MASCA: LLM-based Multi-Agent System for Credit Assessment

Gautam Jajoo
Kairosity
jajoo@kairosity.ai

Atharva Pandey
Kairosity
atpan@kairosity.ai

Pranjal Chitale
Microsoft Research

Abstract

Recent advancements in financial problem-solving have leveraged LLMs and agent-based systems, with a primary focus on trading and financial modeling. However, credit assessment remains an underexplored challenge, traditionally dependent on rule-based methods and statistical models. In this paper, we introduce MASCA, an LLM-driven multi-agent system designed to enhance credit evaluation by mirroring real-world decision-making processes. The framework employs a layered architecture where specialized LLM-based agents collaboratively tackle sub-tasks. We further present a signaling game theory perspective on hierarchical multi-agent systems, offering theoretical insights into their structure and interactions. Our paper also includes a detailed bias analysis in credit assessment, addressing fairness concerns. Experimental results demonstrate that MASCA outperforms baseline approaches, highlighting the effectiveness of hierarchical LLM-based multi-agent systems in financial applications, particularly in credit scoring.

1 Introduction

The financial domain has witnessed a major shift with the introduction of Large Language Models (LLMs), which have demonstrated potential across various financial tasks. Recent studies have showcased the capabilities of advanced LLMs, such as GPT-4, in financial text analysis [7], prediction tasks [18], and financial reasoning [11]. These models have proven particularly effective in processing and analyzing complex financial data, offering insights that were previously challenging to obtain through traditional methods.

Building upon the capabilities of LLMs, autonomous agents leveraging these models to tackle complex financial problems, have emerged as a powerful approach. Autonomous agents leverage LLMs to comprehend, generate, and reason with natural language, and this capability has been extended to the financial domain where they assist in tasks ranging from real-time market analysis to automated trading decisions [15]. Such agents have shown promise not only in processing large volumes of financial data but also in engaging in strategic and collaborative decision-making. However, one area where their potential remains underexplored is credit assessment, a domain that requires processing diverse data sources and navigating dynamic borrower-lender interactions.

Traditional credit assessment and scoring methods, while widely used, face several critical challenges: they rely heavily on historical credit data, overlooking alternative data sources that could provide a more comprehensive view of creditworthiness. Historical data can also inadvertently perpetuate existing biases leading to unfair lending practices [5]. Traditional models operate as “black boxes" in the decision-making processes of these systems, making it difficult to understand for consumers and

regulators to interpret [2]. Static models struggle to adapt quickly to changing economic conditions or evolving financial behaviors.

LLMs are uniquely positioned to address these challenges. Their ability to process unstructured and diverse data sources enables them to incorporate alternative data into credit assessments. Furthermore, their reasoning capabilities can enhance transparency by providing interpretable explanations for decisions. By integrating these models into a multi-agent framework, it becomes possible to create adaptive systems that respond dynamically to changing market conditions while promoting fairness and inclusivity.

This paper makes three contributions:

- **LLM-based multi-agent framework for credit assessment**: improving accuracy, fairness, and adaptability in decision-making.
- **Hierarchical multi-agent structure with Signaling Game Theory**: capturing borrower-lender strategic interactions and information flow.
- **Bias analysis in LLM-based credit assessment**: identifying and mitigating systemic risks in financial decision-making.

2 Related Work

LLMs have demonstrated strong capabilities across various financial applications [8], such as analyzing sentiment in financial news and social media [10], predicting market trends [4], interpreting financial time series data [19, 12], and finding factors that influence stock movements [13]. Their ability to extract relevant financial metrics and ratios from unstructured data has enhanced the speed and accuracy of financial assessments [14].

Multi-agent systems (MAS) have long been used in financial applications for their ability to model complex, dynamic environments [6], [1]. Agents in these systems operate autonomously, interact with each other, and collaborate to achieve shared goals. MAS has been applied to tasks such as algorithmic trading, fraud detection, and dynamic portfolio management.

There has been previous research work on LLM-based agents such as FinMem [21], a trading agent with layered memory to convert the insights gained from memories into investment decisions and FinAgent [22], which proposes a multimodal agent to reason for financial trading. Previous work on LLM-based multi-agent systems include financial decision-making [20] and trading systems [3], [15].

3 Methodology

Our hierarchical multi-agent system (MAS) for credit assessment (Figure 1) mirrors real-world credit teams by decomposing the task into specialized, modular agents. This design ensures Modularity, Explainability, Specialization, and Scalability (MESS). Each layer focuses on distinct aspects of credit evaluation, enabling transparent, accurate, and efficient decision-making.

Data Ingestion & Contextualization Layer: Transforms raw applicant data into structured profiles via three agents: **Data Analyst:** Aggregates, formats, and validates structured/unstructured data, **Contextualizer:** Builds applicant personas, integrating financial and behavioral insights, **Feature Engineer:** Derives key metrics (e.g., DTI, DAR, Credit Utilization, Employment Stability).

Multidimensional Assessment Layer: Conducts parallel evaluation of risks and rewards: The **Risk Team:** contains Risk Modeler (credit history, red flags), Income & Stability Analyst (income consistency, employment history, stress tests), Debt Analyst (debt burden, loan specifics). The **Reward Modeler:** highlights profitability, creditworthiness, and mitigating factors.

Strategic Optimization Layer & Decision Orchestrator: This layer contains **Risk-Reward Optimizer** which balances downsides and upsides via risk-reward ratios, weighted scoring, and scenario simulations. Finally, it synthesizes inputs from all layers to deliver the final approval decision.

Signaling Game Theory: Our framework models borrower-lender interactions as a signaling game. Borrowers send signals (credit history, income, loan details), while agents act as receivers, updating

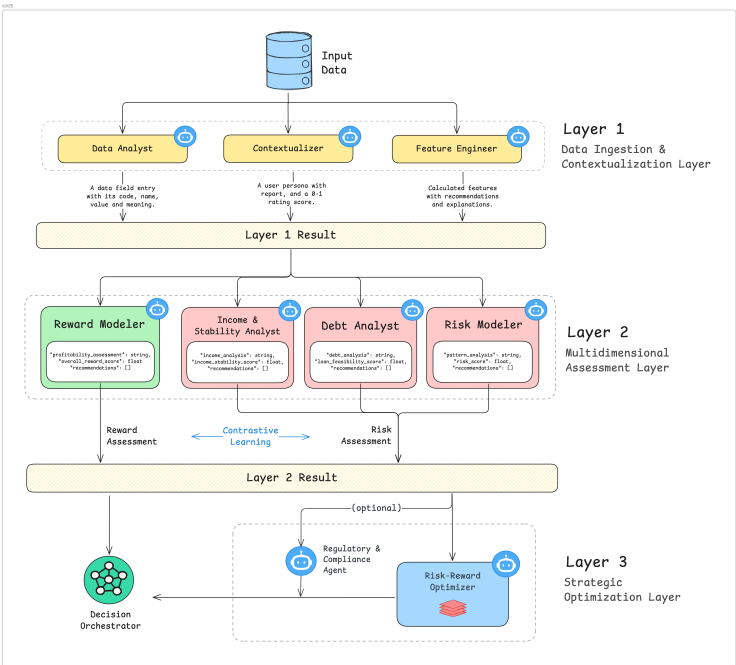


Figure 1: MASCA: The multi agent framework for credit assessment

beliefs based on signals. Hierarchical signaling enables higher-level agents (senders) to guide lower-level ones (receivers), promoting efficient exploration–exploitation trade-offs. The system converges toward Perfect Bayesian Equilibrium, refining beliefs and balancing risk-reward assessments. This mirrors how lenders dynamically infer creditworthiness from borrower signals. Theoretical analysis is presented at E.

4 Experiments

Dataset: We use credit scoring dataset based on the German Credit Dataset used in financial risk assessment provided by the TheFinAI where it benchmarks multiple datasets and tasks on various LLMs [17, 16]. Results on cra-lending dataset are reported in the Appendix B.

Models: Our experiments primarily use GPT [9] family models, specifically *gpt-4o* and *o3-mini*. We consider *o3-mini* to be more effective in reasoning tasks, making it a suitable choice for decision-making and overall assessment within our framework. We also show results using Llama3-70B model.

4.1 Baselines

We compare our framework against multiple baselines:

- **Zero shot performance**: Evaluate the input query with zero-shot baseline for comparison.
- **Chain of Thought(CoT)**: To assess reasoning ability, we prompt the model with *"Think step by step"* and analyze its response trace within the CoT framework.
- **Single Agent performing Multiple Tasks**: A single agent is assigned the responsibility of performing all subtasks.
- **Multi Agent System(OURS)**: We experiment with both homogeneous(same model) and heterogeneous setups(different models).

To evaluate the robustness of our proposed hierarchical framework, we introduce the following ablations:

Table 1: Performance metrics comparing various credit assessment approaches

Evaluation	Accuracy	Precision	Recall	F1 Score
Zero Shot (gpt4o)	45.5%	33.33%	67.69%	44.67%
Zero Shot (o3-mini)	44%	47.73%	59.43%	52.94%
Zero Shot (Llama3-70B)	41.5%	63.20%	27.30%	38.00%
Chain of Thought (gpt-4o)	36%	37.12%	52.13%	43.36%
Single Agent performing multitasks(gpt-4o)	42.5%	28.79 %	64.41%	39.79 %
Single Agent performing multitasks(o3-mini)	45.5%	43.18 %	62.64%	51.12 %
MultiAgent(OURS) (Llama3-70B & o3-mini)	48.50%	**67.90%**	41.70%	51.70%
MultiAgent(OURS) (gpt-4o)	51%	65.18%	55.3%	59.84%
MultiAgent(OURS) (o3-mini)	53.5%	65.12%	63.64%	64.37%
MultiAgent(OURS) (gpt-4o & o3-mini)	**60%**	65.48%	**83.33%**	**73.33%**

- **A single-level architecture with multiple agents:** All agents operate at the same level without a hierarchical structure, independently processing different aspects of the credit assessment task.
- **A two-level architecture with multiple agents:** Agents are organized into two layers, where the first layer performs the initial pre-processing and assessment, while the second layer performs risk and reward assessment.

5 Results and Discussion

Table 1 and 2 highlights that our **hierarchical MAS significantly outperforms baselines**. Combining *GPT-4o* and *o3-mini* yields 60% Accuracy (+15.5% over Zero-Shot GPT-4o), 83.33% Recall (+15.64%), and 73.33% F1 (+20.39%), while even MAS with *o3-mini* alone surpasses all non-MAS setups (+9.5% Accuracy, +13.25% F1). Baseline methods reveal clear limitations: Zero-shot GPT-4o achieves high Recall (67.69%) but low Precision (33.33%), showing over-approval bias. *o3-mini* favors Precision (47.73%) at Recall's cost (59.43%). CoT performs worst overall (36% Accuracy), suggesting reasoning chains propagate errors in credit tasks. **Single-agent multitasking proves suboptimal** (GPT-4o: 42.5% Accuracy, 28.79% Precision). Conflicting priorities hinder decision quality, whereas MAS cross-validation reduces false positives and balances Precision–Recall trade-offs. Ablations (Table 3) show flat architectures yield 9.23% lower F1 than hierarchical systems. ***Division of labor enables specialization, error correction, and refinement***, with later layers validating earlier assessments. Finally, **heterogeneous MAS combining GPT-4o's reasoning with o3-mini's** efficiency achieves the most robust and balanced predictions, reflected in superior Recall and F1 scores.

6 Biasness Perspective: Towards Fair Lending

We analyze potential gender and ethnicity biases in our multi-agent credit assessment system. For gender 2a, accuracy dropped from 65.22% for male applicants to 58.26% when gender was switched to female while keeping all other features constant. Out of 115 samples, 14 cases showed differing outcomes solely due to gender, indicating bias. Even with gender removed, accuracy further declined to 51.30%, suggesting indirect bias from correlated features. Loans approved for female applicants also showed lower confidence scores, though strong attributes like stable employment and positive credit history acted as counterbias factors.

For ethnicity 2b, performance varied across groups. African/Black applicants achieved the highest accuracy (57.5%) but still fell below ground truth (60%), while Asian applicants had the lowest (52.5%, -7.5% below ground truth). All groups underperformed ground truth recall (83.33%), indicating reduced ability to identify creditworthy applicants. The Asian approval rate (52.5%) reached only 87.5% of the baseline, nearing the disparate impact threshold under the 4/5th rule. Additionally, African/Black applicants showed higher recall (75.76%) but lower precision (65.36%), implying approval bias despite higher risk. Overall, results demonstrate that both gender and ethnicity significantly influence system outcomes, reinforcing the need for bias mitigation strategies.

References

[1] S. Abu-Hakima and B. Toloo. "A Multi-Agent Systems Approach for Fraud Detection in Personal Communication Systems". In: *Proceedings of the AAAI Workshop on Artificial Intelligence in Telecommunications* (1997), pp. 1–7.

[2] Philippe Bracke et al. *Machine Learning Explainability in Finance: An Application to Default Risk Analysis*. Working Paper 816. Bank of England, Aug. 2019. DOI: 10.2139/ssrn.3435104. URL: https://ssrn.com/abstract=3435104.

[3] Han Ding et al. *Large Language Model Agent in Financial Trading: A Survey*. 2024. arXiv: 2408.06361 [q-fin.TR]. URL: https://arxiv.org/abs/2408.06361.

[4] George Fatouros et al. "Can Large Language Models beat wall street? Evaluating GPT-4's impact on financial decision-making with MarketSenseAI". In: *Neural Computing and Applications* (Dec. 2024). ISSN: 1433-3058. DOI: 10.1007/s00521-024-10613-4. URL: http://dx.doi.org/10.1007/s00521-024-10613-4.

[5] ANDREAS FUSTER et al. "Predictably Unequal? The Effects of Machine Learning on Credit Markets". In: *The Journal of Finance* 77.1 (2022), pp. 5–47. DOI: https://doi.org/10.1111/jofi.13090. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jofi.13090. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.13090.

[6] Michael Kampouridis et al. "Multi-agent systems for computational economics and finance". In: *AI Communications* 35.4 (Sept. 2022). Ed. by Stefano V. Albrecht and Michael Woolridge, pp. 369–380. ISSN: 0921-7126. DOI: 10.3233/aic-220117. URL: http://dx.doi.org/10.3233/AIC-220117.

[7] Alejandro Lopez-Lira and Yuehua Tang. *Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models*. 2024. arXiv: 2304.07619 [q-fin.ST]. URL: https://arxiv.org/abs/2304.07619.

[8] Yuqi Nie et al. *A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges*. 2024. arXiv: 2406.11903 [q-fin.GN]. URL: https://arxiv.org/abs/2406.11903.

[9] OpenAI et al. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL]. URL: https://arxiv.org/abs/2303.08774.

[10] Yanxin Shen and Pulin Kirin Zhang. *Financial Sentiment Analysis on News and Reports Using Large Language Models and FinBERT*. 2024. arXiv: 2410.01987 [cs.IR]. URL: https://arxiv.org/abs/2410.01987.

[11] Guijin Son et al. *Beyond Classification: Financial Reasoning in State-of-the-Art Language Models*. 2023. arXiv: 2305.01505 [cs.CL]. URL: https://arxiv.org/abs/2305.01505.

[12] Hua Tang et al. *Time Series Forecasting with LLMs: Understanding and Enhancing Model Capabilities*. 2024. arXiv: 2402.10835 [cs.CL]. URL: https://arxiv.org/abs/2402.10835.

[13] Meiyun Wang, Kiyoshi Izumi, and Hiroki Sakaji. *LLMFactor: Extracting Profitable Factors through Prompts for Explainable Stock Movement Prediction*. 2024. arXiv: 2406.10811 [cs.CL]. URL: https://arxiv.org/abs/2406.10811.

[14] Xinlin Wang and Mats Brorsson. "Can Large language model analyze financial statements well?" In: *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*. Ed. by Chung-Chi Chen et al. Abu Dhabi, UAE: Association for Computational Linguistics, Jan. 2025, pp. 196–206. URL: https://aclanthology.org/2025.finnlp-1.19/.

[15] Yijia Xiao et al. *TradingAgents: Multi-Agents LLM Financial Trading Framework*. 2025. arXiv: 2412.20138 [q-fin.TR]. URL: https://arxiv.org/abs/2412.20138.

[16] Qianqian Xie et al. *PIXIU: A Large Language Model, Instruction Data and Evaluation Benchmark for Finance*. 2023. arXiv: 2306.05443 [cs.CL].

[17] Qianqian Xie et al. *The FinBen: An Holistic Financial Benchmark for Large Language Models*. 2024. arXiv: 2402.12659 [cs.CL].

[18] Qianqian Xie et al. *The Wall Street Neophyte: A Zero-Shot Analysis of ChatGPT Over MultiModal Stock Movement Prediction Challenges*. 2023. arXiv: 2304.05351 [cs.CL]. URL: https://arxiv.org/abs/2304.05351.

[19] Xinli Yu et al. *Temporal Data Meets LLM – Explainable Financial Time Series Forecasting*. 2023. arXiv: 2306.11025 [cs.LG]. URL: https://arxiv.org/abs/2306.11025.

[20] Yangyang Yu et al. *FinCon: A Synthesized LLM Multi-Agent System with Conceptual Verbal Reinforcement for Enhanced Financial Decision Making*. 2024. arXiv: 2407.06567 [cs.CL]. URL: https://arxiv.org/abs/2407.06567.

[21] Yangyang Yu et al. *FinMem: A Performance-Enhanced LLM Trading Agent with Layered Memory and Character Design*. 2023. arXiv: 2311.13743 [q-fin.CP]. URL: https://arxiv.org/abs/2311.13743.

[22] Wentao Zhang et al. *A Multimodal Foundation Agent for Financial Trading: Tool-Augmented, Diversified, and Generalist*. 2024. arXiv: 2402.18485 [q-fin.TR]. URL: https://arxiv.org/abs/2402.18485.

A Dataset Information

The results are primarily evaluated on two datasets:

- *german-flare* dataset: 200 test samples of the datatset.
- *cra-lending* dataset: 2690 test samples of the dataset.

There are 20 features/attributes(13 categorical, 7 numerical) present for each query in the test samples. The credit assessment classifies individuals as “good" or “bad" credit risks using historical customer data.

B Results on cra-lending dataset

Table 2: Performance metrics comparing various credit assessment approaches on cra-lending dataset

Evaluation	Accuracy	Precision	Recall	F1 Score
Zero Shot (gpt-4o)	60.5%	88.80%	58.60%	70.60%
Zero Shot (o3-mini)	61%	89.60%	58.60%	70.90%
MultiAgent (OURS) (gpt-4o & o3-mini as expert)	**66.67%**	87.90%	**68.10%**	**76.70%**

C Ablations study

Table 3: Ablations to evaluate the robustness of our proposed hierarchical framework

Evaluation	Accuracy	Precision	Recall	F1 Score
Single-level with multiple agents	46%	59.38%	57.58%	58.46%
Two-level with multiple agents	53.77%	63.70%	70.45%	66.91%

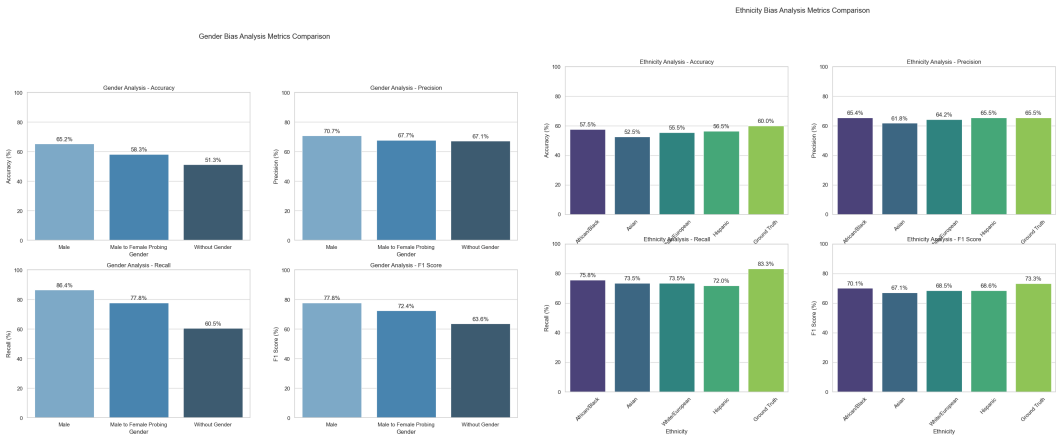
D Biasness Figures

E Signalling Game Theory in Multi Agent Setup

We define our hierarchical MASCA system formally as a signaling game:

$$(T, M, A, \mu, \sigma, U_S, U_R)$$

where:



(a) Gender Bias Analysis (b) Race Bias Analysis

Figure 2: Bias Analysis for Gender and Race

- T: Set of borrower types (creditworthy t_1 / risky t_2)
- M: Signals (from initial layer)
- A: Actions (approve/reject decisions)
- $\mu(t|m)$: Receiver's belief about borrower type t given signal m
- $\sigma(m|t)$: Sender's strategy mapping types to signals
- $U_S(t, a)$: Payoff function for sender (borrower)
- $U_R(t, a)$: Payoff function for receiver (MAS)

E.1 Mapping to MASCA

Element	Our Implementation
Sender	Borrower transmitting financial signals
Receiver	MASCA's hierarchical agents analyzing signals
Types (t)	Borrower creditworthiness: Creditworthy (t_1) vs. Risky (t_2)
Signals (m)	Processed outputs from initial layer
Actions (a)	Approve/Reject decisions by Decision Orchestrator
Beliefs ($\mu(t\|m)$)	Updated risk probabilities via Bayesian inference

E.2 Sender Strategy

A borrower of type $t \in \{t_1, t_2\}$ chooses signal m with probability:

$$\sigma(m|t) = \mathbb{P}(\text{Send } m|\text{Type } t)$$

Example:

- Creditworthy borrowers (t_1) send "Strong Credit History" (m_1) with $\sigma(m_1|t_1) = 1$.
- Risky borrowers (t_2) may mimic m_1 with $\sigma(m_1|t_2) = 0.3$.

E.3 Receiver Beliefs

Posterior probability of borrower type given signal m:

$$\mu(t|m) = \frac{\sigma(m|t) \cdot p(t)}{\sum_{t'} \sigma(m|t') \cdot p(t')}$$

where $p(t)$ is the prior (e.g., 60% creditworthy, 40% risky).

E.4 Receiver Strategy

Decision Orchestrator chooses action $a \in \{Approve, Reject\}$ to maximize expected utility:

$$a^*(m) = \arg\max_a \mathbb{E}[U_R(t,a)|m] = \sum_t \mu(t|m) \cdot U_R(t,a)$$

E.5 Equilibrium Conditions

Sequential Rationality:

- Senders: $\sigma(m|t)$ maximizes $\mathbb{E}[U_S(t,a)|m]$.
- Receivers: $a^*(m)$ optimizes utility given $\mu(t|m)$.

Belief Consistency: Posterior beliefs $\mu(t|m)$ align with sender strategies via Bayes' rule.

Example: If risky borrowers often mimic m_1, then $\mu(t_2|m_1)$ increases, reducing approval rates.

E.6 Equilibrium Types in MASCA

Equilibrium Type	MASCA Implementation	Empirical Evidence
Separating	t_1 and t_2 send distinct signals	83.33% recall [Table 1]
Pooling	Both types send identical signals	9.23% F1 drop [Table 2]

E.7 Case Study: Employment History Signaling Game

Agents:

- Sender: Contextualizer (Data Ingestion & Contextualization Layer)
- Receiver: Income Stability Analyst (Assessment Layer)

Payoff Matrix:

Sender Type	Signal	Receiver Action	Sender Payoff	Receiver Payoff
Stable Employment	Low variance	Approve	5	4
Unstable Employment	High variance	Reject	1	5
Unstable Employment	Mimics Low variance	Approve (Error)	3	-2

Table 4: Signaling game payoff matrix for employment stability and credit approval

E.8 Equilibrium Analysis

Separating Equilibrium:

$$\sigma(\text{Low}|\text{Stable}) = 1, \quad \sigma(\text{High}|\text{Unstable}) = 1$$

- Stable applicants truthfully signal low variance $\rightarrow$ Approval. - Unstable applicants truthfully signal high variance $\rightarrow$ Rejection.

Pooling Equilibrium Failure: If both types send "Low variance":

$$\mu(\text{Unstable}|m = \text{Low}) = \frac{0.4 \cdot 0.3}{0.6 \cdot 1 + 0.4 \cdot 0.3} = 16.7\%$$

Receiver utility for Approve:

$$0.833 \cdot 4 + 0.167 \cdot (-2) = 3.0 < 5$$

Thus Reject dominates, and pooling equilibrium collapses.