ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

Adaptive Evidence-Based Multi-Agent Debate for Robust Financial Statement Fraud Screening

Amelia Singh *, Derek Das, Emily Li, Nathan Yoon

Isenberg School of Management, University of Massachusetts Amherst, USA

* Corresponding Author: amelias2023@umass.edu

Abstract

Multi-agent debate has recently been applied to financial statement fraud detection: specialist agents exchange evidence-grounded claims and a reconciled fraud-risk score is issued after a fixed number of rounds. Two questions remain open: when and for how long the agents should debate, and whether debate protects the fused score when one agent hallucinates red flags. We propose A-EMAD, an adaptive evidence-based multi-agent debate in which accounting, peer, temporal and forensic agents decompose their verdicts exactly into claims on nine fraud-risk constructs. Each round cross-examines every claim against the credibility-weighted evidence of the other agents under a proportional burden of proof. We prove that the credibility update is a sup-norm contraction, so the debate converges to a unique fixed point, and we stop each debate adaptively; a disagreement gate decides which filings are debated at all. On the public AAER-labelled panel of Bao et al. (146,045 firm-years), with out-of-time tests on 2003–2008 over four seeds, A-EMAD performs on par with stacking, RUSBoost and a fixed three-round debate (AUC 0.707 vs. 0.706, 0.714 and 0.708; no difference is significant) while using 54% fewer debate rounds than the fixed schedule. When one agent fabricates alert-level red flags, its credibility falls from 0.835 to 0.35, and A-EMAD loses roughly half as much AUC as stacking ($p < 10^{-4}$). A median pool is even more robust in AUC on the test years, but it trailed stacking by 0.016 AUC on clean development data and cannot indicate which agent or claim failed.

Keywords

Financial statement fraud, multi-agent debate, evidence grounding, credibility, robust aggregation, accounting and auditing enforcement releases.

1. Introduction

Accounting misstatements reach investors through audited filings, and only a fraction of them is ever detected [1]. Regulators and auditors therefore screen filings statistically, ranking firm-years so that a small investigation budget is spent where misstatement risk is highest. The literature has moved from ratio-based scores such as the M-score [2] and the F-score [3] to kernel methods [4] and to ensembles trained on raw financial-statement items, most prominently the RUSBoost model of Bao et al. [5, 6], and further to learners that combine accounting, market and audit variables [7] or financial text [8]. These detectors output a single score. They do not say which piece of evidence drives an alert, nor whether that evidence survives when it is checked against other views of the same firm.

Multi-agent debate offers such a check. Letting several model instances argue over several rounds improves factual accuracy [9] and counteracts the tendency of a single model to entrench its first answer [10], and evidence-based multi-agent debate (EMAD) forces each executor to tie its claims to retrieved evidence [11]. Yue et al. [12] transferred EMAD to listed-company statement fraud with FraudDebate-Agent, in which quantitative, narrative and

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

industry-peer agents feed two anchored executors, a groundedness penalty shrinks unsupported claims and an evidence graph penalizes contradictions before a report aligned with the auditing standard AS 2401 [13] is produced. Their analysis is instructive: the debate was a no-op on most filings and paid off on the minority where the modalities disagreed; the dominant failure was false escalation of benign filings whose boilerplate language was accepted as a red flag; and a learned stopping rule was left for future work.

These observations raise three questions. *When and for how long should agents debate?* A fixed schedule spends rounds on filings where the agents already agree, and without a convergence argument there is no principled point at which to stop. *Does debate protect the fused score against an agent that hallucinates?* Collaborating language-model agents can be swayed by a persuasive adversary [14], and a fraud screen must not let one agent's fabricated alarm push a benign firm into the review queue. *Does debate improve accuracy on clean inputs at all,* compared with standard late fusion [15]? Answering these questions with LLM executors over two decades of filings would be costly and hard to reproduce. We therefore study the debate mechanics with specialist quantitative agents whose claims are exact additive attributions, a setting in which the protocol can be analyzed formally, run on every firm-year and stress-tested under controlled hallucinations. The contributions are as follows.

A-EMAD, an LLM-agnostic adaptive evidence-based debate protocol. Four agents decompose their verdicts into claims on nine fraud-risk constructs; each round cross-examines every claim against the credibility-weighted evidence of the other agents under a proportional burden of proof, and agent credibility is the fixed point of this exchange. We prove that the exchange is a contraction (Proposition 1).

Adaptive control of the debate through a disagreement gate (when), convergence-based stopping (how long) and a credibility-aware arbiter (whom to trust). The adaptive policy matches the accuracy of a fixed three-round schedule with 54% fewer rounds.

A controlled evaluation on the public AAER panel of Bao et al. [5] under their out-of-time protocol, with thirteen baselines, ablations, DeLong tests and a hallucination-injection stress test modeled on the false-escalation failure of debate systems.

An account of what debate buys. On clean inputs its accuracy is statistically indistinguishable from stacking and RUSBoost; under fabricated escalation it identifies the hallucinating agent (credibility 0.35 vs. 0.835) and roughly halves the AUC loss of stacking, while a median pool is a strong but uninformative competitor.

2. Related Work

2.1. Statistical Detection of Misstatements

Early detectors model the probability of manipulation from a handful of accrual and performance ratios [2, 3]. Cecchini et al. [4] mapped raw accounting numbers into a financial kernel, and Bao et al. [5] showed that RUSBoost [6] trained on 28 raw items outperforms ratio-based logit models; they also introduced the rank-aware NDCG@k metric [16] to fraud detection together with an out-of-time protocol with serial-fraud correction, which we adopt. Bertomeu et al. [7] found that accounting variables become informative mainly through interactions with audit and market variables, and Craja et al. [8] combined ratios with the management discussion in a hierarchical attention network. All of these produce one score per firm-year, so evidence from different views is fused implicitly inside one model. Temporal logical attention networks for distributed-system anomalies provide a valuable methodological example of integrating temporal dependencies with interpretable structural signals for event screening [30].

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

2.2. Multi-Agent Debate and Argumentation

Debate between AI systems was proposed as a route to judgments that a weaker verifier can check [17]. In language models, multi-round debate improves reasoning and factuality [9], and Liang et al. [10] reported that an adaptive break of the debate is needed for good performance. Audit-LLM [11] introduced pair-wise EMAD for insider-threat detection, and FraudDebate-Agent extended it to cross-modal statement fraud. Amayuelas et al. [14] showed that a persuasive adversarial agent can degrade a debating collective. Computational argumentation supplies the formal backbone for such exchanges: abstract argumentation defines acceptability through attack relations [18], and quantitative argumentation debates compute the strength of a position from its supporting and attacking arguments [19]. Our cross-examination can be read as a gradual, bipolar evaluation of claims whose base strengths are model attributions.

2.3. Trust, Truth Discovery and Robust Aggregation

Truth discovery estimates source trustworthiness and claim truth by mutual reinforcement: a source is trustworthy if it provides many true claims, and a claim is likely true if trustworthy sources assert it [20]. The credibility iteration of A-EMAD follows this principle for signed, construct-level evidence and adds a convergence guarantee. Stacking [15] and mixtures of experts [21] learn how to weight base models, but stacking weights do not react to the content of an individual decision. Byzantine-robust aggregation replaces averages by order statistics to resist corrupted contributors [22]; we include median pooling as such a baseline.

3. Problem Setting and Evidence Agents

3.1. Task

Each observation is a firm-year i with an AAER-based label $y_i \in \{0, 1\}$. Following Bao et al. [5], the goal is to rank the firm-years of a test year so that a reviewer who inspects the top $k = 1\%$ finds as many misstatements as possible. All models for test year t are trained on fiscal years 1991 to $t - 2$.

3.2. Specialist Agents

The agents are quantitative evidence models with explicit decision rules, not separately prompted or fine-tuned large language models. The *Accounting agent* is a LightGBM classifier [23] on the 28 raw items and 14 ratios of the Bao et al. panel. The *Peer agent* adopts the Industry Peer agent of FraudDebate-Agent [12]: for each firm-year it retrieves $K = 20$ peers of the same fiscal year and measures the Ledoit–Wolf-shrunk [24] Mahalanobis distance of the firm's 14 ratios r_i from the peer mean μ_i,

$$D_i = \sqrt{(\mathbf{r}_i - \boldsymbol{\mu}_i)^\top \hat{\Sigma}_i^{-1} (\mathbf{r}_i - \boldsymbol{\mu}_i)}, \quad \hat{\Sigma}_i = (1 - \delta_i) S_i + \delta_i \frac{\mathrm{tr}(S_i)}{p} I_p \tag{1}$$

Where S_i is the peer sample covariance, $p = 14$ and δ_i the analytic shrinkage intensity. Because the public panel carries no industry codes, peers are retrieved by business-profile similarity, in the spirit of text-based network industries [25]: nearest neighbors in a rank-Gaussianized profile of log assets, asset turnover, cost of sales and the receivable, inventory, PP&E, cash and depreciation intensities. The agent sees D_i, the per-ratio peer z-scores, peer-relative levels of 13 asset-scaled items and the profile distance to the peers (29 features). The *Temporal agent* compares a firm with its own past: year-over-year changes of 25 asset-scaled items, growth of assets, sales, shares and price, the deviation of each ratio from its trailing three-year mean scaled by its own volatility, an aggregate self-deviation and the length of the history (45 features). The *Forensic agent* is an L2-regularized logit on the seven F-score variables of

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

Dechow et al. [3]. Tree agents use 100 trees with seven leaves. Out-of-fold predictions on the training window come from five-fold cross-validation grouped by firm, and every agent is calibrated by Platt scaling [26] on its out-of-fold margins.

3.3. Construct-Level Claims

TreeSHAP [27] decomposes the log-odds of a tree ensemble exactly, and a logit decomposes trivially, so every calibrated verdict ℓ_a splits into a base value b_a and feature contributions. Summing the contributions within each of the K = 9 constructs of Table I yields the claims of agent a:

$$\ell_a = b_a + \sum_{k=1}^{K} \phi_{a,k} \tag{2}$$

A positive $\varphi_{a,k}$ asserts that construct k points toward misstatement and a negative one asserts the opposite. The ontology maps every feature of every agent to exactly one construct, so the four agents speak a common language: an accrual claim of the Accounting agent and an accrual claim of the Temporal agent concern the same economic question seen through different reference frames, which is the triangulation an auditor performs when comparing a balance with prior years and with industry data.

Table I. Evidence ontology: the nine risk constructs

Construct	Examples of mapped evidence
REV Revenue & receivables	sales, receivables, Δreceivables, Δcash sales; peer and self deviations
ACC Accrual quality	working-capital and RSST accruals in all four views
INV Inventory & cost of sales	inventory, cost of goods sold, Δinventory
AST Asset structure	total and current assets, PP&E, soft assets, depreciation index
PRF Performance pressure	income, ΔROA, Δcash margin, EBIT, retained earnings
FIN Financing & leverage	debt levels and issuance, stock sales, preferred stock, interest
LIQ Liquidity & cash flow	cash, current liabilities, payables, taxes payable, ΔFCF
MKT Market valuation	price, shares, common equity, book-to-market
GLB Global anomaly	peer distance Di, profile distance, self-deviation, history length

4. Adaptive Evidence-Based Debate

4.1. Cross-Examination and Credibility

Claims of different agents live on different scales, so each claim is first normalized by the out-of-fold standard deviation $\sigma_{a,k}$ of the agent's claims on that construct:

$$e_{a,k} = \tanh\left(\frac{\phi_{a,k}}{c_u\, \sigma_{a,k}}\right) \tag{3}$$

Constructs that an agent never speaks on have $e_{a,k}$ = 0. In round r, claim (a, k) is cross-examined against the other agents' evidence on the same construct, weighted by their current credibility:

$$s_{a,k}^{(r)} = \operatorname{sgn}(\phi_{a,k}) \sum_{b \neq a} w_b^{(r-1)}\ e_{b,k} \tag{4}$$

A positive score means corroboration and a negative one contradiction. EMAD treats a claim as grounded only if cited evidence supports it [11]. We encode this as a proportional burden of proof, so that a strong claim needs strong corroboration and an uncorroborated claim is not taken at face value:

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

$$g_{a,k}^{(r)} = \sigma\left(\frac{s_{a,k}^{(r)} - \beta\,|e_{a,k}|}{T}\right) \tag{5}$$

An agent's credibility is the evidence-mass-weighted share of its claims that are grounded,

$$w_a^{(r)} = \sum_{k=1}^{K} \pi_{a,k}\; g_{a,k}^{(r)}, \qquad \pi_{a,k} = \frac{|\phi_{a,k}|}{\sum_{k'} |\phi_{a,k'}|} \tag{6}$$

Which mirrors the groundedness-weighted confidence of FraudDebate-Agent, here computed per construct and fed back into the next round. All agents start with $w_a^{(0)} = 1$. The exchange has the mutual-reinforcement structure of truth discovery [20]: agents whose claims are corroborated gain weight when the claims of the others are examined, and agents whose claims are contradicted lose it.

4.2. Convergence

Let F: $[0, 1]^M \to [0, 1]^M$ collect the updates (6) for M agents.

Proposition *1: For any $T > (M - 1)/4$, F is a contraction in the sup-norm,*

$$\|F(\mathbf{w}) - F(\mathbf{w}')\|_\infty \le \frac{M-1}{4T} \|\mathbf{w} - \mathbf{w}'\|_\infty \tag{7}$$

So the debate has a unique fixed point $w\star$, reached from any initialization at a geometric rate, and stopping at $\|w^{(r)} - w^{(r-1)}\|_\infty < \varepsilon$ guarantees $\|w^{(r)} - w\star\|_\infty \le q\varepsilon/(1 - q)$ with $q = (M - 1)/(4T)$.

Proof: The logistic function has slope at most $1/4$, $|e_{b,k}| \le 1$, and the burden term in (5) does not depend on w. Hence $|g_{a,k}(w) - g_{a,k}(w')| \le (4T)^{-1} \Sigma_{b \ne a} |w_b - w'_b| \le q\|w - w'\|_\infty$. Because the weights $\pi_{a,k}$ sum to one over k, the same bound holds for F_a. Banach's fixed-point theorem gives existence, uniqueness and the geometric rate, and the stopping bound follows from the triangle inequality.

The modulus q is a worst case. The effective modulus of a firm-year scales with $\Sigma_{b \ne a}|e_{b,k}|$, so filings with weak or one-sided evidence settle within one or two rounds while contested filings need more. A fixed number of rounds therefore over-serves the first group and may truncate the second, which motivates instance-wise stopping. With $M = 4$ and the temperature $T = 1.2$ used below, $q \le 0.625$.

4.3. Adaptive Policies

When. A firm-year is debated only if the agents disagree about whether it deserves an alert. With τ_a the $(1 - q_g)$ quantile of agent a's out-of-fold verdicts and $q_g = 10\%$, the gate is

$$\mathcal{G}_i = \mathbf{1}\left[\max_a(\ell_a - \tau_a) \ge 0 > \min_a(\ell_a - \tau_a)\right] \tag{8}$$

Undisputed filings, flagged by all agents or by none, keep their opening verdicts ($w_a = 1$, no revision). How *long.* A debated filing runs until $\|w^{(r)} - w^{(r-1)}\|_\infty < \varepsilon = 10^{-3}$, with a safety cap of 30 rounds that was never reached. *Whom to trust.* After the debate, escalating claims that failed cross-examination are withdrawn in proportion to their lack of groundedness,

$$\tilde{\ell}_a = \ell_a - \lambda \sum_k \max\left(\phi_{a,k}, 0\right) \max\left(0, 1 - 2g_{a,k}^{\star}\right) \tag{9}$$

And an L2-regularized logistic arbiter combines the revised verdicts with credibility-weighted, centered verdicts,

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

$$z_i = \theta_0 + \sum_{a=1}^{M} \left[\alpha_a \tilde{\ell}_a + \gamma_a \frac{w_a^\star}{\bar{w}_a} \left(\tilde{\ell}_a - \bar{\ell}_a \right) \right] \quad (10)$$

Where $\bar{w}_a$ and $\bar{\ell}_a$ are the mean credibility and verdict of agent a on the training window, and $w\star_a/\bar{w}_a = 1$ for undisputed filings. Only escalating claims are revised because false escalation is the failure that wastes review capacity. Gate thresholds, evidence scales, credibility norms and arbiter weights are fitted on out-of-fold debate states of the training window and never on the test year. Algorithm 1 summarizes the procedure and Fig. 1 the architecture.

Algorithm 1: A-EMAD for one firm-year i
Input: verdicts ℓ_a, claims $\varphi_{a,k}$; fitted τ_a, $\sigma_{a,k}$, $\bar{w}_a$, $\bar{\ell}_a$ and arbiter weights
1: if gate (8) is closed then $w\star \leftarrow 1$, $\ell_a \leftarrow \bar{\ell}_a$; go to 6
2: $w^{(0)} \leftarrow 1$; $r \leftarrow 0$
3: repeat $r \leftarrow r + 1$; compute $s^{(r)}$, $g^{(r)}$, $w^{(r)}$ by (4)–(6)
4: until $\|w^{(r)} - w^{(r-1)}\|_\infty < \varepsilon$ or $r = R_{max}$
5: $w\star \leftarrow w^{(r)}$; revise escalating claims by (9)
6: return score z_i by (10) and the claim report $\{(\varphi_{a,k}, g\star_{a,k}), w\star_a\}$

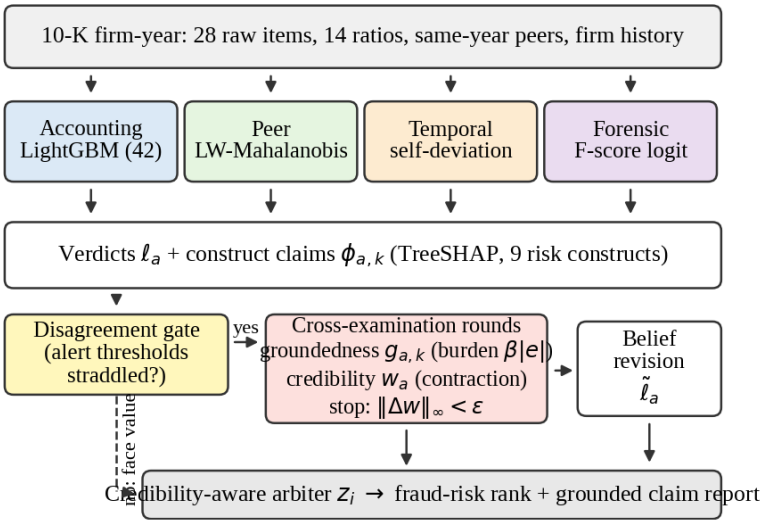


Figure 1. Architecture of A-EMAD. Four specialist agents emit verdicts and construct-level claims; a disagreement gate decides whether a firm-year is debated; cross-examination rounds update claim groundedness and agent credibility until convergence; escalating claims that fail are withdrawn before a credibility-aware arbiter issues the risk score and the claim report.

5. Experimental Setup

5.1. Data and Protocol

We use the public replication panel of Bao et al. [5], the same panel used by FraudDebate-Agent: 146,045 Compustat firm-years from 1990 to 2014 with 28 raw items, 14 ratios and AAER-based labels. Fiscal years 1991–2008 contain 108,399 firm-years and 837 misstatements (0.77%). We follow Bao et al.'s out-of-time protocol: for each test year $t \in \{2003, \ldots, 2008\}$, models are trained on 1991 to $t - 2$, and training frauds that belong to an AAER case also present in the test year are relabelled as non-fraud to prevent serial-fraud leakage. Table II lists the samples. All design decisions, including the A-EMAD hyperparameters, were made on the development years 2000

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

and 2001 (each trained on 1991 to $t - 2$), the period Bao et al. used for tuning; the test years were evaluated once with the frozen configuration. Results are averaged over four seeds that change the fold assignment, the LightGBM subsampling and the RUSBoost undersampling.

Table II. Out-of-time test samples

Test year	Training firm-years (frauds)	Test firm-years	Test frauds	k (1%)
2003	67,166 (402)	5,981	69	60
2004	73,233 (467)	5,934	58	59
2005	79,214 (538)	5,863	45	59
2006	85,148 (660)	5,908	33	59
2007	91,011 (720)	5,868	30	59
2008	96,919 (758)	5,612	26	56
Total	–	35,166	261	352

5.2. Baselines

Single models: the logit on the 14 ratios and RUSBoost on the 28 raw items, configured as in Bao et al.'s released code (300 trees, learning rate 0.1, minimum leaf size 5); each agent alone; and an early-fusion LightGBM on the union of all agent features (116 inputs). *Fusion layers* on the same agent outputs: mean and median pooling of standardized verdicts, stacking with an L2 logit [15], and a flat-evidence logit that receives all 36 construct claims in addition to the verdicts. *Debate variants:* a fixed-round debate that runs exactly $R = 3$ rounds for every firm-year, the schedule used by FraudDebate-Agent [12], with A-EMAD's update rules; and a debate-feature arbiter, which feeds raw debate statistics (corroboration, credibility and evidence-graph masses) to a higher-capacity model.

5.3. Hallucination Injection

To test whether debate protects the fused score, we corrupt the output of one agent on a random fraction $\rho \in \{5, 10, 20\}\%$ of the test firm-years, either always the Accounting agent, which is the most influential, or a randomly chosen agent. *False escalation* mimics the dominant failure reported for FraudDebate-Agent [12]: the agent fabricates red-flag claims on two randomly chosen constructs it covers, large enough to lift its verdict 0.5 log-odds above its own top-1% alert level. *Misattribution* replaces the agent's verdict and claims by those of another random firm-year, as when an agent confuses filings. Each setting is drawn three times per seed and year. All fusion layers are trained on clean data only. An oracle that knows the corrupted firm-years and drops the affected agent for them gives an upper reference.

5.4. Metrics and Implementation

We report AUC, NDCG@1% [16], and sensitivity and precision in the top 1%, computed per test year exactly as in Bao et al.'s evaluation code and averaged over 2003–2008. AUC differences on the pooled test predictions are tested with DeLong's test [28] using the fast algorithm of Sun and Xu [29], NDCG differences with a firm-year bootstrap (500 resamples), and robustness differences with Wilcoxon signed-rank tests over seed × year × draw units. A-EMAD uses $c_u = 2$, $T = 1.2$, $\beta = 1$, $\lambda = 1$, $q_g = 10\%$ and $\varepsilon = 10^{-3}$. These values were selected from 96 configurations by a rule fixed before testing: among the configurations within 0.001 development AUC of stacking on clean inputs, take the best rank sum of AUC and NDCG@1% under development-time corruption. The code uses Python, LightGBM 4.7 and scikit-learn 1.8 and runs on a single CPU core; the debate itself adds negligible run time compared with training the agents.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

6. Results

6.1. Screening Accuracy on Clean Inputs

Table III. Screening accuracy on the 2003–2008 test years (mean of yearly values; mean ± s.d. over four seeds)

Method	AUC	NDCG@1%	Sens@1% (%)	Prec@1% (%)
Logit, 14 ratios	0.691±0.000	0.029±0.000	3.7±0.0	2.0±0.0
RUSBoost, 28 raw items	0.714±0.004	0.017±0.002	2.2±0.2	2.0±0.2
Accounting agent	0.698±0.003	0.017±0.002	2.1±0.1	1.8±0.1
Peer agent	0.657±0.005	0.029±0.009	2.8±0.7	2.4±0.6
Temporal agent	0.636±0.005	0.035±0.003	3.8±0.4	2.9±0.3
Forensic agent	0.664±0.000	0.038±0.000	5.0±0.0	2.8±0.0
Early-fusion LightGBM	0.700±0.005	0.038±0.006	3.3±0.4	2.7±0.3
Mean pooling	0.707±0.003	0.029±0.007	2.9±0.6	2.3±0.5
Median pooling	0.706±0.004	0.032±0.006	3.4±0.6	2.8±0.4
Stacking (L2 logit)	0.706±0.002	0.023±0.006	2.4±0.4	1.9±0.3
Flat-evidence logit	0.691±0.004	0.023±0.005	2.3±0.1	1.9±0.1
Fixed-round debate, R = 3	0.708±0.003	0.026±0.006	2.8±0.3	2.1±0.3
A-EMAD (ours)	0.707±0.003	0.024±0.008	2.4±0.6	1.8±0.5

Table III reports the main comparison, and three findings stand out. First, the test years are harder than the development years: the fusion layers and RUSBoost reach AUCs between 0.69 and 0.71, whereas stacking and A-EMAD reached 0.753 on the development years 2000–2001. Second, A-EMAD (AUC 0.707 ± 0.003) is on par with stacking (0.706), median pooling (0.706) and the fixed three-round debate (0.708); none of these differences is significant in any seed (DeLong, all $p > 0.05$). RUSBoost has the highest mean AUC (0.714), but its advantage over A-EMAD on the pooled predictions is not significant in any seed either. A-EMAD is significantly better than the ratio logit (+0.039 pooled AUC, $p < 0.01$ in all seeds), mean pooling (+0.015, $p < 0.05$ in all seeds) and the Accounting agent alone (+0.009, $p < 0.05$ in three of four seeds). Third, NDCG@1% is low for every method, because each test year contributes only 26 to 69 frauds and 56 to 60 review slots, so one or two hits move the metric. Its seed-to-seed standard deviation reaches 0.009, and for seed 0 the bootstrap interval of A-EMAD minus stacking, [−0.012, 0.008], contains zero. The single-view Forensic and Temporal agents attain the highest NDCG@1% despite the lowest AUCs, a reminder that top-1% metrics computed on a handful of positives are fragile. Fig. 4 shows the share of frauds recovered as the budget grows.

6.2. Efficiency of the Adaptive Debate

The gate sends 30.1% of the firm-years to debate. Debated filings converge after 4.6 rounds on average (median 5, maximum 7), and every debated filing meets the ε criterion, whereas only 2.7% of the filings have converged after the three rounds of the fixed schedule. Overall, A-EMAD spends 1.39 rounds per firm-year, 54% fewer than the fixed schedule (3.0) and 69% fewer than running every filing to convergence (4.52), at the same AUC (Table IV). The left panel of Fig. 2 shows the geometric decay of the credibility updates predicted by Proposition 1. In an LLM instantiation, where each round costs one call per agent, the policy translates directly into fewer calls; FraudDebate-Agent reports about eight calls per filing at three rounds.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

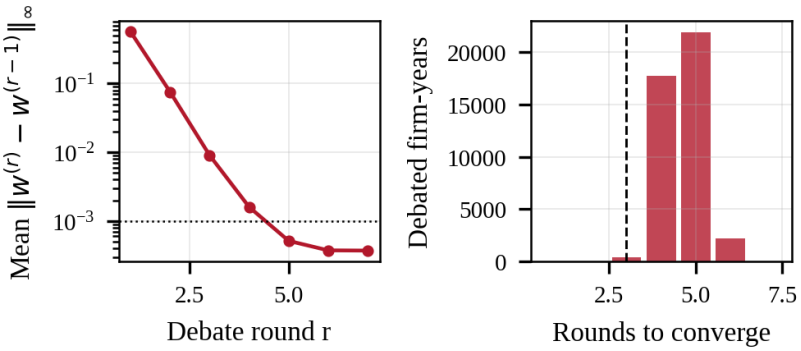


Figure 2. Debate dynamics on the test years. Left: mean sup-norm change of agent credibilities per round when every firm-year is debated (log scale); the dotted line marks ε. Right: rounds needed by the firm-years that A-EMAD debates; the dashed line marks the fixed schedule R = 3.

Table IV. Ablation, robustness and cost of A-EMAD (AUC@esc: AUC under 10% false escalation by the Accounting agent; rounds per firm-year on clean inputs)

Variant	AUC	NDCG@1%	AUC@esc	Rounds
A-EMAD (full)	0.707	0.024	0.691	1.39
w/o gate (debate all)	0.707	0.025	0.690	4.52
single round, R = 1	0.708	0.027	–	0.30
fixed R = 3, no gate	0.708	0.026	0.690	3.00
w/o burden of proof (β = 0)	0.707	0.026	0.682	1.31
symmetric revision	0.707	0.024	–	1.39
debate-feature arbiter	0.704	0.022	–	–
w/o Accounting agent	0.694	0.025	–	–
w/o Peer agent	0.706	0.021	–	–
w/o Temporal agent	0.702	0.018	–	–
w/o Forensic agent	0.704	0.020	–	–

6.3. Robustness to Hallucinating Agents

Under false escalation, the credibility of the hallucinating agent falls to 0.35 on the corrupted filings against 0.835 for honest outputs, and the gate routes 99% of the corrupted filings to cross-examination (30% of clean ones), because a single fabricated alarm opposed by the other agents is exactly a straddle of the alert thresholds. Fig. 3 and Table V show the consequence for ranking quality. When the Accounting agent escalates 10% of the firm-years, stacking falls from 0.706 to 0.676 AUC while A-EMAD falls from 0.707 to 0.691; at 20% the losses are 0.053 and 0.028. Across the six escalation settings A-EMAD loses between 38% and 55% as much AUC as stacking, and the paired difference is positive in 83–99% of the seed × year × draw units (Wilcoxon $p < 10^{-4}$ in every setting). The burden of proof is essential: without it (β = 0), AUC at 20% escalation by the Accounting agent drops to 0.662. The gains in NDCG@1% over stacking are small (0.000 to 0.004) and significant in only one of the six settings.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)



Figure 3. AUC under injected false escalation by the Accounting agent (left) and by a randomly chosen agent (right). Values at 0% are clean results; all fusion layers are trained on clean data. The oracle knows the corrupted firm-years.

Table V. Robustness at ρ = 10% corrupted firm-years: AUC (NDCG@1%)

Fusion layer	Clean	Escalation, Accounting	Escalation, random	Misattribution, Accounting
Stacking (L2 logit)	0.706	0.676 (0.025)	0.695 (0.021)	0.699 (0.022)
Mean pooling	0.707	0.693 (0.027)	0.697 (0.027)	0.704 (0.027)
Median pooling	0.706	0.702 (0.030)	0.704 (0.031)	0.703 (0.031)
Flat-evidence logit	0.691	0.671 (0.020)	0.682 (0.021)	0.687 (0.023)
Fixed-round debate	0.708	0.690 (0.027)	0.702 (0.024)	0.700 (0.025)
A-EMAD, β = 0	0.707	0.682 (0.027)	0.698 (0.026)	0.701 (0.024)
A-EMAD	0.707	0.691 (0.026)	0.703 (0.024)	0.699 (0.023)
Oracle	–	0.704 (0.024)	0.706 (0.024)	0.704 (0.023)

Median pooling is even more robust in AUC (0.702 under 10% escalation by the Accounting agent, close to the oracle at 0.704), because with four agents of similar accuracy a single outlier has no influence on the median. Two caveats temper this. On clean development data the median trailed stacking by 0.016 AUC (0.737 vs. 0.753), whereas A-EMAD matched stacking in both periods (0.753 on development), so its clean accuracy depends on how evenly the agents perform. And the median cannot tell an auditor which agent failed or which claim was withdrawn, while A-EMAD reports both. Misattribution is a weaker attack for every method (AUC losses of about 0.02 or less at 20%) and is harder to detect: a copied verdict is internally consistent, and its credibility drops only to 0.78–0.80.

6.4. Ablations

Table IV also reports ablations. Removing the Accounting agent costs the most (AUC 0.694), followed by the Temporal (0.702) and Forensic (0.704) agents, while removing the Peer agent changes AUC by only 0.001. The debate-feature arbiter (0.704) trails stacking on the development years because feeding raw corroboration statistics to a higher-capacity arbiter overfits the few hundred positives per training window, which is why A-EMAD applies groundedness structurally through (9). Symmetric revision, which also amplifies corroborated claims, leaves clean accuracy unchanged. A single round (R = 1) is as accurate as the converged debate on clean inputs, and dropping the gate changes neither clean accuracy nor robustness noticeably but more than triples the number of rounds.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

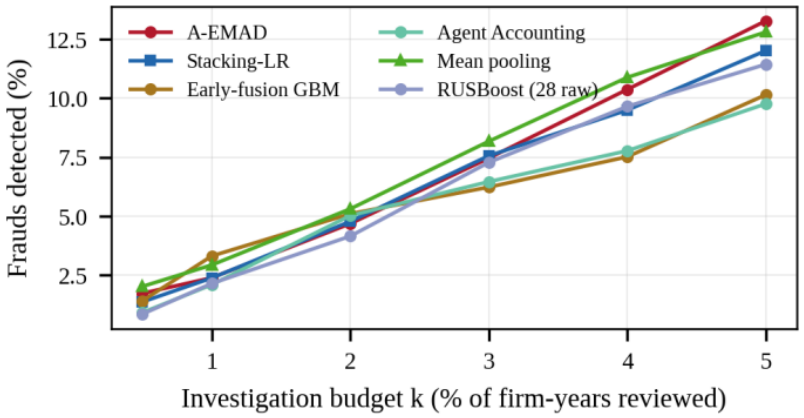


Figure 4. Share of frauds recovered as the investigation budget grows from 0.5% to 5% of the firm-years (test years, mean over four seeds).

6.5. Discussion

Why does debate not improve accuracy on clean inputs? In this panel the Accounting agent carries most of the signal (it receives the largest arbiter weights and its removal costs the most), and its strongest claims, which load on firm size and scale-related items, have no counterpart in the peer-relative and self-relative views. Cross-examination therefore has little to correct on honest outputs, consistent with debate being idle on most filings. What debate provides is a per-filing, per-claim record of which evidence survived and how far each agent can be trusted, and a defense against an agent that fabricates alarms. For LLM-based screening the results suggest three design rules: debate only contested filings and stop on convergence rather than after a fixed number of rounds; demand corroboration in proportion to a claim's strength; and keep an order-statistic fallback such as the median next to the arbiter, since it is hard to beat when the agents are of comparable quality.

The study has a defined scope. The agents are quantitative models rather than language models, so persuasion effects of LLM executors [14] are not captured, and the corruptions are synthetic; misattribution already shows that consistent but wrong evidence is hard to expose by cross-examination alone. The public panel contains no filing text and no industry codes, so the narrative modality is not modeled and peers are defined by business profiles. AAER labels record detected misstatements only [1], and all results come from one market and the 2003–2008 test period. Neuro-symbolic constraint verification for internet-finance transaction execution also illustrates how explicit rules can support verifiable language-model decisions in sensitive financial workflows [31].

7. Conclusion

A-EMAD turns evidence-based multi-agent debate into an adaptive, analyzable protocol: claims are exact construct-level attributions, cross-examination applies a proportional burden of proof, credibility is the unique fixed point of a contractive exchange, and a gate and a convergence rule decide when and for how long to debate. On two decades of AAER-labelled filings, A-EMAD matches the accuracy of stacking, RUSBoost and a fixed three-round debate with 54% fewer rounds than the latter, identifies agents that fabricate red flags and roughly halves the AUC loss that such agents inflict on stacking. On clean inputs, debate is statistically indistinguishable from the strongest fusion baselines, and a median pool remains a strong robust reference. Future work will instantiate the agents with language models over filing text, learn the burden of proof per construct, and combine credibility-weighted arbitration with order-statistic fallbacks.

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

References

[1] Dyck, A., Morse, A., & Zingales, L. (2010). Who blows the whistle on corporate fraud? The Journal of Finance, 65(6), 2213–2253. https://doi.org/10.1111/j.1540-6261.2010.01614.x

[2] Beneish, M. D. (1999). The detection of earnings manipulation. Financial Analysts Journal, 55(5), 24–36. https://doi.org/10.2469/faj.v55.n5.2296

[3] Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting material accounting misstatements. Contemporary Accounting Research, 28(1), 17–82. https://doi.org/10.1111/j.1911-3846.2010.01041.x

[4] Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010). Detecting management fraud in public companies. Management Science, 56(7), 1146–1160. https://doi.org/10.1287/mnsc.1100.1174

[5] Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. Journal of Accounting Research, 58(1), 199–235. https://doi.org/10.1111/1475-679X.12292

[6] Chen, J., Han, X., Zhang, X., & Yang, Y. (2023). Exact concept-level explanations for early detection of financial misstatements. Eurasia Journal of Science and Technology, 5(2), 34–48.

[7] Bertomeu, J., Cheynel, E., Floyd, E., & Pan, W. (2021). Using machine learning to detect misstatements. Review of Accounting Studies, 26(2), 468–519. https://doi.org/10.1007/s11142-020-09563-8

[8] Craja, P., Kim, A., & Lessmann, S. (2020). Deep learning for detecting financial statement fraud. Decision Support Systems, 139, 113421. https://doi.org/10.1016/j.dss.2020.113421

[9] Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. In Proceedings of the 41st International Conference on Machine Learning (pp. 11733–11763). PMLR.

[10] Liang, T., et al. (2024). Encouraging divergent thinking in large language models through multi-agent debate. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 17889–17904). https://doi.org/10.18653/v1/2024.emnlp-main.992

[11] Song, C., Ma, L., Zheng, J., Liao, J., Kuang, H., & Yang, L. (2024). Audit-LLM: Multi-agent collaboration for log-based insider threat detection. arXiv:2408.08902.

[12] Yue, X., Yang, J., & Liu, W. (2026). FraudDebate-Agent: A multi-agent LLM framework with an evidence-based debate mechanism for financial statement fraud detection. Mathematics, 14(15), 2695. https://doi.org/10.3390/math14152695

[13] Public Company Accounting Oversight Board. AS 2401: Consideration of Fraud in a Financial Statement Audit. PCAOB Auditing Standards.

[14] Amayuelas, A., Yang, X., Antoniades, A., Hua, W., Pan, L., & Wang, W. Y. (2024). MultiAgent collaboration attack: Investigating adversarial attacks in large language model collaborations via debate. In Findings of the Association for Computational Linguistics: EMNLP (pp. 6929–6948). https://doi.org/10.18653/v1/2024.findings-emnlp.407

[15] Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

[16] Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems, 20(4), 422–446. https://doi.org/10.1145/582415.582418

[17] Irving, G., Christiano, P., & Amodei, D. (2018). AI safety via debate. arXiv:1805.00899.

[18] Teng, D. (2025). RSRBench: Reproducibility and performance benchmarking for regional seismic risk workflows. Academic Journal of Architecture and Civil Engineering, 3(1), 40–52.

[19] Rago, A., Toni, F., Aurisicchio, M., & Baroni, P. (2016). Discontinuity-free decision support with quantitative argumentation debates. In Proceedings of the 15th International Conference on Principles of Knowledge Representation and Reasoning (pp. 63–73).

[20] Yin, X., Han, J., & Yu, P. S. (2007). Truth discovery with multiple conflicting information providers on the web. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1048–1052). https://doi.org/10.1145/1281192.1281309

ISSN: 3080-9843 (Print), ISSN: 3080-9851 (Online)

[21] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. Neural Computation, 3(1), 79–87. https://doi.org/10.1162/neco.1991.3.1.79

[22] Blanchard, P., El Mhamdi, E. M., Guerraoui, R., & Stainer, J. (2017). Machine learning with adversaries: Byzantine tolerant gradient descent. Advances in Neural Information Processing Systems, 30, 119–129.

[23] Ke, G., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

[24] Ledoit, O., & Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis, 88(2), 365–411. https://doi.org/10.1016/S0047-259X(03)00096-4

[25] Hoberg, G., & Phillips, G. (2016). Text-based network industries and endogenous product differentiation. Journal of Political Economy, 124(5), 1423–1465. https://doi.org/10.1086/688176

[26] Platt, J. C. (2000). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In A. J. Smola, P. Bartlett, B. Schölkopf, & D. Schuurmans (Eds.), Advances in Large Margin Classifiers (pp. 61–74). MIT Press.

[27] Lundberg, S. M., et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[28] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics, 44(3), 837–845. https://doi.org/10.2307/2531595

[29] Sun, X., & Xu, W. (2014). Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. IEEE Signal Processing Letters, 21(11), 1389–1393. https://doi.org/10.1109/LSP.2014.2337313

[30] Zhou, F. (2026). Moment-constrained side-force allocation for wind-assisted merchant ships: Hydrodynamic performance of retractable anti-leeway fins and flap rudders. Innovation and Technology Studies, 3(1), 27–41.

[31] Qin, Y., Rhee, M., Teng, D., Zhang, C., & Zi, B. (2026). Neuro-symbolic constraint verification for LLM-driven internet finance transaction execution. Scientific Reports. https://doi.org/10.1038/s41598-026-65233-w