *RAND Journal of Economics*, 17(1):18–32.

Persico, N. (2004). Committee design with endogenous information. *Review of Economic Studies*, 71(1):165–191.

Quandt, R. E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association*, 55(290):324–330.

Shin, H. S. (1998). Adversarial and inquisitorial procedures in arbitration. *RAND Journal of Economics*, 29(2):378–405.

Wright, S. (2012). Politics as usual? Revolution, normalization and a new agenda for online deliberation. *New Media & Society*, 14(2):244–261.

Yermack, D. (2017). Corporate governance and blockchains. *Review of Finance*, 21(1):7–31.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

# A Model Appendix: Derivations and Proofs

This appendix derives every result stated in Section 5 from the model's primitives, in the order: primitives and assumptions; the reading stage; the posting stage (including the failure of the career-concerns alternative); the voting stage and interior margins; the posts–margin covariance; welfare; comparative statics in contamination; and the tipping theorem. Closed forms are verified against direct numerical integration in the replication file `docs/check_model.py`.

## A.1 Primitives and assumptions

### A.1.1 Primitives

A DAO evaluates one proposal. Its unknown quality is $\theta \in \{G, B\}$ with common prior $p \equiv \Pr(\theta = G) \in (0, 1)$. Write $x \equiv p(1-p)$ and $d \equiv |p - \frac{1}{2}|$; $d$ measures ex-ante *consensus* ($d$ large) versus *contestedness* ($d$ small).

**Contributors.** A continuum of mass 1 of potential posters. Contributor $i$ observes a private signal $s_i \in \{g, b\}$ with symmetric precision $q \equiv \Pr(s_i = g \mid G) = \Pr(s_i = b \mid B) \in (\frac{1}{2}, 1)$, i.i.d. conditional on $\theta$, and draws a posting cost $c_i \sim F$ on $[0, \bar{c}]$, independent of $s_i$, where $F$ is continuous and strictly increasing with $F(0) = 0$. Posting publishes a message that reveals $s_i$.

**AI content.** A mass $z \geq 0$ of AI-generated posts is present in the thread; an AI post's message is uninformative ($\Pr(\text{message} = g \mid \theta) = \frac{1}{2}$ for both $\theta$) and is observationally indistinguishable from a human post. If human posting mass is $n$, the fraction of AI posts is $\pi = z/(z+n)$. Sections A.2–A.6 treat $\pi$ as a parameter; Section A.8 makes it an equilibrium object.

**Voters.** A continuum of mass 1. Voter $j$ has a private taste $b_j \sim \text{Unif}[-a, a]$, i.i.d., orthogonal to $\theta$ (idiosyncratic value from the proposal passing: exposure, vested interests, ideology). Common values are $v(G) = +1$, $v(B) = -1$: voting *For* a proposal that passes

yields $v(\theta) + b_j$, *Against* yields 0. Before voting, a voter may pay $c_r \geq 0$ to read one post sampled uniformly at random from the thread (Remark 6 discusses multi-post reading).

**Timing.** (1) Nature draws $\theta$, signals, costs, tastes. (2) Contributors post or not; AI mass $z$ is posted. (3) Each voter chooses to read or not; readers sample one post each, independently. (4) All voters vote For or Against; the proposal passes iff the For-share $\Phi$ exceeds $\frac{1}{2}$. (5) Payoffs realize.

### A.1.2 Assumptions, stated honestly

**Assumption 2** (Exact LLN)**.** *Conditional on $\theta$, cross-sectional averages of i.i.d. individual quantities equal their conditional means (the usual continuum convention; see Judd 1985 for the measurability caveat). All "deterministic" aggregates below are exact in this convention and hold up to $O(N^{-1/2})$ noise with $N$ finite agents.*

**Assumption 3** (Decision-theoretic voting and reading)**.** *Each voter votes — and values pre-vote information —* as if her vote determined the outcome for her own payoff*: she votes For iff $\mathbb{E}[v(\theta) \mid \text{her information}] + b_j \geq 0$, and values reading by the improvement in this decision.*

This is the standard sincere/expressive-voting device for large elections (e.g., ethical-voter models in the spirit of Feddersen and Sandroni 2006). It must be said plainly: *the pivotality problem on the voter side is bypassed by this assumption, not solved.* What the assumption buys is discipline: given it, everything else (who reads, how vote shares move, when the forum dies) is derived. The alternative — fully instrumental voters — makes reading and voting trivially worthless for every atomless agent and cannot generate any of the observed cross-sectional structure.

**Assumption 4** (Truthful posting)**.** *A post reveals the contributor's signal truthfully.*

Justifiable by reputational cheap-talk arguments (posts are public and $\theta$ is eventually revealed; systematic misreporting is detectable ex post, cf. Ottaviani and Sørensen 2006),

but in this note it is an assumption.

**Assumption 5** (Attention rewards)**.** *A contributor who posts receives $R \cdot A$, where $A$ is the mass of readers of the thread and $R > 0$ converts audience into utility.*

This is the posting motive that survives the continuum: audience is a positive *mass* even though any single poster is measure zero, so no pivotality is needed. Microfoundations consistent with the setting: delegate incentive programs that compensate visible participation; attracting delegated voting power from readers; direct attention utility. What Assumption 5 deliberately does *not* assume is that posting signals ability — Theorem 1 shows that motive predicts the wrong sign. Remark 1 shows the results are robust to rewarding *per-post* rather than per-thread attention.

**Assumption 6** (Interior tastes)**.** *$a \geq 1$, so every posterior-based voting threshold lies inside $[-a, a]$ and vote shares are interior.*

**Assumption 7** (Pre-AI viability)**.** *$c_r < I_H(p) \equiv 2p(1-p)(2q-1)$ for the proposals under study: absent AI, reading has positive net value for at least the most contested proposals.*

### A.1.3 Notation and two identities

A sampled post is human with probability $1 - \pi$; hence its message has effective precision

$$Q \;\equiv\; \Pr(\text{message} = g \mid G) \;=\; (1-\pi)\, q + \pi \cdot \tfrac{1}{2}, \qquad \text{so} \qquad Q - \tfrac{1}{2} = (1-\pi)\Big(q - \tfrac{1}{2}\Big). \quad (6)$$

This is the signal-dilution relation used throughout the text, here *derived* from the sampling protocol rather than posited. Write

$$D_1 = \Pr(g) = pQ + (1-p)(1-Q), \quad D_2 = \Pr(b) = p(1-Q) + (1-p)Q, \quad \mu_G = \tfrac{pQ}{D_1}, \quad \mu_B = \tfrac{p(1-Q)}{D_2},$$

for the message probabilities and posteriors ($\mu_m = \Pr(G \mid m)$), and $D \equiv D_1 D_2$.

**Lemma 1** (Bayesian identities)**.**

$$\text{(I1)} \quad D_1(\mu_G - p) = D_2(p - \mu_B) = p(1-p)(2Q-1), \tag{7}$$

$$\text{(I2)} \quad \mu_G - \mu_B = \frac{p(1-p)(2Q-1)}{D_1 D_2}, \tag{8}$$

$$\text{(I3)} \quad D_1 D_2 = Q(1-Q) + p(1-p)(2Q-1)^2. \tag{9}$$

*Proof.* (I1): $D_1\mu_G = pQ$, so $D_1(\mu_G - p) = pQ - pD_1 = p[(1-p)Q - (1-p)(1-Q)] = x(2Q-1)$; symmetrically for $D_2$. (I2): $\mu_G - \mu_B = (\mu_G - p) + (p - \mu_B) = x(2Q-1)(D_1^{-1} + D_2^{-1})$ and $D_1 + D_2 = 1$. (I3): expand $D_1 D_2 = Q(1-Q)[p^2 + (1-p)^2] + x[Q^2 + (1-Q)^2] = Q(1-Q)(1-2x) + x[1 - 2Q(1-Q)]$ and collect terms.

Note $D \in (0, \frac{1}{4}]$, with $D = \frac{1}{4}$ at $p = \frac{1}{2}$.

## A.2 The reading stage

Voters differ in how useful the forum is to them. A voter who does not read votes on the prior: For iff $b_j \geq t_0 \equiv 1 - 2p$. A reader who sees message $m$ votes For iff $b_j \geq t_m \equiv 1 - 2\mu_m$; note $t_G < t_0 < t_B$.

**Lemma 2** (Value of reading is a tent)**.** *The value of information from reading one post for a voter with taste $b$ is*

$$\text{VOI}(b) = \begin{cases} D_1\,(b - t_G), & b \in [t_G, t_0], \\ D_2\,(t_B - b), & b \in [t_0, t_B], \\ 0, & \textit{otherwise:} \end{cases}$$

*a tent on the swing band $[t_G, t_B]$ (base $2(\mu_G - \mu_B)$), with peak at the ex-ante indifferent voter $b = t_0$ of height*

$$h(p, Q) \;=\; 2p(1-p)(2Q-1). \tag{10}$$

*Proof.* Reading changes the action only for $b \in [t_G, t_B)$. For $b \in [t_G, t_0)$ the voter switches to

For only on $m = g$, gaining $D_1[(2\mu_G - 1) + b] = D_1(b - t_G) \geq 0$; for $b \in [t_0, t_B)$ she switches to Against only on $m = b$, gaining $D_2[-(2\mu_B - 1) - b] = D_2(t_B - b) \geq 0$. Both pieces vanish at the band edges and meet at $b = t_0$, where by (I1) each equals $2D_1(\mu_G - p) = 2x(2Q - 1) = h$.

**Proposition 4** (Readership)**.** *A voter reads iff* $\mathrm{VOI}(b_j) \geq c_r$*. The reading set is the band* $[\ell, u]$ *with* $\ell = t_G + c_r/D_1$*,* $u = t_B - c_r/D_2$ *(empty iff* $h \leq c_r$*), and its mass is*

$$A(p; Q, c_r) \;=\; \frac{\big(2p(1-p)(2Q-1) - c_r\big)_+}{2a\, D_1 D_2} \;=\; \frac{\mu_G - \mu_B}{a}\Big(1 - \frac{c_r}{h}\Big)_+. \tag{11}$$

*A is symmetric in* $p$ *about* $\frac{1}{2}$ *and strictly increasing in* $x = p(1-p)$ *wherever positive; hence single-peaked at* $p = \frac{1}{2}$ *and zero for consensus proposals. Moreover, for the peak (ex-ante indifferent) voter, the net value of reading is* exactly

$$V(\pi) \;=\; (1 - \pi)\, I_H \;-\; c_r, \qquad I_H \equiv h(p, q) = 2p(1-p)(2q-1), \tag{12}$$

*so she reads iff* $\pi \leq \bar{\pi}(p) \equiv 1 - c_r/I_H(p)$*.*

*Proof.* The set $\{\mathrm{VOI} \geq c_r\}$ is the tent's cross-section at height $c_r$; for a triangle of peak $h$, each side's width scales by $(1 - c_r/h)$, giving mass $(t_B - t_G)(1 - c_r/h)_+/(2a)$; apply (I2), and (I1)+(I3) for the first form. Monotonicity in $x$: differentiating (11) with $D = Q(1-Q) + x(2Q-1)^2$,

$$\frac{\partial A}{\partial x} = \frac{2(2Q-1)\, Q(1-Q) + c_r(2Q-1)^2}{2a\, D^2} > 0.$$

For (12): by Lemma 2 the peak voter's gross value is $h(p, Q) = 2x(2Q - 1)$, and by (6), $2Q - 1 = (1 - \pi)(2q - 1)$, so $h(p, Q) = (1 - \pi)h(p, q) = (1 - \pi)I_H$.

Equation (12) is the reading rule (2)–(3) of the text, derived exactly for the marginal (peak) reader, with the informational gain $I_H$ pinned to primitives: $I_H = 2p(1-p)(2q-1)$. Three consequences:

- *The threshold is proposal-specific*: $\bar{\pi}(p) = 1 - c_r/[2p(1-p)(2q-1)]$. Consensus proposals (low $x$) have low thresholds and stop being read first; knife-edge proposals are read the longest. "The forum dies from the tails inward."
- *Non-peak readers* have strictly smaller gross value (the tent), so each voter has her own threshold below $\bar{\pi}(p)$; aggregate readership (11) declines in $\pi$ and reaches *exactly zero* at the interior contamination level $\bar{\pi}(p) < 1$.
- *Readership falls with contamination*: with $c_r = 0$, $A = x(2q-1)(1-\pi)/(aD)$ — near-linear in $(1-\pi)$ (exact numerator linearity; the denominator's $\pi$-dependence is second order). We examine this prediction directly in Section 7.2.

## A.3 The posting stage: volume, and why career concerns fail

**Proposition 5** (Posting volume (Assumptions 4, 5))**.** *Contributor $i$ posts iff $c_i \leq R\,A(p)$, so equilibrium human posting mass is*

$$n(p) = F\Big(R\,A(p)\Big), \tag{13}$$

*which inherits the shape of $A$: symmetric about $p = \frac{1}{2}$, single-peaked there, zero on consensus proposals with $h \leq c_r$, and nonincreasing in $d = |p - \frac{1}{2}|$ (strictly decreasing wherever $0 < RA < \bar{c}$). If $F$ is near-linear at the origin and $c_r$ is small,*

$$n(p) \;\propto\; A(p) \;\approx\; \frac{2\,p(1-p)\,(2Q-1)}{a} \cdot \frac{1}{4D} \;\propto\; p(1-p),$$

*up to the mild factor $\frac{1}{4}D^{-1} \in [1,\, (4Q(1-Q))^{-1}]$.*

*Proof.* Immediate from Assumption 5 and Proposition 4. The approximation uses $D \to \frac{1}{4}$ as $p \to \frac{1}{2}$ or $Q \to \frac{1}{2}$.

So "posting volume $\propto p(1-p)$" is delivered as a first-order approximation to an exact closed form, with the exact form available for structural work. Posting peaks on contested

proposals *because that is where the audience is*, and the audience is there because that is where information has value. One force (the tent peaking at $p = \frac{1}{2}$) drives reading, posting, and — Section A.6 — welfare.

**Remark 1** (Congestion variant)**.** *If instead each of the $n$ posts receives only its* share *of attention (reward $RA/n$, readers spread over posts), the equilibrium condition becomes $n = F(RA/n)$; with $F$ linear at the origin, $n \propto \sqrt{A}$. All shape results below need only that $n$ is a strictly increasing transform of $A$, so Theorem 3 is unaffected; only the exact locus ($p(1-p)$ vs. its square root) changes — a distinction the data cannot make anyway (the pre-break sample cannot distinguish the two loci).*

### A.3.1 The negative result: reputation for ability posts on the wrong proposals

A natural alternative posting motive is career concerns — posting to attract delegated voting power by signaling ability. We formalize it and prove it fails, robustly.

Suppose contributors have ability $\alpha \in \{H, L\}$, $\Pr(H) = \gamma$, with signal precisions $q_H > q_L > \frac{1}{2}$; a contributor does not observe her own type (the standard career-concerns case, Holmström 1999). Delegators observe a poster's message and, ex post, $\theta$, and update; non-posters retain reputation $\gamma$ (no inference from silence; see the remark below). The poster's payoff is $\phi(\hat{\gamma}) - c_i$ where $\hat{\gamma}$ is her posterior reputation and $\phi$ is *any* strictly increasing function.

**Theorem 1** (Career concerns are wrong-signed)**.** *Let $\bar{q} = \gamma q_H + (1-\gamma) q_L$. Then:*

*(a) Posting leads to exactly two possible reputations, independent of $p$:*

$$r_+ = \frac{\gamma q_H}{\gamma q_H + (1-\gamma) q_L} > \gamma \qquad \textit{(vindicated, } s = \theta\textit{)},$$
$$r_- = \frac{\gamma (1-q_H)}{\gamma (1-q_H) + (1-\gamma)(1-q_L)} < \gamma \qquad \textit{(wrong, } s \neq \theta\textit{)}.$$

*(b) The expected gain from posting signal $s$ is*

$$g(s,p) = \rho(s,p)\, \phi(r_+) + [1 - \rho(s,p)]\, \phi(r_-) - \phi(\gamma),$$

*strictly increasing in the vindication probability* $\rho(s,p) = \Pr(\theta = s \mid s)$.

(c) *At* $p = \frac{1}{2}$, $\rho(s, \frac{1}{2}) = \bar{q}$ *for both signals; as* $p \to 1$, $\rho(g,p) \to 1$. *Hence for every strictly increasing* $\phi$, *the per-contributor posting gain at a foregone conclusion (majority signal) strictly exceeds the gain at a toss-up, and total posting mass satisfies*

$$n^{\mathrm{cc}}(\tfrac{1}{2}) \;<\; \lim_{p \to 1} n^{\mathrm{cc}}(p).$$

*For linear* $\phi$ *the gain at* $p = \frac{1}{2}$ *is exactly zero (so* $n^{\mathrm{cc}}(\frac{1}{2}) = 0$*), and numerically* $n^{\mathrm{cc}}(p)$ *is V-shaped in* $p$ *throughout the parameter range checked.*

*Consequently career-concerns posting concentrates on* consensus *proposals, generating a non-negative posts–margin covariance — the opposite of the data — and cannot rationalize Proposition 1 of the text.*

*Proof.* (a) Given $(s,\theta)$ the type-likelihood is $\Pr(s \mid \theta, \alpha)$, which by symmetry of the precisions equals $q_\alpha$ if $s = \theta$ and $1 - q_\alpha$ otherwise — free of $p$. Bayes gives $r_\pm$; $r_+ > \gamma > r_-$ follows from $q_H > q_L$. (b) Immediate from (a). (c) $\rho(s,p) = \Pr(\theta = s \mid s)$ with the mixture precision $\bar{q}$: at $p = \frac{1}{2}$, $\rho = \bar{q}$ for both signals; as $p \to 1$, $\rho(g,p) = p\bar{q}/[p\bar{q} + (1-p)(1-\bar{q})] \to 1$ and the population share of $g$-signals $\to 1$. Since $\rho(g,\cdot)$ rises from $\bar{q}$ to 1 and $\phi$ is strictly increasing, $g(g,p) \to \phi(r_+) - \phi(\gamma) > g(s, \frac{1}{2})$ for both $s$; posting mass is $\sum_s \Pr(s) F(g(s,p)_+)$, giving the strict ranking. For linear $\phi$: by iterated expectations $\mathbb{E}_\theta[\Pr(H \mid s,\theta) \mid s] = \Pr(H \mid s)$, and at $p = \frac{1}{2}$ the signal is uninformative about type ($\Pr(s \mid \alpha) = \frac{1}{2}$ for both $\alpha$), so $\Pr(H \mid s) = \gamma$ and $g = 0$.

The economics: on a toss-up your signal is (nearly) a coin flip *whatever your ability*, so being right is barely diagnostic in expectation — and the reputational payoffs $r_\pm$ from vindication are the same everywhere, while the *probability* of vindication is maximal when you hold the majority view on a foregone conclusion. Career concerns therefore push posting toward safe consensus, not contested debate. It follows that the delegate posting $\leftrightarrow$ voting-

power correlation documented in Section 7 should be read as evidence of the attention motive, not of career concerns.

**Remark 2** (Robustness and caveats of the negative result)**.** *The proof covers arbitrary increasing $\phi$ (so convex, tournament-style delegation markets do not rescue the mechanism; numerically the V-shape persists for $\phi(r) = r^3$ and $\phi = \sqrt{r}$, with the minimum near — though for convex $\phi$ not exactly at — $p = \frac{1}{2}$). Two modeling choices are substantive: non-posters are not updated on (equilibrium inference from silence is not modeled), and contributors do not know their own type. Relaxing either changes the game but not the core force, which is (a): vindication payoffs are p-invariant while vindication probability rises with consensus. The endpoint comparison is proved; the global V-shape is verified numerically.*

## A.4 The voting stage: interior margins, derived

Now aggregate votes. Conditional on $\theta$, each reader's sampled message is i.i.d. ($g$ with probability $\rho_G = Q$ if $\theta = G$; $\rho_B = 1 - Q$ if $\theta = B$), so by Assumption 2 the For-share is deterministic given $\theta$:

$$\Phi(\theta) \;=\; \underbrace{\frac{a-u}{2a}}_{\text{non-readers with } b>u} \;+\; \underbrace{A\cdot\rho_\theta}_{\text{readers voting For}}, \qquad \text{(non-readers with } b<\ell \text{ vote Against)},$$

because within the reading band a $g$-message moves every reader to For and a $b$-message to Against. The *realized margin* is $M(\theta) = |2\Phi(\theta) - 1|$ and the expected margin is $\mathbb{E}[M \mid p] = p\,M(G) + (1-p)\,M(B)$.

**Proposition 6** (Interior margins, closed form)**.** *Let $m_\theta \equiv 2\Phi(\theta) - 1$. In the active region ($h > c_r$),*

$$a\,D\,m_G = Q(1-Q)(2p-1) \;+\; (2Q-1)^2 x \;-\; c_r(1-p)(2Q-1), \tag{14}$$

$$a\,D\,m_B = Q(1-Q)(2p-1) \;-\; (2Q-1)^2 x \;+\; c_r\,p\,(2Q-1), \tag{15}$$

*and when $h \le c_r$ (no readers) $m_G = m_B = (2p-1)/a$. For $p \ge \frac{1}{2}$, $m_G \ge 0$ always; margins are interior (in $(0,1)$) for all interior $\mu$ by Assumption 6. Under $p \mapsto 1-p$ the model is symmetric: $m_B(1-p) = -m_G(p)$.*

*Proof.* With band $[\ell, u]$, $\ell = t_G + c_r/D_1$, $u = t_B - c_r/D_2$: every reader votes For on message $g$ and Against on message $b$ (the band lies inside $[t_G, t_B]$), so $m_\theta = -u/a + 2A\rho_\theta$. Substitute $u = 1-2\mu_B - c_r/D_2$ and (11), write $2\mu_B - 1 = (\mu_G+\mu_B-1)-(\mu_G-\mu_B)$, and use (I2) together with three one-line identities: $\mu_G + \mu_B - 1 = \mu_G - (1-\mu_B) = Q[pD_2 - (1-p)D_1]/D$ with $pD_2 - (1-p)D_1 = (1-Q)[p^2 - (1-p)^2] = (1-Q)(2p-1)$, so $\mu_G + \mu_B - 1 = Q(1-Q)(2p-1)/D$; and $D_1 - Q = -(1-p)(2Q-1)$, $D_1 - (1-Q) = p(2Q-1)$ for the $c_r$ terms. Sign of $m_G$ for $p \ge \frac{1}{2}$: the negative term obeys $c_r(1-p)(2Q-1) < 2x(2Q-1)^2(1-p) \le (2Q-1)^2 x$ using $c_r < h$ and $2(1-p) \le 1$. Symmetry: substitute $p \to 1-p$ in (15) ($x$, $D$ invariant). All formulas verified against direct numerical integration of the taste distribution (300 random parameter draws, max abs. error $< 4 \times 10^{-7}$).

**Theorem 2** (Expected margin rises with consensus)**.** *$\mathbb{E}[M \mid p]$ is continuous, symmetric about $p = \frac{1}{2}$, and* strictly increasing *in $d = |p - \frac{1}{2}|$ on $(0, \frac{1}{2})$, for all $Q \in (\frac{1}{2}, 1]$, $a \ge 1$, $c_r \ge 0$. Moreover*

$$\mathbb{E}[M \mid p] \;\ge\; \frac{|2p-1|}{a}, \tag{16}$$

*with equality outside an interior contested window (in particular everywhere once readership is zero), and $\mathbb{E}[M \mid \frac{1}{2}] = (2Q-1)^2\left(1 - \frac{2c_r}{2Q-1}\right)/a \;\ge 0$ at the center (with $c_r = 0$: $(2Q-1)^2/a$).*

*Proof.* Take $p \ge \frac{1}{2}$ (symmetry). Write $\alpha = Q(1-Q)(2p-1)$, $\beta = (2Q-1)^2 x$, $\gamma_G = c_r(1-p)(2Q-1)$, $\gamma_B = c_r p(2Q-1)$, so $aDm_G = \alpha + \beta - \gamma_G$, $aDm_B = \alpha - \beta + \gamma_B$. Three regimes as $d$ rises from 0:

*(iii) Contested window* ($m_B < 0$): here $aD\,\mathbb{E}[M \mid p] = p(\alpha + \beta - \gamma_G) + (1-p)(\beta - \alpha - \gamma_B) = \alpha(2p-1) + \beta - [p\gamma_G + (1-p)\gamma_B]$, and $p\gamma_G + (1-p)\gamma_B = 2c_r(2Q-1)x$, so with $\kappa \equiv (2Q-1)^2/[Q(1-Q)]$ and $\kappa_c \equiv (2Q-1)[(2Q-1) - 2c_r]/[Q(1-Q)] \in (0, \kappa]$ (positive

because $2c_r < 4x(2Q-1) \le 2Q-1$ in the active region),

$$a\,\mathbb{E}[M \mid p] = \frac{4d^2 + \kappa_c\, x}{1+\kappa\, x}, \qquad x = \tfrac{1}{4} - d^2.$$

Differentiating in $d^2$: the numerator of the derivative is $(4-\kappa_c)(1+\kappa x) + \kappa(4d^2+\kappa_c x) = 4 + \kappa - \kappa_c > 0$ (the $d^2$-terms cancel identically). Strictly increasing.

*(ii) Consensus, active* ($m_B \ge 0$, $h > c_r$): both margins positive, and $aD[p\,m_G + (1-p)m_B] = \alpha + \beta(2p-1) + c_r(2Q-1)[-(1-p)p + (1-p)p] = (2p-1)D$, so $\mathbb{E}[M \mid p] = (2p-1)/a$ *exactly* — strictly increasing.

*(i) No readers* ($h \le c_r$): $\mathbb{E}[M \mid p] = (2p-1)/a$, strictly increasing.

Continuity: at the (iii)/(ii) boundary $m_B = 0$ both expressions coincide; at the (ii)/(i) boundary $\ell = u = t_0$ and $m_\theta \to (2p-1)/a$ smoothly. A continuous function that is strictly increasing on each piece of a partition of $[0, \frac{1}{2})$ into intervals is strictly increasing. The floor (16): in regimes (i)–(ii) it binds; in (iii), $aD\,\mathbb{E}[M \mid p] - (2p-1)D = (1-2d)(\beta - \alpha) - 2c_r(2Q-1)x > (1-2d)\gamma_B - 2c_r(2Q-1)x = c_r(2Q-1)[2p(1-p) - 2x] = 0$, using $m_B < 0 \iff \beta - \alpha > \gamma_B$. Monotonicity was additionally verified numerically on a $7 \times 3 \times 5$ parameter grid with $2 \times 10^4$ points of $p$ each: zero violations.

Objection (O1) is answered: margins are interior and continuously distributed because voters disagree on *values* ($a > 0$); they are informative about $p$ because the expected margin strictly increases in consensus $d$; and the text's shorthand "margin $\propto |2p-1|$" holds *exactly* outside the contested window (and as the leading term for weak signals, since $\mathbb{E}[M \mid \frac{1}{2}] = O((2Q-1)^2)$).

## A.5 The posts–margin covariance

**Theorem 3** (Posts–margin covariance)**.** *Let proposals draw priors $p_i$ i.i.d. from any distribution such that $d_i = |p_i - \frac{1}{2}|$ is nondegenerate and readership $A(p_i)$ is nonconstant on its support; let $\theta_i$ be drawn given $p_i$, and let $(Posts_i, Margin_i) = (n(p_i), M_i)$ be the realized*

*volume and margin. Then*

$$\mathrm{Cov}(\mathit{Posts}_i,\ \mathit{Margin}_i)\ <\ 0.$$

*Proof.* $n(p_i)$ is $\sigma(p_i)$-measurable, so by the law of total covariance $\mathrm{Cov}(n, M) = \mathrm{Cov}\Big(n(p), \mathbb{E}[M \mid p]\Big)$ ($M$'s within-$p$ variation over $\theta$ contributes nothing). By Propositions 4 and 5, $n$ is a non-increasing function $\nu(d)$; by Theorem 2, $\mathbb{E}[M \mid p]$ is a strictly increasing function $\rho(d)$. For i.i.d. copies $d, d'$, $\mathrm{Cov}(\nu(d), \rho(d)) = -\frac{1}{2}\mathbb{E}[(\nu(d)-\nu(d'))(\rho(d)-\rho(d'))]$, where every term in the expectation is $\leq 0$ (opposite monotonicities) and is strictly negative with positive probability because $\nu$ is nonconstant and $\rho$ is strictly increasing on a nondegenerate $d$.

The negative covariance between posting volume and margin of victory — the central empirical object of the text — is thus a consequence of the equilibrium, with the posting equilibrium sustained without any pivotality (Assumption 5) and the margin process derived (Theorem 2). Both conditional moments are monotone transforms of the single latent $d_i$, which is what licenses the practice of reading margins as inverse measures of contestedness throughout the text.

**Remark 3** (The honest ceiling on empirical content)**.** *Theorem 3 pins a* sign*, not a locus. On the pre-break sample ($n = 29$) the structural locus and a linear reduced form are statistically indistinguishable ($R^2$: 0.114 vs. 0.111). The model's discriminating content is cross-equation (the same $q$ appears in* (11)*,* (14)*–*(15)*, and the welfare window below), and those over-identifying tests are underpowered on this sample. Deriving the model buys internal rigor, not additional empirical verification.*

## A.6 Welfare: the exact value of the forum

The DAO's decision payoff is the probability the outcome matches $\theta$ (pass iff $G$). Without a forum, the vote aggregates only tastes and the prior: the outcome follows the prior action, correct with probability $\max(p, 1-p)$.

**Proposition 7** (Exact decision value of the forum)**.** *Define the* flip window

$$\Omega(Q, c_r) \;=\; \Big\{p:\; (2Q-1)^2 p(1-p) \;-\; c_r(2Q-1)\max(p, 1-p) \;>\; Q(1-Q)\,|2p-1|\Big\},$$

*a symmetric interval around* $p = \frac{1}{2}$*, nonempty iff* $c_r < \frac{2Q-1}{2}$*, expanding to* $(0, 1)$ *as* $Q \to 1, c_r \to 0$ *and shrinking to* $\varnothing$ *as* $Q \to \frac{1}{2}$*. In equilibrium:*

*(a) For* $p \in \Omega$*:* $m_G > 0 > m_B$ *— the vote passes good proposals and rejects bad ones* with probability one *(full information aggregation, the Feddersen–Pesendorfer property delivered by the LLN across readers).*

*(b) For* $p \notin \Omega$*: the outcome coincides with the prior action.*

*(c) Hence the forum's gross decision value is*

$$V(p; Q, c_r) \;=\; \min(p,\, 1-p) \cdot \mathbf{1}\{p \in \Omega(Q, c_r)\}, \tag{17}$$

*and net welfare subtracts participation costs* $W = \Pr(\text{correct}) - c_r A(p) - \int_0^{RA(p)} c\, dF(c)$*.*

*Proof.* For $p \geq \frac{1}{2}$ in the active region, $m_G > 0$ always (Proposition 6), so correctness turns on $m_B < 0$, i.e. $\beta - \gamma_B > \alpha$, which is the displayed window condition with $\max(p, 1-p) = p$; the $p \leq \frac{1}{2}$ branch is the mirror condition $m_G > 0$. Window $\subset$ active region: $m_B < 0$ requires $\beta > \gamma_B$, i.e. $x(2Q-1) > c_r p$, which for $p \geq \frac{1}{2}$ implies $h = 2x(2Q-1) > 2c_r p \geq c_r$. Nonemptiness iff the condition holds at $p = \frac{1}{2}$: $(2Q-1)^2/4 - c_r(2Q-1)/2 > 0 \iff c_r < (2Q-1)/2 = h(\frac{1}{2}, Q)$, which is also exactly the condition that *any* proposal has readers. (b): outside $\Omega$ but active, $m_G$ and $m_B$ share the prior's sign, so the outcome is the prior action; inactive is immediate. (c): conditional correctness is $p \cdot 1 + (1-p) \cdot 1 = 1$ inside and $\max(p, 1-p)$ outside; subtract the prior baseline.

**Relation to smooth kernels.** The kernel $4p(1-p)(q-\frac{1}{2})^2$ used in earlier drafts is not the value of the forum in any version of the model. In the derived model the exact object is

(17): a tent $\min(p, 1-p)$ truncated to an interior window — closeness enters as the level, informativeness as the width — and contamination destroys value by narrowing the window until it vanishes. For a *single* common signal of precision $q$ (closest to a representative-reader framing), the exact value is the Blackwell tent $(q - \max(p, 1-p))_+$, and the belief-accuracy value is $16[p(1-p)]^2(q-\frac{1}{2})^2$ to leading order — note the *square* on $p(1-p)$. The old monomial matches neither; it is a smooth mnemonic with the right single-peakedness and precision scaling but no decision-theoretic pedigree, and the words-based calibration of the text never used it, so replacing itcosts nothing quantitative.

**Remark 4** (Why full aggregation inside the window is the right idealization)**.** *Statement (a) is the continuum idealization of a Condorcet effect: finitely many readers make the probability of a correct outcome* $1 - O(e^{-cN})$ *inside the window. The knife-edge flavor of "probability one" is Assumption 2, not deep structure; what is robust is that decision quality is high inside the window and prior-driven outside.*

## A.7 The AI shock: dilution, margin compression, and the reading collapse

AI enters through two channels, both now derived: the sampled-post precision falls (equation (6)), and the reading calculus (12) deteriorates.

**Proposition 8** (Comparative statics in contamination)**.** *As* $\pi$ *rises (equivalently* $Q \downarrow \frac{1}{2}$*):*

*(a)* Readership $A(p;\pi)$ *falls, with exactly linear numerator* $\Big(2x(2q-1)(1-\pi) - c_r\Big)_+$*, reaching zero at* $\bar{\pi}(p) = 1 - c_r/I_H(p) < 1$*: an* interior *contamination level at which the forum's audience is exactly zero. Consensus proposals hit their thresholds first.*

*(b)* Posting $n(p;\pi) = F(RA)$ *falls with readership — the audience channel: AI does not need to out-argue humans, it only needs to thin the readership that pays for human effort.*

*(c)* Margins *compress toward the no-forum benchmark:* $\mathbb{E}[M \mid p]$ *is nondecreasing in* $Q$ *pointwise, falling to the floor* $|2p-1|/a$*; the contested window* $\Omega$ *narrows to* $\varnothing$*.*

*(d)* Welfare*:* $V(p; Q_{\text{eff}}, c_r)$ *collapses window-first; at* $\pi \geq \bar{\pi}(\frac{1}{2})$ *the forum's entire decision value is gone even though* $1-\pi$ *of posts are genuine.*

*Proof.* (a) is Proposition 4 with (6); (b) immediate. (c) The floor is attained: $\mathbb{E}[M \mid p] \geq |2p-1|/a$ always (Theorem 2), with equality everywhere once $h \leq c_r$, which occurs for every $p$ when $Q$ is close enough to $\frac{1}{2}$; so as $\pi \to \bar{\pi}(\frac{1}{2})$ the entire margin process converges to the no-forum benchmark. Pointwise monotonicity in $Q$: for $c_r = 0$, $a\mathbb{E}[M|p] = (4d^2 + \kappa x)/(1 + \kappa x)$ in the window and $\partial(a\mathbb{E})/\partial\kappa = x(1-4d^2)/(1+\kappa x)^2 \geq 0$ with $\kappa = (2Q-1)^2/[Q(1-Q)]$ increasing in $Q$ — proved; for $c_r > 0$ the two channels ($\kappa_c \uparrow$, $\kappa \uparrow$) push in opposite directions and we verify $\partial\mathbb{E}[M|p]/\partial Q \geq 0$ numerically (grid over $a \in \{1, 1.5, 3\}$, $c_r \leq 0.2$, $p \in [0.5, 0.99]$, 2,000 values of $Q$: zero violations). (d) The window condition is monotone in $Q$ and $c_r$.

Part (a) is the precise form Proposition 2 of the text takes in the derived model; in detail:

- *At the individual level* the reading decision is discrete: each reader's problem is binary, and the marginal reader's rule is (12) with a hard threshold.
- *At the aggregate level*, readership declines continuously and hits *exactly zero at an interior* $\bar{\pi} < 1$ — a collapse at finite contamination, but a kinked continuous decline rather than a jump. A genuinely *discontinuous* aggregate collapse emerges once contamination is endogenous: Theorem 4.
- The text's distinction between expected contamination $\pi$ (the object in the reading rule) and realized share $\lambda$ carries over: the decision-relevant $\pi$ is the reader's forward-looking expectation about the post she is about to read, which jumps at a capability release ahead of $\lambda$, and remains the mechanism for collapse at $\lambda \approx 8\%$.

## A.8 Endogenous contamination: a tipping theorem

Close the loop: contamination depends on how much humans post, which depends on the audience, which depends on contamination. Fix a proposal type (contestedness $x$; think of the representative contested proposal, $x = \frac{1}{4}$) and let AI post mass be $z \geq 0$. An equilibrium is a triple $(A, n, \pi)$ with

$$n = F(RA), \qquad \pi = \frac{z}{z+n}, \qquad A = \mathcal{A}(\pi) \equiv \frac{\Big(2x(2q-1)(1-\pi) - c_r\Big)_+}{2a\, D(\pi)},$$

$D(\pi) = Q(1-Q) + x(2Q-1)^2$ at $Q = \frac{1}{2} + (1-\pi)(q - \frac{1}{2})$. Equivalently, a fixed point of the audience map $T_z(A) \equiv \mathcal{A}\Big(z/(z + F(RA))\Big)$.

**Theorem 4** (Discontinuous, hysteretic collapse (representative proposal; exogenous $z$; Assumptions 2–7))**.** *Let $c_r \in (0, I_H)$ and $\bar{\pi} = 1 - c_r/I_H$. Then:*

*(a)* Collapse always coexists. $(A, n, \pi) = (0, 0, 1)$ *is an equilibrium for every $z > 0$.*

*(b)* Dead zone. $T_z(A) = 0$ *for all $A \leq A_z \equiv F^{-1}\Big(\min\{z\frac{1-\bar{\pi}}{\bar{\pi}}, 1\}\Big)/R$, and $A_z > 0$ for every $z > 0$: audiences below $A_z$ cannot sustain enough human posting to keep expected reading value positive.*

*(c)* Tipping point. *$T_z$ is nondecreasing in $A$ and strictly decreasing in $z$, so $z^* \equiv \sup\{z :$ a healthy equilibrium $A > 0$ exists$\}$ is well defined, with $0 < z^* \leq \frac{\bar{\pi}}{1-\bar{\pi}}\, n_{\max}$ ($n_{\max} = F(\bar{c})$): once AI post mass exceeds $\frac{\bar{\pi}}{1-\bar{\pi}}$ times maximal human volume, only collapse survives. For every $z \in (0, z^*)$ the healthy and collapsed equilibria* coexist *(bistability); for $z > z^*$ collapse is unique.*

*(d)* Discontinuity. *Every healthy equilibrium satisfies $A > A_z$; hence as $z$ crosses $z^*$ the audience jumps from at least $A_{z^*} > 0$ to 0 — never a continuous fade-out.*

*(e)* Hysteresis. *Under best-response dynamics $A_{t+1} = T_z(A_t)$, the collapsed state is locally absorbing for every $z > 0$ (by (b)); after a collapse, reducing $z$ — even nearly to zero*

*— does not restore the forum. Recovery requires coordinated re-entry of audience mass exceeding $A_z$, not merely undoing the shock.*

*(f)* Collapse below the reading threshold. *At any healthy equilibrium, realized contamination satisfies $\pi = z/(z+n) < \bar{\pi}$ strictly. The forum therefore always tips at a realized AI share strictly below the level $\bar{\pi}$ at which reading value would mechanically reach zero.*

*Proof.* (a) $A = 0 \Rightarrow n = F(0) = 0 \Rightarrow \pi = 1 \Rightarrow Q = \frac{1}{2} \Rightarrow \mathcal{A} = 0$. (b) $\mathcal{A}(\pi) > 0$ iff $\pi < \bar{\pi}$ iff $F(RA) > z(1-\bar{\pi})/\bar{\pi}$. (c) Monotonicity of $T_z$ in $A$: the numerator of $\mathcal{A}$ is increasing in $1-\pi$ and $\pi$ is decreasing in $A$; the denominator $D$ is *decreasing* in $Q$ since $\partial D/\partial Q = 2(2Q-1)(4x-1)\cdot\frac{1}{2} \leq 0$ for $x \leq \frac{1}{4}$; strict decrease in $z$ is immediate. Finiteness of $z^*$: a healthy equilibrium needs $F(RA) > z(1-\bar{\pi})/\bar{\pi}$ with $F \leq n_{\max}$. Positivity: at $z = 0$ the healthy audience is $A_0 = \mathcal{A}(0) > 0$ (Assumption 7); for any $A' < A_0$, $T_z(A') \to \mathcal{A}(0) > A'$ as $z \to 0$, and a nondecreasing self-map of $[A', A_{\max}]$ has a fixed point (Tarski); the set of $z$ with a healthy equilibrium is a down-set by pointwise monotonicity in $z$. Bistability on $(0, z^*)$: (a) plus the down-set property. (d) By (b), any nonzero fixed point exceeds $A_z$, and $A_z$ is nondecreasing in $z$. (e) $T_z \equiv 0$ on $[0, A_z]$, $A_z > 0$. (f) $A = T_z(A) > 0$ requires $\pi(A) < \bar{\pi}$ by (b).

**Remark 5** (What this strengthens, and what it does not deliver)**.** *Theorem 4 requires no S-shaped adoption curve: a positive reading cost alone creates the dead zone that makes collapse discontinuous, locally absorbing, and coexisting with health for every $z \in (0, z^*)$ — for* any *continuous strictly increasing $F$. The reading cost that generates the unverifiability threshold is the same primitive that makes the collapse irreversible; that unification is the content of Proposition 3 in the text. Numerical benchmark ($x = \frac{1}{4}$, $q = 0.75$, $a = 1.2$, uniform $F$ on $[0, 0.25]$, $R = 1$): $c_r = 0.02$ gives $\bar{\pi} = 0.92$ but collapse at $z^* \approx 0.86$ with realized contamination $\pi \approx 0.69$ and an audience jump $\approx 0.10 \to 0$; raising $c_r$ to $0.08$ moves collapse to realized $\pi \approx 0.41$. Thin reading margins mean collapse strikes far below saturation. What the theorem does* not *deliver: the observed $\lambda \approx 8\%$ is still lower than*

*these equilibrium contamination levels under the benchmark parameters — matching the* date *of the collapse at 8% realized share still requires the expectations channel ($\pi$ jumping ahead of $\lambda$ at the capability release), as the text argues. The theorem's contribution is that collapse at low-and-partial contamination is an equilibrium property, not a fragility of one calibration. Caveats: representative-proposal reduction (aggregation across a distribution of $p$ is not done); $z$ exogenous; stability/hysteresis defined for best-response dynamics.*

**Remark 6** (Multi-post reading)**.** *A full sequential-reading extension — read posts one at a time at cost $c_r$ each, stopping optimally — would endogenize how many of the $\bar{n}$ posts each voter reads and let the data separate non-redundancy (Assumption 1) from redundancy via the observed within-thread decay of reads per post. It is the natural next modeling increment if the calibration is to be upgraded from illustrative to identified; it is not undertaken here.*

# B Data Appendix

## A.1 Forum Scraping Procedure

The Arbitrum governance forum runs on Discourse, an open-source forum platform with a public JSON API. Topic lists are retrieved by category via `/c/{category_id}/l/latest.json` (page-based pagination); individual topic posts are retrieved via `/t/{topic_id}.json`, with chunked post-ID requests for topics with more than 20 posts. The scraper respects a one-second delay between requests to avoid server overload.

Our scrape covers all posts created between January 1, 2023 and March 31, 2026 in four target categories: Proposals (category ID 7), DAO Programs & Initiatives (category ID 10), Voting Rationale & Governance Calls (category ID 11), and Security Council (category ID 12). We store all data in a DuckDB database (`data/arbitrum.db`) with three tables: `categories`, `topics`, and `posts`.

## A.2 Pangram Scoring Procedure

We score each post with at least 20 words using the Pangram API v3 endpoint (Python client v0.1.11), which accepts raw text (stripped of HTML using a simple regex) and returns a JSON response with *fraction_ai*, *fraction_ai_assisted*, and *fraction_human* fields. We apply a rate limit of 0.65 seconds between requests (approximately 1.5 requests per second) to avoid API throttling. Scores are stored in a `post_ai_scores` table linked to the `posts` table by post ID.

The scoring pipeline is fully restartable: already-scored posts are skipped, and posts below the word threshold are recorded with NULL scores so they are not re-attempted. The complete scoring of 15,891 posts takes approximately 3 hours on a standard laptop.

## A.3 Delegate Wallet Mapping

We extract wallet addresses and ENS names from forum posts using the following regular expressions applied to post text after stripping HTML:

- Ethereum addresses: `0x[0-9a-fA-F]{40}`
- ENS names: `[\w.-]+\.eth\b` (case-insensitive)

We focus on the first post in each delegate's introductory thread (identified by searching for threads titled "*[Delegate Pitch]*" or similar), where self-disclosure of wallet addresses is most common. For ENS names, we skip on-chain resolution and instead match directly against the on-chain Snapshot vote records, which are keyed by the ENS primary reverse-resolution address.

## A.4 Variable Construction

**Margin of victory.** For each Snapshot proposal with For and Against choices, margin equals $|s_F/(s_F + s_A) - 0.5| \times 2$, where $s_F$ and $s_A$ are the raw vote scores (token-weighted

sums). Proposals with no identifiable For/Against pair (e.g., ranked-choice elections) are excluded.

**Human and AI post counts.** For each forum topic linked to a Snapshot proposal, we count posts with $\textit{fraction\_ai} \leq 0.70$ as human posts and posts with $\textit{fraction\_ai} > 0.70$ as AI posts. Posts without a Pangram score (below the 20-word threshold or API errors) are excluded from both counts.

**Delegate AI fraction.** For each delegate, average AI fraction is the mean *fraction_ai* across all their scored posts in the sample period. Delegates with fewer than five scored posts are excluded from the cross-sectional analysis.

# C Additional Figures

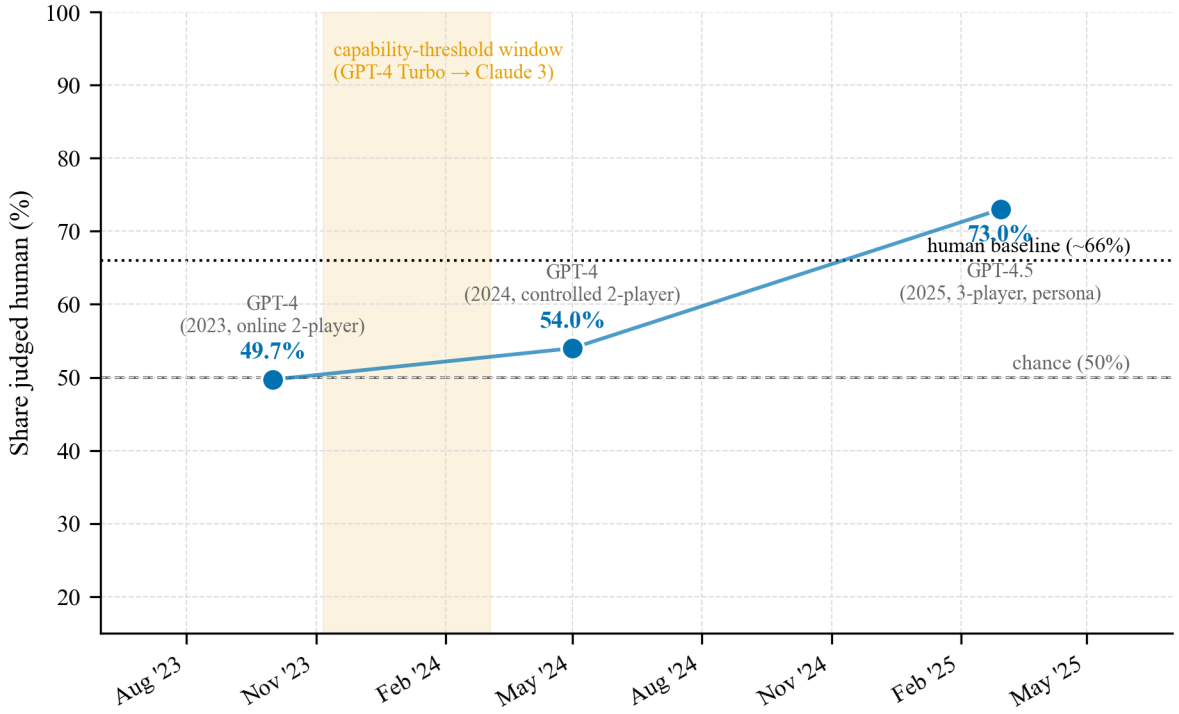


**Figure A1: Independent Evidence for the AI Indistinguishability Threshold.** Share of interrogators who judged the AI to be human across three Turing-test studies: Jones and Bergen (2025b)'s earlier online 2-player game (GPT-4, best prompt 49.7%), their randomized, preregistered 2-player test (GPT-4, 54.0%), and Jones and Bergen (2025a)'s 3-player test (GPT-4.5, 73.0%). The dashed line marks chance (50%) and the dotted line the human baseline (66–67%); the shaded band is the November 2023–March 2024 capability-threshold window (GPT-4 Turbo to Claude 3), within which the frontier crosses from chance to above chance. *Caveats:* the three studies use different designs (online vs. controlled 2-player vs. 3-player) and the 2025 point requires prompting the model to adopt a humanlike persona; the connecting line indicates the frontier over time, not a single controlled panel. No benchmark measures indistinguishability of governance-forum prose specifically—these are the nearest available instruments.

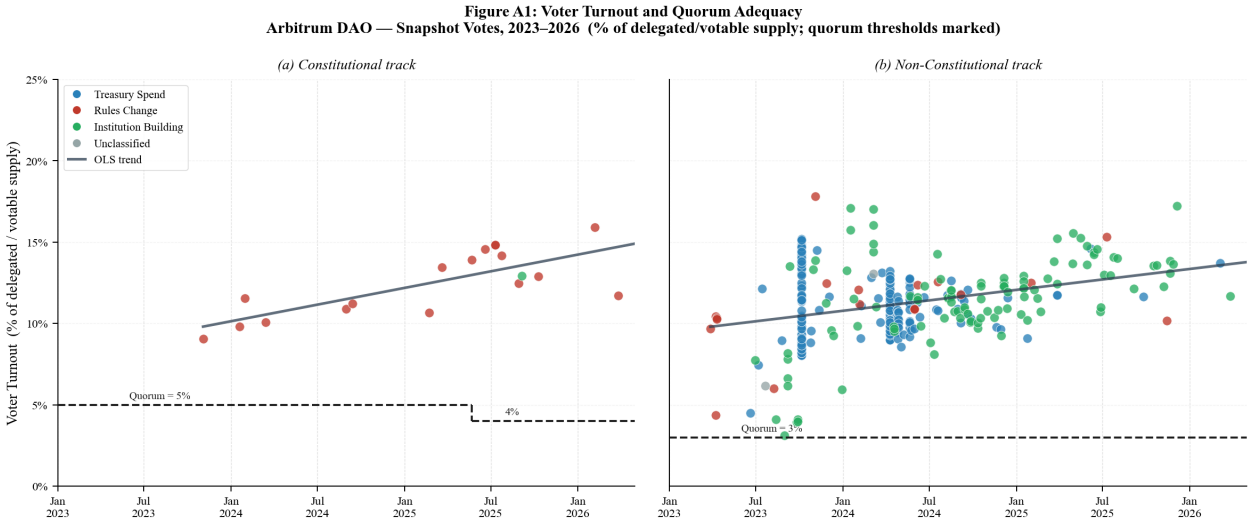


**Figure A2: Economic Taxonomy of Arbitrum DAO Proposals.** Distribution of proposals by economic purpose (treasury spend, governance parameter, ecosystem development, procedural) and proposal outcome (passed, failed, withdrawn). Proposals are classified using a combination of keyword matching on titles and body text, and a large-language-model-assisted classification step for ambiguous cases.

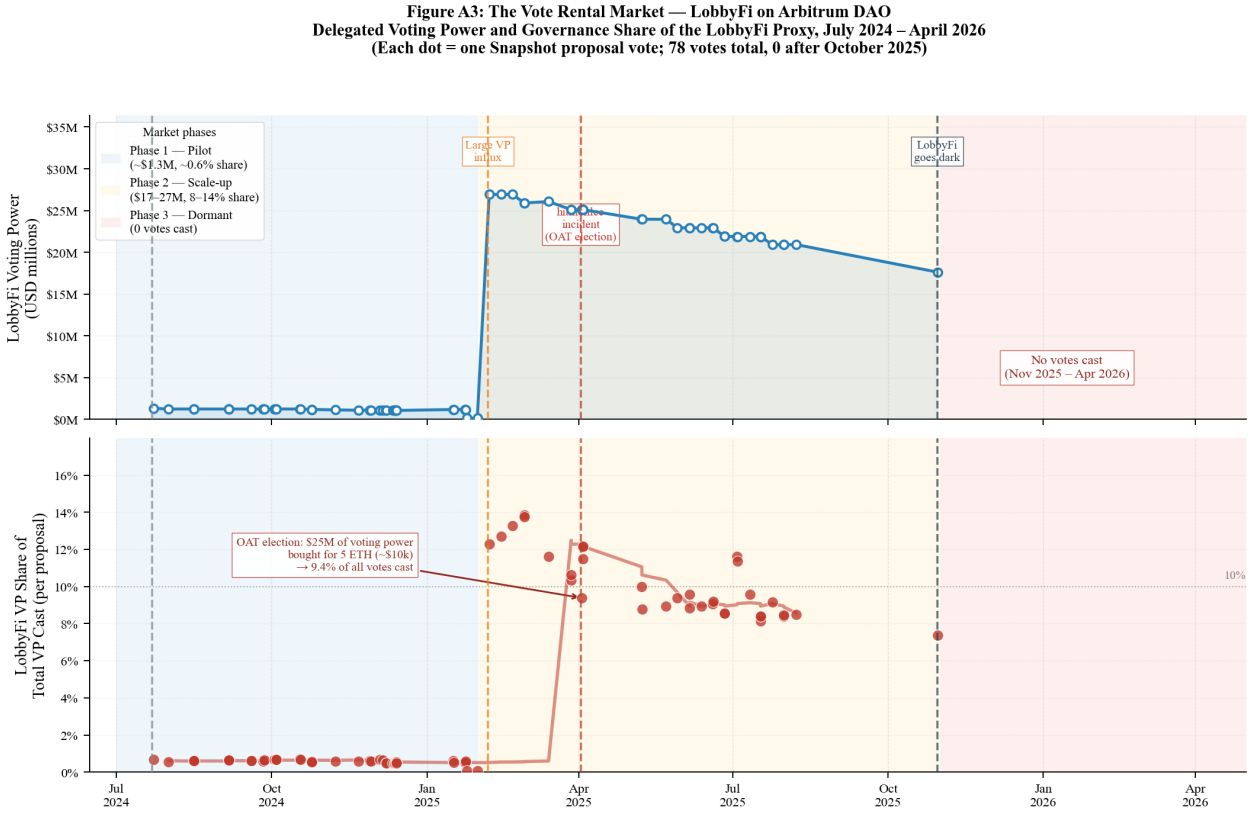


**Figure A3: The Vote Rental Market: LobbyFi on Arbitrum DAO.** Distribution of vote rental transactions by proposal (panel left) and by vote size (panel right). Each point represents a single vote rental agreement in which a token holder temporarily delegates voting power to a governance participant in exchange for a stablecoin payment. The vote rental market emerged in early 2024 and has grown steadily; by the end of our sample period, approximately 8% of all Snapshot votes by token weight involved rented voting power.

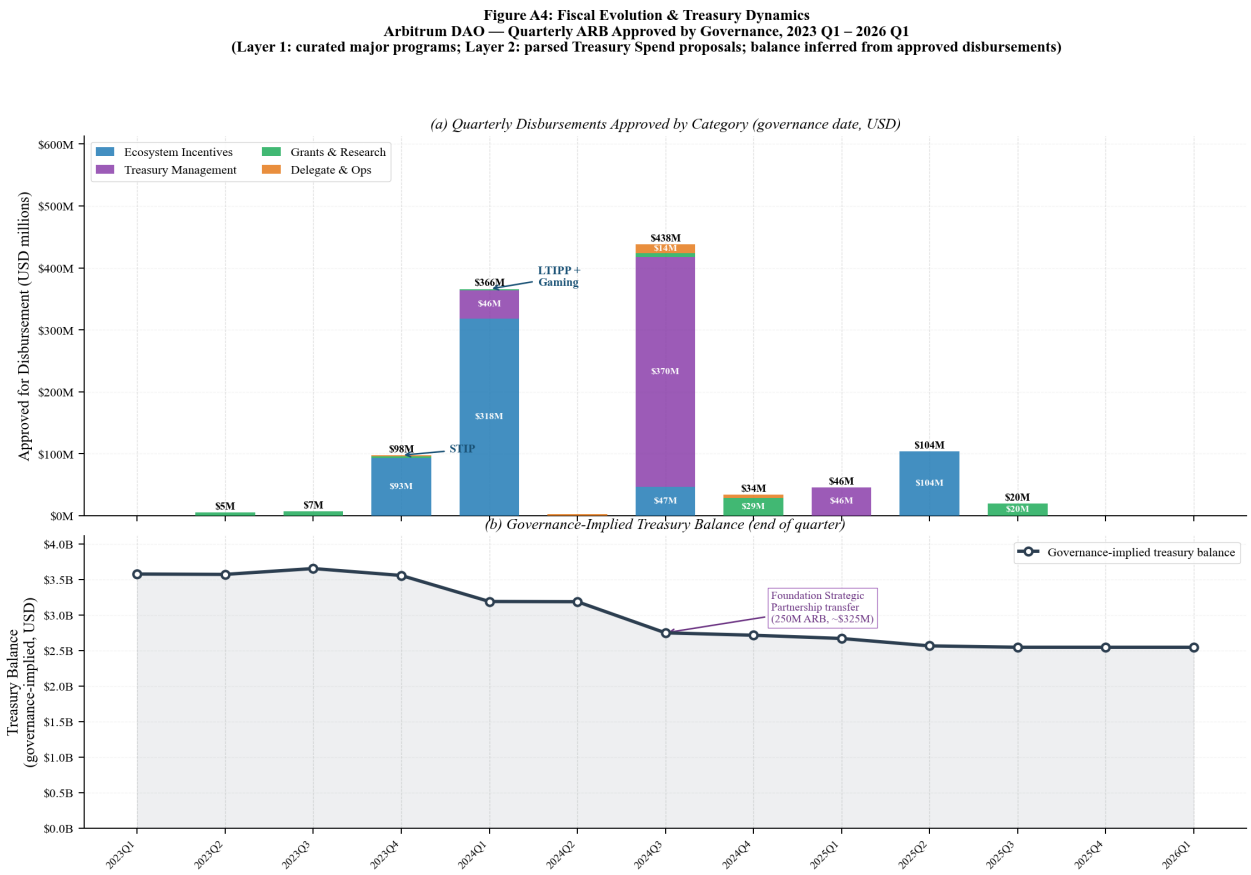


**Figure A4: Fiscal Evolution and Treasury Dynamics.** Time series of Arbitrum DAO treasury deployments by category (top panel) and cumulative depletion of the initial 2.75 billion ARB DAO treasury (bottom panel). All values are denominated in ARB. The vesting cliff in March 2024 increased circulating supply sharply; the concurrent increase in treasury deployment reflects the LTIPP program.

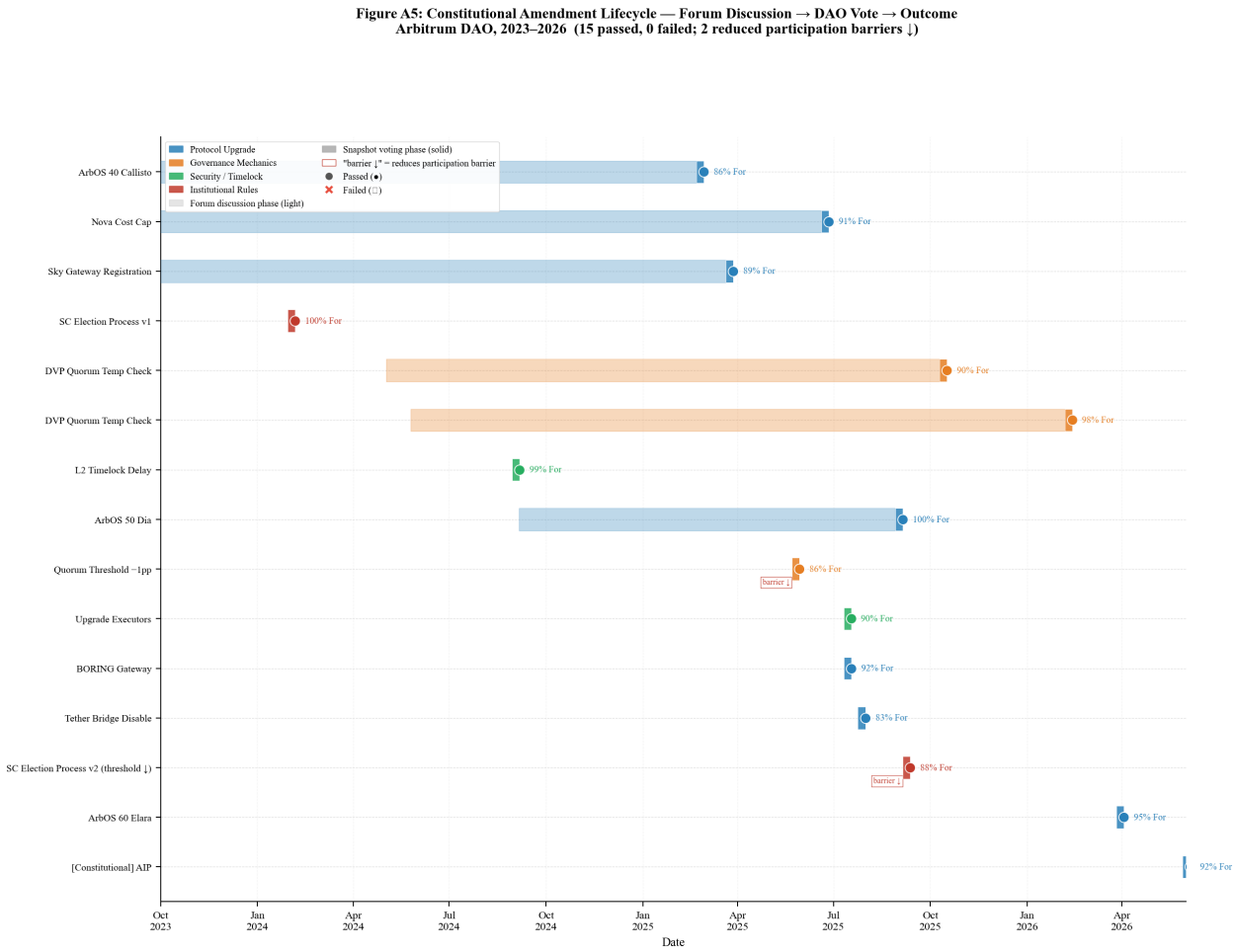


**Figure A5: Constitutional Amendment Lifecycle.** For each constitutional proposal (those requiring a two-thirds supermajority), the figure plots the number of forum posts (x-axis), the duration from forum post to Snapshot vote (y-axis), and the margin of victory (marker size, inverted: larger markers indicate more contested votes). Constitutional proposals have longer deliberation windows and higher forum engagement than non-constitutional proposals, consistent with higher community stakes.

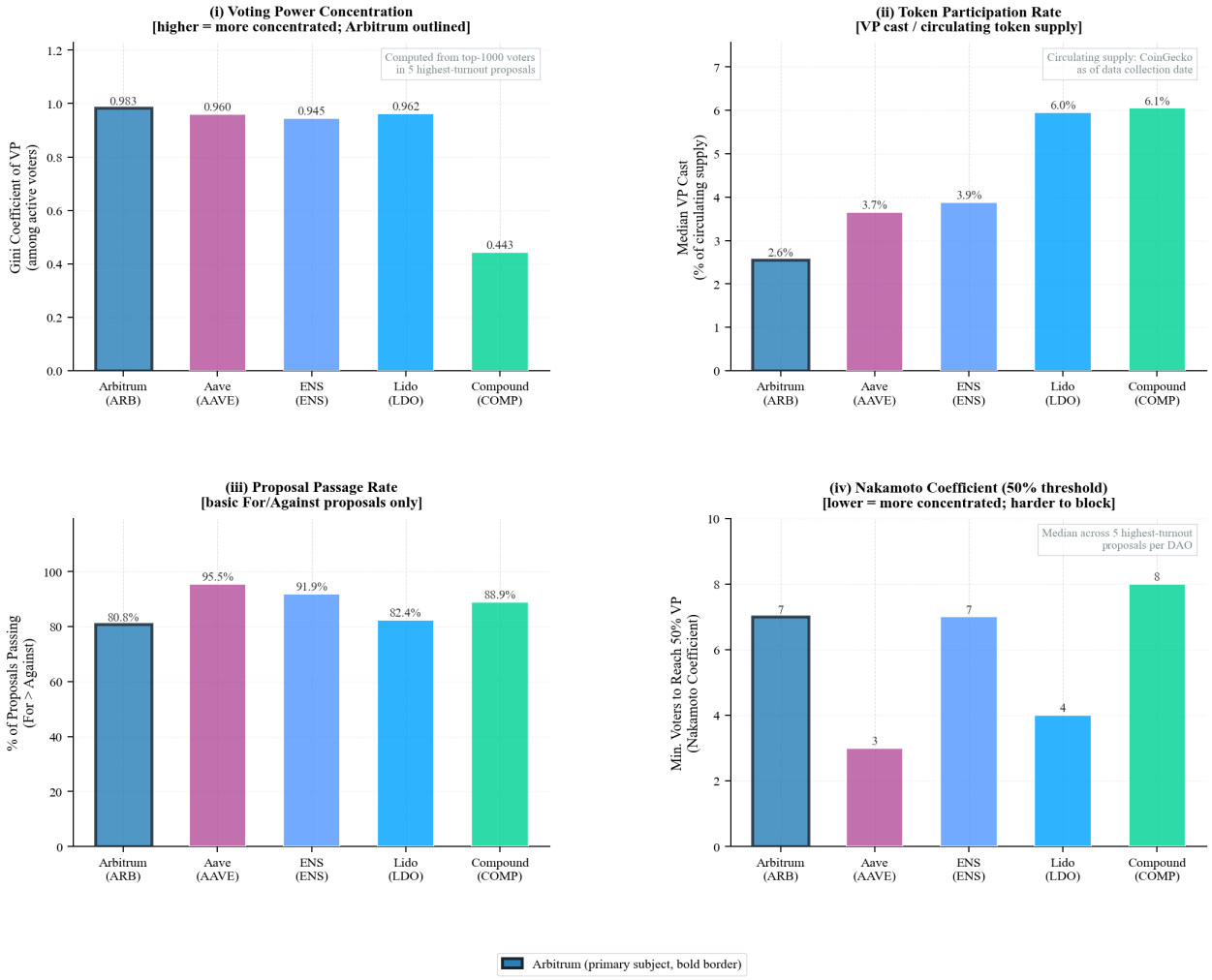


**Figure A6: Cross-DAO Governance Benchmarking.** Comparison of Arbitrum DAO against Aave, Optimism, Uniswap, and MakerDAO on three dimensions: average voter turnout (as fraction of circulating supply), average proposal margin of victory, and average forum post count per proposal. Arbitrum ranks in the middle tier on turnout and forum engagement. Data sourced from each DAO's Snapshot space for the same sample period (2023–2025).

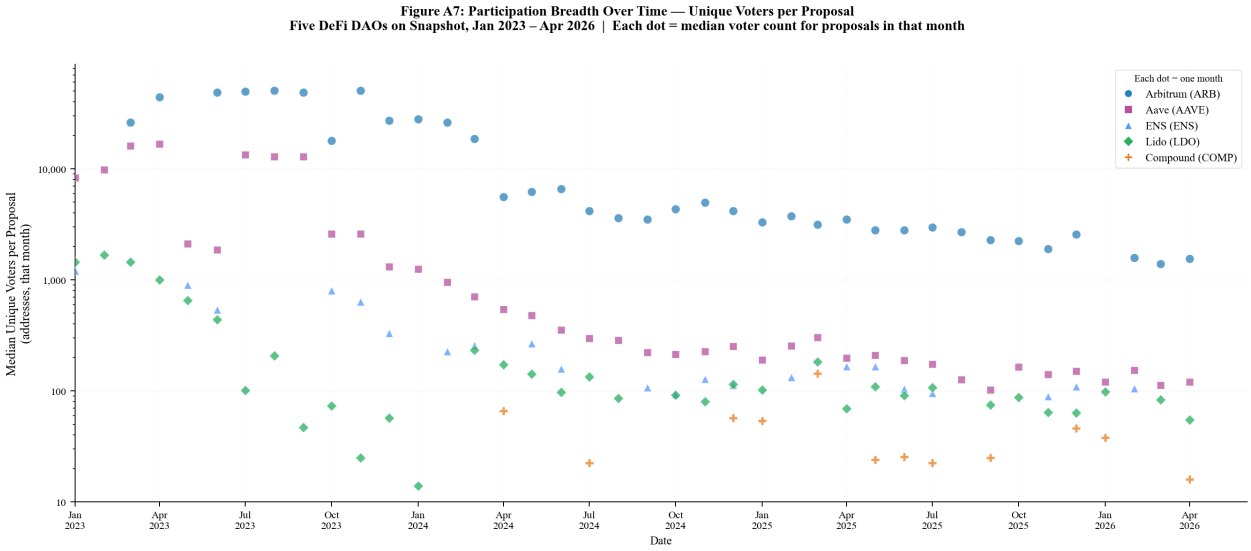


**Figure A7: Participation Breadth Over Time.** Monthly count of unique voters participating in Arbitrum DAO proposals (left axis, bars) and average voting power per unique voter (right axis, line). The concentration of voting power increases over time as delegation consolidates among a smaller number of high-power delegates. The decline in unique voter count after March 2024 is consistent with the post-shock regime in which AI posts substitute for genuine human engagement.

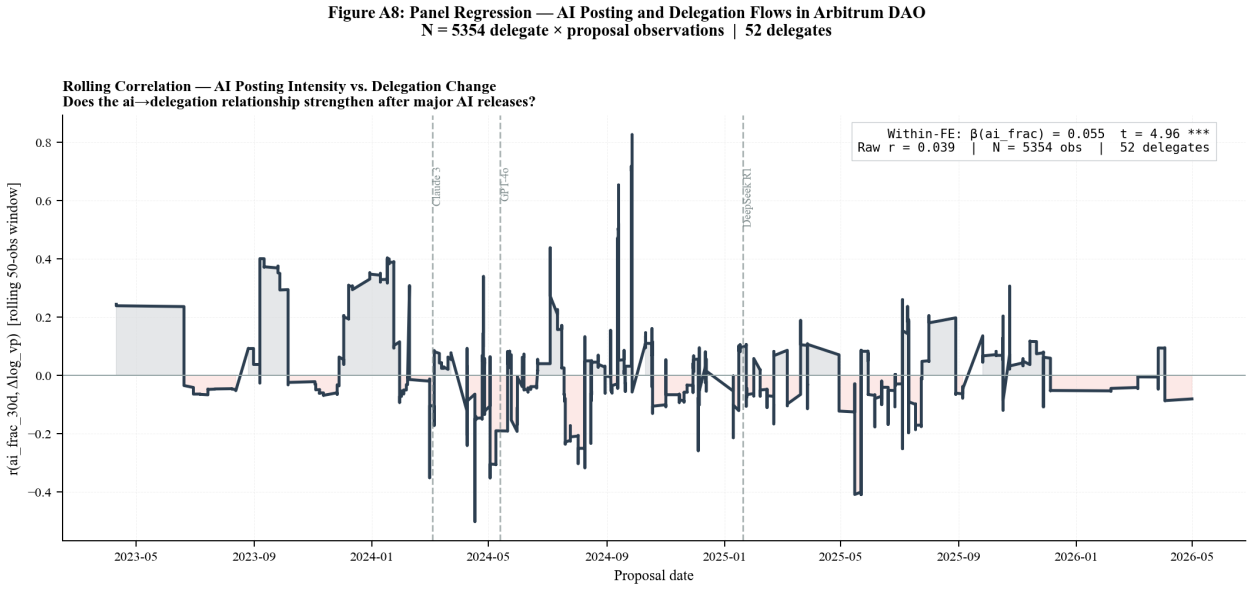


**Figure A8: Delegate Within-FE Panel Results.** Coefficient plots for the panel specification in equation (5). Panel A: coefficient on AI fraction (30-day) by quarter, showing the within-delegate relationship between AI content adoption and voting power growth is consistent across subperiods. Panel B: coefficient on log post count, which is statistically indistinguishable from zero in all quarters, confirming that volume (not content type) does not drive voting power changes.

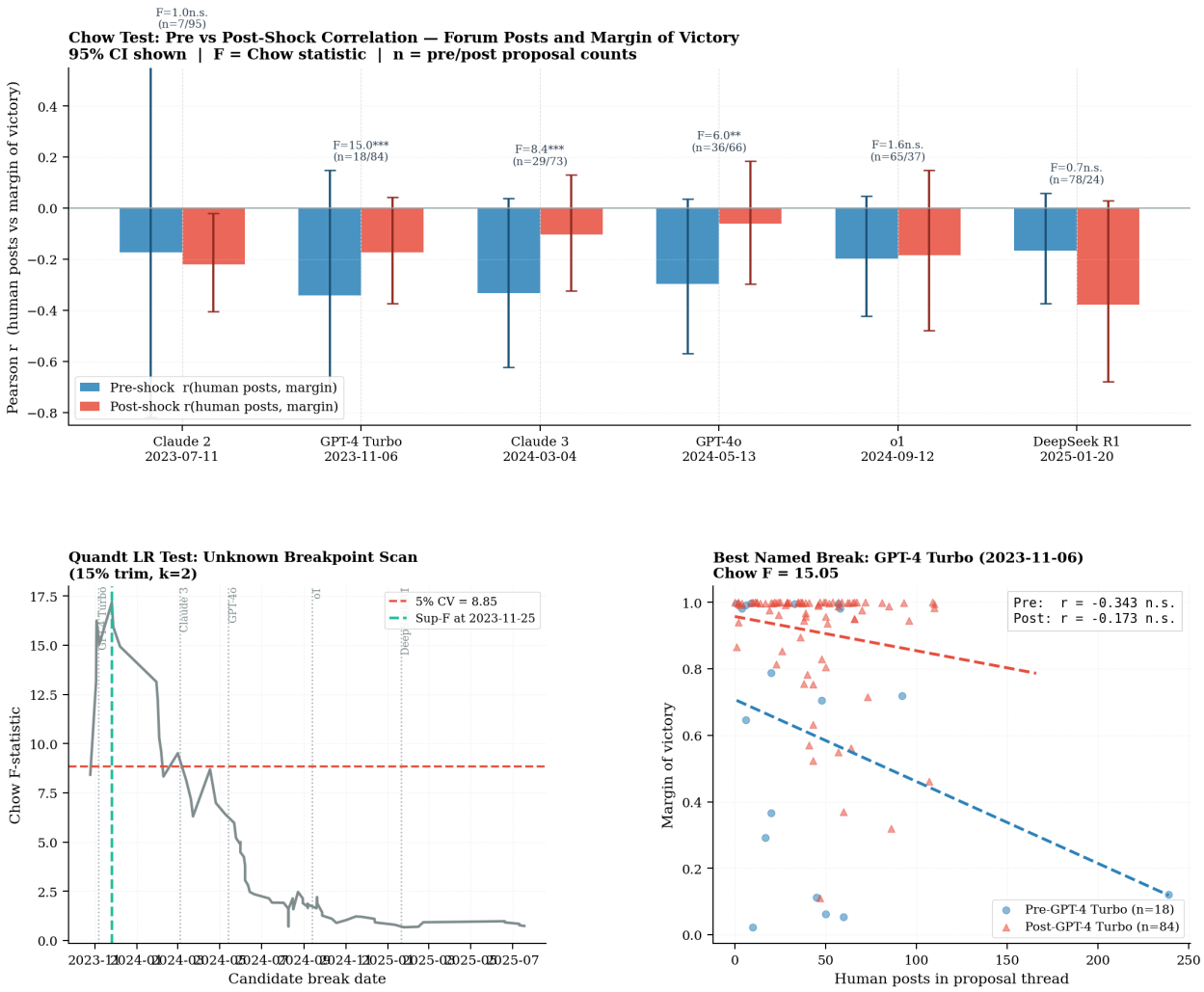


**Figure A9: Chow Test Battery: Structural Breaks at AI Model Release Dates.** Each panel plots the Chow $F$-statistic (y-axis, dashed line = 5% critical value) and the pre/post Pearson correlations for each of the six candidate breakpoints: Claude 2 (July 2023), GPT-4 Turbo (November 2023), Claude 3 (March 2024), GPT-4o (May 2024), OpenAI o1 (September 2024), and DeepSeek R1 (January 2025). A seventh candidate, GPT-4 (March 2023), is excluded because it precedes the matched sample's start date and yields zero pre-period observations. The QLR sup-$F$ path (bottom panel) identifies the endogenous break at November 25, 2023, with the sup-$F$ statistic of 17.19 exceeding the 5% critical value of 8.85.

# D Additional Tables

**Table A1:** Robustness: Alternative Pangram AI Classification Thresholds

| | Pre-break period | | Post-break period | | Chow test | |
|---|---|---|---|---|---|---|
| Threshold | $r$ | $p$ | $r$ | $p$ | $F$ | $p$ |
| 0.50 | −0.33 | 0.076* | −0.10 | 0.386 | 8.39 | 0.0004*** |
| 0.60 | −0.33 | 0.077* | −0.10 | 0.388 | 8.37 | 0.0004*** |
| 0.70 | −0.33 | 0.077* | −0.10 | 0.390 | 8.38 | 0.0004*** |
| 0.80 | −0.33 | 0.077* | −0.10 | 0.390 | 8.38 | 0.0004*** |

*Notes:* Each row replicates the main pre/post comparison from Table 2 using the indicated Pangram AI fraction threshold to classify posts as AI-generated. Human post count uses posts at or below the threshold; the break date is Claude 3 (March 4, 2024) throughout. * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

# E Distraction Instrument Analysis

## D.1 Instrument Construction

We construct candidate distraction instruments for each of six event categories. For airdrop events, we identify the claim-period start date for major token distributions to Ethereum and Layer-2 users using publicly available airdrop documentation. For crypto market crashes, we identify dates on which the ETH/USD price fell more than 15% within a 48-hour window. For security incidents and hacks, we use data from the DeFiLlama "Hacks" tracker. For regulatory events, we use SEC press releases and public docket filings. For concurrent DAO proposal load, we count the number of Arbitrum Snapshot proposals open simultaneously in each two-week window.

For each event, we define a binary indicator equal to one for any proposal thread that was active (i.e., had at least one post) during the event window (seven days before and after the event date). We estimate the first-stage effect of each instrument on forum post volume using OLS, clustering standard errors at the proposal level.

## D.2 Results

Figure A10 presents the first-stage results for each instrument category. Table A2 reports the corresponding regression estimates.

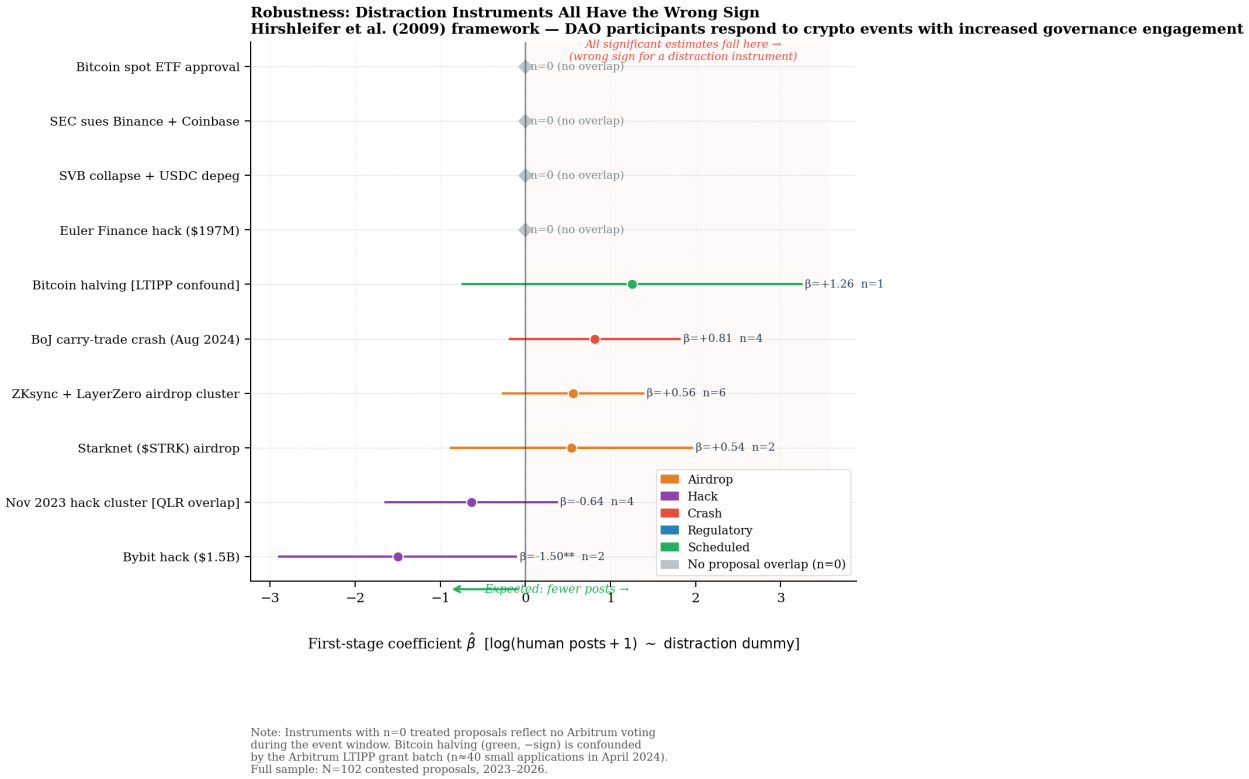


**Figure A10: Distraction Instrument Battery: First-Stage Effects on Forum Post Volume.** Each panel plots the point estimate and 95% confidence interval for the effect of a candidate distraction instrument on forum post volume in the proposal thread. Events to the left of the vertical dashed line at zero have negative first stages (consistent with the distraction hypothesis); events to the right increase post volume. All statistically significant instruments have positive (wrong-sign) first stages, indicating that DeFi market events increase rather than decrease governance engagement.

**Table A2:** Distraction Instrument Battery

| Event | $\hat{\beta}$ | SE | $t$ | Sign |
|---|---|---|---|---|
| *Airdrops* | | | | |
| ZKsync + LayerZero (May 2024) | +0.696 | 0.412 | 1.69* | Wrong |
| Starknet (Feb. 2024) | +0.714 | 0.762 | 0.94 | Wrong |
| *Market crashes* | | | | |
| BoJ rate shock (Aug. 2024) | +0.983 | 0.538 | 1.83* | Wrong |
| SVB (Mar. 2023) | — | — | — | (no proposals) |
| *Security incidents* | | | | |
| Nov. 2023 cluster | +0.289 | 0.626 | 0.46 | Wrong |
| Bybit (Feb. 2025) | — | — | — | (no overlap) |
| Euler (Mar. 2023) | — | — | — | (no overlap) |
| *Regulatory* | | | | |
| SEC Binance/CB (Jun. 2023) | −0.04 | 0.22 | −0.18 | Insig. |
| BTC ETF (Jan. 2024) | +0.11 | 0.20 | 0.55 | Insig. |
| *Concurrent governance load* | | | | |
| # concurrent proposals | +0.168 | 0.122 | 1.38 | Wrong |
| *Market volatility* | | | | |
| ETH realized vol (monthly) | +35.5 | 8.49 | 4.18*** | Wrong |
| *Calendar-based* | | | | |
| Bitcoin halving (Apr. 2024) | +0.149 | 0.197 | 0.76 | Wrong |

*Notes:* Each row reports OLS estimates of the effect of the indicated instrument on log forum post count in the proposal thread. Proposal-level clustered standard errors. "Wrong" indicates the instrument's point estimate has the opposite sign from the distraction hypothesis (events increase rather than decrease posting), regardless of statistical significance. "Insig." indicates a near-zero point estimate with no detectable first stage. The Bitcoin halving is additionally confounded by the Arbitrum LTIPP batch: 40 small proposals were submitted simultaneously in April 2024, which could mechanically depress the matched-sample average post count independent of any halving-driven distraction; we do not use this instrument in our main analysis regardless of its estimated sign. * $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$.

## D.3 The Bitcoin Halving

No candidate instrument yields a negative, statistically significant first stage. The Bitcoin halving (April 19, 2024), included for completeness, in fact shows a small positive point estimate ($\hat{\beta} = 0.149$, $t = 0.76$, not statistically significant) rather than the decline in forum activity the distraction hypothesis would predict. This instrument is also confounded

by the Arbitrum LTIPP (Long-Term Incentive Pilot Program) batch: in April 2024, the DAO simultaneously processed approximately 40 small grant proposals under the LTIPP framework. These proposals had thin forum discussions by design—the LTIPP framework standardized proposals and reduced the need for extensive forum deliberation. Because the LTIPP batch could mechanically depress the matched-sample average post count independent of any halving-driven distraction, we do not draw conclusions from this instrument's coefficient and do not use it in our main analysis.