# CREDITAGENT: HIERARCHICAL MULTI-AGENT SYSTEMS FOR HIGH-STAKES FINANCIAL DECISION-MAKING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Credit risk assessment requires both accurate default prediction and a traceable decision process over heterogeneous financial evidence. Large Language Models (LLMs) can reason over such evidence in natural language. However, in high-stakes underwriting a monolithic LLM is exposed to noisy inputs, can miss rare but decisive risk cues, and offers limited control over policy-dependent decision thresholds. To address these issues, we introduce CREDITAGENT, a hierarchical credit review system with three stages: evidence governance, specialist risk analysis, and decision fusion. Layer 1 combines two coarse-filter agents with an XGBoost-guided field gate, Layer 2 runs six fine-tuned specialist agents that write structured assessments to a shared blackboard, and Layer 3 fuses these assessments with an attention-based scorer, a threshold policy trained by Group Relative Policy Optimization (GRPO), and deterministic hard-stop rules. Our central evaluation holds the adapted backbone, specialist-agent outputs, hard-stop rules, and data split fixed and varies only the fusion strategy. On a held-out set of 6,000 personal credit cases from a Chinese financial institution, GRPO fusion reaches 83.32% accuracy and a Business Efficiency Coefficient of 0.7647, compared with 78.40% and 0.6862 for average fusion, and the largest gain is on approval-boundary cases. The gain remains when the static fusion baselines receive tuned thresholds, and the hard-stop rules account for 0.41 points of accuracy. CREDITAGENT also outperforms cost-sensitive XGBoost and CatBoost with tuned thresholds on the same held-out cases, by 3.15 points in accuracy and 9.75 points on approval-boundary cases. These results indicate that structured agent hierarchies provide an interpretable framework for high-stakes financial decision support and address the limitations of monolithic model deployments.

## 1 INTRODUCTION

Automated credit risk assessment is a high-stakes problem that requires synthesizing structured quantitative signals with institution-specific qualitative evidence (Moro et al., 2016; Xu, 2024; Tavasoli et al., 2025). Traditional machine learning models such as XGBoost (Chen & Guestrin, 2016) remain strong on tabular benchmarks (Simonetto et al., 2024), but they do not natively express how macroeconomic context, borrowing history, inquiry velocity, and fraud cues should be combined in a committee-style credit decision (Song et al., 2023; Hoang & Wiegratz, 2023; Tian et al., 2024).

In practice, credit review therefore still depends heavily on human analysts and static rule stacks (Ben Abdelaziz & Mrad, 2023; Hong et al., 2023; Guo et al., 2024). As shown in Figure 1, this workflow is labor-intensive, difficult to scale, and vulnerable to inconsistent judgment. Large Language Models (LLMs) make richer reasoning possible in financial workflows (Yao et al., 2023; Li & Ouyang, 2024), but directly deploying a monolithic LLM for credit decisions still suffers from the following issues: The first challenge is ***hallucination***

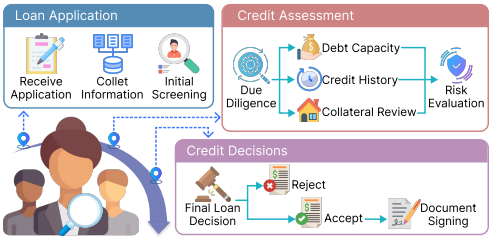



Figure 1: Human-in-the-loop credit review workflow at a real-world financial institution.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Taxonomy of related work. **Arch.**: Architecture. **Trans.**: Transaction Log. **CR**: Credit Report. **#Feat.**: Number of Features. **#Samp.**: Number of Samples. **STI**: Same-Tier Interaction. **H**: Hierarchical. **C**: Centralized. **SM**: Single Model.

| Work | Modality | Arch. | Trans. | CR | #Feat. | #Samp. | STI | Decision Logic |
|---|---|---|---|---|---|---|---|---|
| XGBoost (Chen & Guestrin, 2016) | Tabular | SM | × | × | 39 | 300k | – | Discriminative |
| Omega[2] (Majumdar et al., 2025) | Hybrid (RAG+GBDT) | C | × | × | 24 | 7.8k | – | Logistic |
| MASCA (Jajoo et al., 2025) | Textual | H | × | × | 20 | 200 | × | Majority Vote |
| MADeN (Wang et al., 2025b) | Dynamic | H | × | × | 6 | 975 | ✓ | Interaction-based |
| **Our Work** | **Multimodal** | **H** | ✓ | ✓ | **100+** | **6,000** | ✓ | **Score Fusion (Agent+GBDT)** |

***under noisy financial evidence*** (Jin et al., 2025). Credit files contain thousands of raw fields, many of which are redundant or irrelevant, and a model that reads all of them can ground its judgment on noise. Furthermore, this challenge is compounded by ***signal dilution***: rare but decisive cues such as fraud flags or inquiry spikes are buried inside a long prompt and receive little weight (Gandhar et al., 2024; Popa & Sîrbu, 2025). Finally, a single response offers ***weak traceability and weak threshold control***. It rarely exposes a stable committee-style reasoning trail for regulatory review (Agosto et al., 2025), and it does not expose an explicit operating threshold that can be aligned with the asymmetric cost of lending errors. Existing multi-agent financial systems address specialization, but they usually stop short of an explicit governance-to-reasoning-to-fusion decomposition, leaving the boundary between evidence filtering, specialist judgment, and final policy control underspecified (Table 1).

To address the aforementioned issues, we propose CREDITAGENT, a hierarchical multi-agent system for automated credit review that decomposes the pipeline into specialized and noise-resilient layers. The core idea of CREDITAGENT lies in mirroring a credit-committee workflow through three stages (Park et al., 2023), each of which targets one of the issues above. **Layer ❶: Dynamic Data Governance.** We couple filtering agents with machine-learning-based feature selection so that downstream reasoning agents receive compact but decision-relevant evidence (Fang et al., 2023). **Layer ❷: Multi-Agent Domain Expertise Review.** We organize six specialist agents around major credit-review dimensions (Madero, 2025). Each agent operates within a bounded scope, generates evidence-backed risk judgments, and publishes intermediate findings to a shared blackboard, reducing the diluted attention of a single-agent design. **Layer ❸: Policy-Aware Score Fusion.** The final layer synthesizes the specialist outputs into a binary decision through attention-based score fusion, a case-specific threshold learned by policy optimization, and deterministic hard-stop vetoes (Wright et al., 2025; McCracken et al., 2014; Amirizaniani et al., 2025). In this decomposition, Layer 1 primarily addresses noisy evidence, Layer 2 addresses signal dilution through specialization, and Layer 3 exposes an explicit adjudication step for operating-point control.

To operationalize this agentic infrastructure, CREDITAGENT is trained in two stages: (i) supervised fine-tuning (SFT) (Dong et al., 2024) on Seed-OSS-36B-Instruct[1] to instill structured credit reasoning in the specialist agents, and (ii) reinforcement learning with Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to align the final decision layer with an asymmetric business utility. On a real-world evaluation set of 6,000 Chinese personal credit cases, the resulting system achieves 83.32% accuracy and a Business Efficiency Coefficient of 0.7647. It also exceeds the evaluated zero-shot LLM baselines (Xie et al., 2024), including GPT-5.1[2], Claude-4.5[3], and Gemini-3-Flash[4]; since these proprietary models receive no task adaptation, they characterize untuned LLM behavior under the same input schema and prompting protocol.

The main contributions of this paper are: ❶ ***Evidence Governance.*** We introduce a hybrid machine-learning-and-LLM feature governance loop that filters noisy evidence before specialist reasoning, addressing hallucination under noisy evidence. Coarse-filter agents and an XGBoost-guided feature-selection loop reduce 2,515 raw fields to 100 expert-validated fields. ❷ ***Hierarchical Committee with Policy-Aware Fusion.*** We propose a hierarchical architecture for risk assessment that separates data governance, domain-specific reasoning, and final adjudication, addressing signal dilution through specialization. We further define a Business Efficiency Coefficient (BEC) to quantify the trade-off between default prevention and approval yield. ❸ ***Real-World Credit Dataset.*** We construct

[1] https://seed.bytedance.com/en/blog/seed-oss-open-source-models-release
[2] https://openai.com/index/gpt-5/
[3] https://www.anthropic.com/claude/sonnet
[4] https://deepmind.google/models/gemini/

a 6,000-case credit dataset with 100+ fields, including credit-bureau and anti-fraud records, and outcome labels for both approved and rejected applicants, which public credit datasets do not provide (Appendix D.8). ❹ ***Multi-Agent Fusion under Diverse Risk Regimes.*** We evaluate multi-agent fusion across three risk regimes and show that hierarchical fusion improves over single-agent and simpler fusion variants on the same adapted backbone, especially on approval-boundary cases where policy-aware fusion matters most, while yielding a structured audit trail by design.

## 2 Related Works

Automated credit assessment has moved from static statistical scoring (Tan & Xie, 2025; Zhang et al., 2015) to LLM-assisted and multi-agent reasoning systems (Naili & Lahrichi, 2022; Pathi, 2025; Sun et al., 2024). We organize the literature into three lines most relevant to our setting: mainstream machine learning for credit, multi-agent financial systems (Yu et al., 2024; Zhang et al., 2024a; Fatemi & Hu, 2024), and LLM-based decision fusion (Sun et al., 2025). An extended survey with per-work tables is given in Appendix J.

***Machine Learning and Multi-Agent Financial Reasoning.*** Credit scoring has long relied on Gradient Boosted Decision Trees (GBDTs), including XGBoost (Chen & Guestrin, 2016) and CatBoost (Prokhorenkova et al., 2019), which model nonlinear structure in tabular data well (Bhatore et al., 2020) but do not expose committee-style rationale over heterogeneous financial evidence (Wang et al., 2020b). For complex decision problems where a single LLM is unreliable, multi-agent systems introduce role specialization and collaborative reasoning (He et al., 2025a; Huang et al., 2025). CAMEL (Li et al., 2023a), AgentVerse (Chen et al., 2023), and AutoGen (Wu et al., 2024) study reusable communication protocols and generic role decomposition, whereas MetaGPT (Hong et al., 2023) focuses on software-style planning and execution. In finance, closer systems include manager–analyst hierarchies for investment reasoning (Yu et al., 2024), plan-and-act pipelines for coordinated task execution (Eren Erdogan et al., 2025), recommendation frameworks with financial specialization (Pathi, 2025), and structured asset-review systems for due-diligence workflows (Wan et al., 2024). However, these systems still under-specify the governance-to-decision chain most relevant to underwriting: they usually emphasize role decomposition, but they rarely separate evidence filtering, specialist reasoning, and final policy control as distinct stages, so the agents can reach "consensus on noise" (Dorri et al., 2018).

***LLM Reasoning and Decision Fusion.*** To reduce variance and hallucination, compound AI systems fuse multiple models or reasoning traces (Li et al., 2023b; Yao et al., 2025; Shi et al., 2025). Techniques such as LLM-Fusion (Li & Ouyang, 2024) and GRPO (DeepSeek-AI, 2024) improve alignment with multi-step reasoning, but they are usually evaluated as general aggregation or optimization tools. Our setting is narrower: specialist agents first emit structured scores and evidence chains, and the final layer performs weighted score fusion rather than simple voting (Chen et al., 2019). The relevant comparison is therefore not generic ensemble quality alone, but whether the fusion layer can preserve specialist structure while aligning the operating threshold with asymmetric lending utility.

***Positioning of* CreditAgent.** Relative to the closest prior systems, CreditAgent combines three components that rarely appear together: an explicit evidence-governance stage absent from Omega$^2$, specialist committee reasoning with shared-state and hard-stop logic beyond MASCA-style role decomposition, and business-aware fusion beyond interaction-focused systems such as MADeN. We therefore position CreditAgent as a governance-to-reasoning-to-fusion system rather than as a generic multi-agent wrapper for financial prompting; Table 1 summarizes this distinction.

## 3 Preliminary

We formalize credit review as a hierarchical decision function $\mathcal{F} : \mathcal{X} \to \{0, 1\}$, where $\mathcal{X}$ denotes heterogeneous credit evidence. Each case has a ground-truth default label $y \in \{0, 1\}$ ($1 =$ Default), and the system outputs a decision $\hat{Y} \in \{0, 1\}$ ($1 =$ Reject, $0 =$ Approve); symbols are listed in Appendix A. Each specialist agent $a_j$ emits a raw creditworthiness score $r_j \in \{1, \ldots, 10\}$ (larger is safer; Appendix O), and all fusion operates on the normalized risk $q_j = 1 - r_j/10$ (larger is riskier), so a larger $q_j$ always makes rejection more likely (worked example in Appendix D.1).

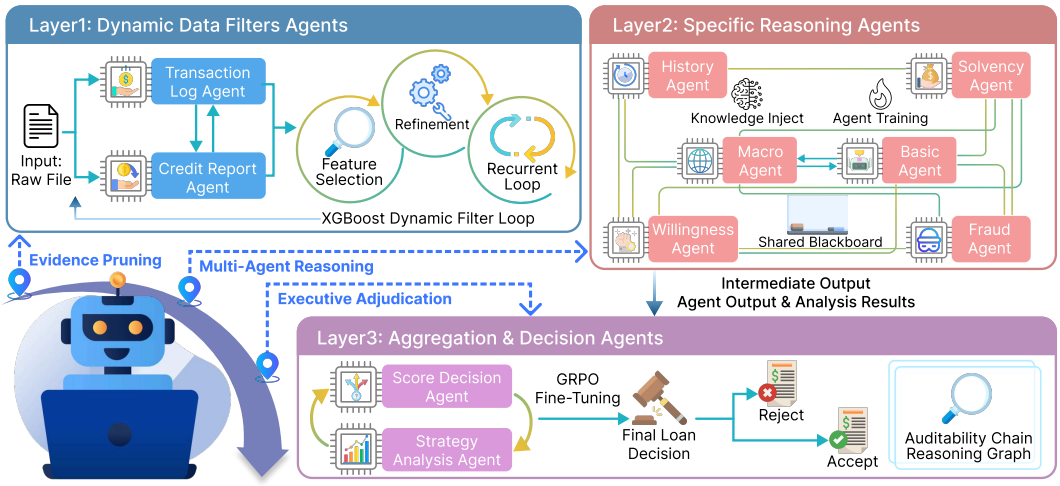



Figure 2: Overview of CREDITAGENT. **Layer 1** reduces the raw evidence with coarse-filter agents and a recurrent feature-selection loop. **Layer 2** runs a committee of six specialist agents over a shared blackboard. **Layer 3** fuses the specialist outputs into a traceable loan decision through score fusion, a learned threshold, and hard-stop rules.

## 4 METHODOLOGY

Figure 2 shows the three layers of CREDITAGENT, Algorithm 1 summarizes the end-to-end pipeline, and the rest of this section explains why each stage is needed and how it operates. The framework decomposes $\mathcal{F}$ into data governance ($\mathcal{Z}_1$), domain expertise ($\mathcal{Z}_2$), and decision fusion ($\mathcal{Z}_3$).

### 4.1 THREE-LAYER CREDIT ANALYSIS SYSTEM

***Layer 1: Dynamic Data Governance.*** The first layer acts as a noise-reduction bottleneck. Given a raw feature vector $\mathbf{x} \in \mathcal{X}$, it extracts a refined evidence vector $\mathbf{x}'$ that preserves decision-relevant signals while reducing token redundancy. Two auxiliary coarse-filter agents, a transaction-log agent $a_{\text{txlog}}$ and a credit-report agent $a_{\text{creport}}$, first discard fields that are malformed, duplicated, or semantically non-informative (*e.g.*, identifiers and session hashes). A Dynamic Data Filter Agent $G_\psi$ then runs a **Recurrent Feature Selection** loop with a GBDT predictor:

$$\mathbf{x}' = \mathbf{x} \odot \mathbb{I}\left[\text{XGB}(\mathbf{x}, y) > \gamma\right], \qquad (1)$$

where $\odot$ is the Hadamard product, $\mathbb{I}[\cdot]$ is the indicator function, $\text{XGB}(\mathbf{x}, y)$ is the feature-importance vector of an XGBoost model trained on historical default labels $y$, and $\gamma$ is a sparsity threshold. $G_\psi$ dynamically adjusts $\gamma$ so that the cardinality of $\mathbf{x}'$ satisfies the context-window constraints of the downstream LLM agents without losing risk signals. The loop reduces 2,515 raw fields to 150, and domain experts then validate the final 100 fields (Appendix C).

**Algorithm 1** CREDITAGENT pipeline

**Input:** Raw evidence $\mathbf{x}$; XGBoost model; filter agent $G_\psi$; coarse-filter agents $a_{\text{txlog}}, a_{\text{creport}}$; specialists $\mathcal{A}$; fusion head MLP; policy $\pi_\theta$.
**Output:** Decision $\hat{Y} \in \{0, 1\}$ (1=Reject), fused risk $S_{\text{risk}}$, audit chain $\{o_j\}_{j=1}^{6}$.
**/* Layer 1: Dynamic Data Governance */**
$\mathbf{x} \leftarrow \text{CoarseFilter}(\mathbf{x} \mid a_{\text{txlog}}(\mathbf{x}),\ a_{\text{creport}}(\mathbf{x}))$
$\mathbf{x}' \leftarrow \mathbf{x} \odot \mathbb{I}[\text{XGB}(\mathbf{x}, y) > \gamma]$ *Eq. (1)*
$S_{\text{shared}} \leftarrow \emptyset$
**/* Layer 2: Multi-Agent Domain Reasoning */**
**for** agent $a_j \in \mathcal{A}$ **do**
  $o_j = \{r_j, e_j\} \leftarrow a_j(P_j \mid S_{\text{shared}}),\ P_j = \mathcal{T}(\mathbf{x}'_j, g(\mathbf{x}'_j))$
  $S_{\text{shared}} \leftarrow [S_{\text{shared}} \| o_j]$
**end for**
**/* Layer 3: Score Fusion & Decision */**
$q_j \leftarrow 1 - r_j/10;\ \alpha_j \leftarrow \text{softmax}_j(\text{MLP}(E(e_j)))$
$S_{\text{risk}} \leftarrow \sigma\left(\sum_j \alpha_j q_j\right)$ *Eq. (2)*
$\tau^* \leftarrow \pi_\theta(S_{\text{shared}})$ *Eq. (3)*
**if** $S_{\text{risk}} \geq \tau^*$ **or** $\exists k \in \mathcal{V} : V_k = 1$ **then** *Eq. (4)*
  $\hat{Y} \leftarrow 1$ ▷ Reject
**else**
  $\hat{Y} \leftarrow 0$ ▷ Approve
**end if**
**return** $\hat{Y}, S_{\text{risk}}, \{o_j\}_{j=1}^{6}$

***Layer 2: Multi-Agent Domain Reasoning.*** Layer 2 maps $\mathbf{x}'$ to a set of specialist assessments. Because credit risk is multi-factor and interdependent, we use a **state-aware routing framework** over six specialist agents $\mathcal{A}$ (Macro, Basic, History, Solvency, Willingness, Fraud; Table 2). To ensure expert-level grounding, each agent is fine-tuned on our *Credit-Instruct* dataset (Section 4.3),

which aligns the backbone with expert reasoning paths, chain-of-thought style supervision, and domain-specific knowledge injection. The output of each agent $a_j$ is a tuple $o_j = \{r_j, e_j\}$, where $r_j \in \{1, \ldots, 10\}$ is the raw creditworthiness score for a specific dimension (*e.g.*, solvency), and $e_j \in \mathcal{E}^*$ is the **evidence chain** in natural language, with $\mathcal{E}$ the token vocabulary of the constrained JSON output space. The blackboard is updated as $S_{t+1} = [S_t \parallel o_{t+1}]$ with $S_0 = \emptyset$, so that downstream agents can condition their responses on upstream findings through the shared blackboard.

***Layer 3: Probabilistic Aggregation and Policy-Driven Gating.*** The final layer resolves conflicting specialist opinions and converts them into one decision. The Score Decision Agent computes a global risk score from the normalized risks $q_j$ using a weighted attention-based aggregation:

$$S_{\text{risk}} = \sigma\Big(\textstyle\sum_{j=1}^{n} \alpha_j \, q_j\Big), \qquad \alpha_j = \frac{\exp\big(\text{MLP}(E(e_j))\big)}{\sum_k \exp\big(\text{MLP}(E(e_k))\big)}, \tag{2}$$

where $n = |\mathcal{A}| = 6$, $E : \mathcal{E}^* \to \mathbb{R}^{d_e}$ is the frozen backbone embedding of an evidence chain, $\text{MLP}(\cdot)$ outputs a scalar relevance logit for each evidence chain, and $\sigma(z) = 1/(1 + e^{-z})$ is the sigmoid function. In practice, $E(\cdot)$ is the frozen backbone embedding used by the Layer 3 scorer, and the MLP is the only trainable fusion head in this step. Unlike static scoring systems, both the fusion weights $\alpha_j$ and the final decision threshold are input-dependent.

***Optimization via GRPO.*** To maximize business utility, we fine-tune the Layer 3 adjudicator with GRPO. We formulate threshold selection as a contextual bandit in which the policy $\pi_\theta$ predicts a single policy-adjusted threshold $\tau^* = \pi_\theta(S_{\text{shared}})$ from the shared committee state and the aggregate risk profile of each case. As shown in Eq. (3), we optimize $\pi_\theta$ with GRPO (Shao et al., 2024), which maximizes the relative reward within a group of sampled threshold actions:

$$J(\theta) = \mathbb{E}\Big[\textstyle\sum_{i=1}^{G} \min\big(\rho_i(\theta)\hat{A}_i,\ \text{clip}(\rho_i(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i\big)\Big], \tag{3}$$

where $G$ is the group size, $\rho_i(\theta)$ is the importance ratio between the current and reference policy for the $i$-th sampled threshold, $\hat{A}_i$ is the reward advantage relative to the group mean, and $\epsilon$ is the clipping radius. The reward is the cost-sensitive utility $\text{BEC}_{\text{pre}}$ (Eq. (7)) plus a stability regularizer. This optimization penalizes both overly conservative and overly aggressive gating, aligning the final decision with the trade-off between default prevention and approval volume.

***Decision rule and hard-stops.*** The final decision is

$$\hat{Y} = \begin{cases} 1 \text{ (Reject)}, & S_{\text{risk}} \geq \tau^* \ \vee \ \exists k \in \mathcal{V} : V_k = 1, \\ 0 \text{ (Approve)}, & \text{otherwise}, \end{cases} \tag{4}$$

where $\mathcal{V} \subseteq \mathcal{A}$ is the set of veto-enabled agents and $V_k \in \{0, 1\}$ is a deterministic hard-stop indicator for catastrophic patterns such as confirmed fraud, active litigation, or identity inconsistency. The veto is not triggered by free-form narrative text; it is triggered by explicit rule matches on the constrained JSON output of agent $k$. In the evaluated system the model-side hard-stop fires when $r_{\text{fraud}} \leq 3$ (Appendix D.5). This design also yields a structured decision trail, because the final decision can be traced back to the selected evidence, the specialist outputs, and any veto that fired.

### 4.2 Credit Evaluation Dataset

The evaluation set contains 6,000 real-world Chinese personal credit records designed to test default prediction and risk reasoning using structured credit-bureau data. It focuses on six core dimensions: borrower profiles, repayment capacity, willingness, credit history, fraud risk, and auxiliary signals. By highlighting multi-lender patterns and inquiry spikes, it evaluates whether a model can identify latent behavioral risk under realistic underwriting evidence. It also records outcomes for rejected applicants (Appendices D.7 and D.8); field definitions and statistics are given in Appendices M and N.

### 4.3 Risk-Oriented Agent Optimization

To address the semantic distortion that can arise when structured financial features are mapped into the linguistic reasoning space (Fazlija & Harder, 2022), we use three agent-level enhancements. First, a **risk-oriented textualization (ROT)** operator $\mathcal{T}$ maps the agent-relevant feature slice $\mathbf{x}'_j \subseteq \mathbf{x}'$ to a

structured prompt; it converts numeric evidence into bounded risk statements that the specialist agents can reason over more consistently. A risk-discretization function $g : \mathbb{R} \to \{\text{Low, Medium, High}\}$, defined by business-driven quantiles (Sadok et al., 2022), labels each value, and the prompt $P_j = \mathcal{T}(\mathbf{x}'_j, g(\mathbf{x}'_j))$ uses causal templates such as "Solvency risk [High], the Debt-to-Income ratio is 85%" to stabilize the model's internal reasoning. Second, we enforce an explicit risk reasoning mechanism through a tripartite *Credit-CoT* template: (i) *Risk Identification*, which extracts risk-bearing signals from the textualized input; (ii) *Evidence Aggregation*, which weighs those signals against offsetting positive indicators and upstream blackboard context; and (iii) *Decision Justification*, which emits a bounded JSON rationale together with the scalar score. Constrained decoding enforces the target JSON schema. Third, we build the *Credit-Instruct* dataset of $(input, CoT_reasoning, decision)$ triples (Appendix F) and adapt the backbone with PEFT via QLoRA (Han et al., 2024):

$$\min_{\theta} \mathcal{L}_{CE}(y, \hat{y}) + \beta \cdot \text{DPO}(O_A, O_B), \tag{5}$$

where $y$ is the ground-truth supervision label of the agent task, $\hat{y}$ is the predicted label, $\beta$ controls the strength of preference alignment, and $\text{DPO}(O_A, O_B)$ is a Direct Preference Optimization term over expert pairwise rankings of a preferred output $O_A$ and a dispreferred output $O_B$. The cross-entropy term supervises decision correctness, while the DPO term encourages preference-aligned reasoning traces.

### 4.4 Business Efficiency Coefficient

AUC (Huang & Ling, 2005) and F1 (Chicco & Jurman, 2020) do not reflect the asymmetric cost of lending errors. We therefore introduce the **Business Efficiency Coefficient (BEC)** as a task-specific utility metric and report it together with standard predictive metrics. The average cost over $N$ cases is

$$\text{AvgCost} = \frac{1}{N} \sum_{i=1}^{N} \left[ C_{\text{FN}}\, \mathbb{I}(y_i = 1, \hat{Y}_i = 0) + C_{\text{FP}}\, \mathbb{I}(y_i = 0, \hat{Y}_i = 1) \right], \tag{6}$$

with $C_{\text{FN}} = 3$ and $C_{\text{FP}} = 1$, following the cost ratio given by the collections expert in Appendix L. With $\text{MaxCost} = \max(C_{\text{FN}}, C_{\text{FP}})$,

$$\text{BEC}_{\text{pre}} = 1 - \frac{\text{AvgCost}}{\text{MaxCost}}, \qquad \text{BEC} = \max\left(0, \text{BEC}_{\text{pre}} \times (1 - \lambda_{\text{bec}} \cdot \sigma_{\text{agents}})\right), \tag{7}$$

where the second factor is a consistency penalty (Stein & Mizzi, 2007), $\sigma_{\text{agents}}$ is the standard deviation of the six normalized specialist risks $\{q_j\}$, and $\lambda_{\text{bec}} = 0.1$. This formulation rewards high utility only when the committee also reaches a relatively stable consensus; Appendix B gives the supporting utility analysis. Because the Layer 3 reward is built from $\text{BEC}_{\text{pre}}$, we report Accuracy and per-group accuracy as independent checks; Appendix D.3 decomposes BEC into its two factors.

## 5 Experiments

### 5.1 Experimental Settings

**Datasets.** The credit dataset of Section 4.2 is designed to stress "risk-boundary" decisions in real-world lending. Its 6,000 temporally later, expert-validated cases (2023) are partitioned into three semantic groups of 2,000 cases (Table 2, bottom) that emphasize different underwriting regimes: G1 (approved, later defaulted) and G3 (rejected, later defaulted) should be rejected, and G2 (non-defaulters in a low-approval region) should be approved (Appendix D.6). The SFT stage uses a separate 40,000-case corpus with a strict cutoff before the evaluation window, so the reported evaluation does not reuse adaptation data.

Table 2: Top: specialist agents in Layer 2. Bottom: evaluation groups (2,000 cases each).

| **Agent** | **Core Input** | **Focus** | **Risk Metric** |
|---|---|---|---|
| Macro | PESTEL | Systematic | Mkt. sensitivity |
| Basic | KYB/KYC | Identity | Fraud suspicion |
| History | Cr. bureau | Velocity | Default risk |
| Solvency | Cash flow | Assets/debt | Liquidity buf. |
| Willingness | Behavioral | Intent | Sincerity |
| Fraud | Anomaly | Adversarial | Synth. patterns |
| **Group** | **Business context** | | **Correct** |
| G1 | High-pass risk | | Reject |
| G2 | Low-approval recovery | | Approve |
| G3 | Rejection stability | | Reject |

**Baselines.** We compare against three proprietary LLMs (GPT-5.1, Claude-Sonnet-4.5, Gemini-3-Flash) called as zero-shot APIs, ten open-source

Table 3: Performance comparison on the 6,000-case evaluation set. Subscripts show absolute difference relative to CREDITAGENT (GRPO) baseline only within the adapted-backbone comparison block. The upper blocks serve as untuned reference models under a matched input protocol. All CREDITAGENT variants use the Seed-OSS-36B-Instruct backbone except CREDITAGENT(DS), which uses DeepSeek-R1-Distill-32B. DST is Dempster-Shafer theory fusion. Avg denotes the uniform average of the six specialist scores. Threshold-tuned Avg and DST are in Table 10. Std. is the standard deviation of G1–G3 accuracy. Underlined values indicate the best performance.

| Model Name | BEC ↑ | ACC.↑(%) | Std.↓ | Acc. (G1)(%) | Acc. (G2)(%) | Acc. (G3)(%) |
|---|---|---|---|---|---|---|
| Gemini-3-Flash-Preview | 0.5803 | 69.71 | 0.1040 | 57.53 | 68.67 | 82.94 |
| Claude-Sonnet-4.5 | 0.6138 | 70.99 | 0.1392 | 70.31 | 54.30 | 88.37 |
| GPT-5.1 | 0.5671 | 66.36 | 0.0767 | 56.09 | 68.48 | 74.52 |
| Seed-OSS-36B-Instruct | 0.5612 | 70.51 | 0.2496 | 86.84 | 35.25 | 89.45 |
| Llama-3.1-8B-Instruct | 0.3056 | 45.03 | 0.1612 | 34.28 | 67.82 | 33.00 |
| Llama-3.3-70B-Instruct | 0.5192 | 58.27 | 0.0560 | 50.63 | 60.29 | 63.89 |
| Qwen3-4B-Instruct | 0.5634 | 63.91 | 0.1479 | 67.75 | 44.19 | 79.79 |
| Qwen3-30B-Thinking | 0.4757 | 63.12 | 0.3207 | 84.25 | 17.80 | 87.30 |
| Qwen3-30B-Instruct | 0.5798 | 68.72 | 0.2119 | 81.50 | 38.85 | 85.80 |
| DeepSeek-R1-Dist-32B | 0.5433 | 68.85 | 0.2685 | 84.60 | 31.05 | 90.90 |
| DeepSeek-R1-Dist-8B | 0.5602 | 61.41 | 0.0758 | 57.91 | 54.38 | 71.94 |
| XuanYuan2-70B-Chat | 0.1508 | 39.37 | 0.3210 | 16.24 | 84.77 | 17.10 |
| Fin-R1 | 0.5382 | 57.64 | 0.1089 | 63.05 | 42.45 | 67.43 |
| CREDITAGENT(SFT)(DS) | $0.6013_{\downarrow 0.1634}$ | $72.23_{\downarrow 11.09}$ | $0.2170_{\uparrow 0.0958}$ | $84.45_{\downarrow 6.05}$ | $41.74_{\downarrow 24.51}$ | $90.50_{\downarrow 2.70}$ |
| CREDITAGENT(SFT)(Seed) | $0.6190_{\downarrow 0.1457}$ | $73.66_{\downarrow 9.66}$ | $0.2012_{\uparrow 0.0800}$ | $83.50_{\downarrow 7.00}$ | $45.62_{\downarrow 20.63}$ | $91.85_{\downarrow 1.35}$ |
| CREDITAGENT (SFT)+Avg | $0.6862_{\downarrow 0.0785}$ | $78.40_{\downarrow 4.92}$ | $0.1313_{\uparrow 0.0101}$ | $78.95_{\downarrow 11.55}$ | $62.05_{\downarrow 4.20}$ | $94.20_{\uparrow 1.00}$ |
| CREDITAGENT (SFT)+DST | $0.7055_{\downarrow 0.0592}$ | $76.55_{\downarrow 6.77}$ | $0.0550_{\downarrow 0.0662}$ | $72.10_{\downarrow 18.40}$ | $73.25_{\uparrow 7.00}$ | $84.30_{\downarrow 8.90}$ |
| **CREDITAGENT (GRPO)** | **0.7647** | **83.32** | 0.1212 | **90.50** | **66.25** | **93.20** |

LLMs including the financial models XuanYuan2 and Fin-R1 run locally, and tabular XGBoost and Random Forest on the same 100-field subset and temporal split. All LLM baselines receive the same input schema, output parsing target, and prompting protocol. We use the proprietary models mainly to characterize untuned LLM behavior on this task. The controlled comparison is within the CREDITAGENT family: SFT-only variants, SFT specialists with uniform average fusion (Avg), SFT specialists with Dempster–Shafer fusion (DST), and the full GRPO fusion. More details can be found in Appendix G; cost-sensitive tabular baselines with tuned thresholds are reported in Appendix D.4.

**Implementation Details.** All CREDITAGENT variants use Seed-OSS-36B-Instruct as the backbone except CREDITAGENT(DS), which uses DeepSeek-R1-Distill-32B. Specialists are adapted with QLoRA SFT plus DPO (Eq. (5)). The GRPO stage uses weight decay 0.1, clip ratio $(0.2, 0.28)$, top-$p$ 0.7, and $n = 4$ sampled responses per prompt. Avg and DST use the fixed threshold 6.0 on the 1–10 scale (Appendices G and D.1).

**Evaluation Metrics.** Standard metrics alone do not capture asymmetric financial error costs, so we adopt a triple-metric evaluation: **BEC**, **Accuracy**, and the **Stability Factor (Std.)**. BEC (Eq. (7)) converts false-negative and false-positive costs into a normalized $[0, 1]$ utility score with $C_{\mathrm{FN}} : C_{\mathrm{FP}} = 3 : 1$, so that larger values indicate better business utility under asymmetric lending risk. Accuracy, reported overall and per group (G1–G3), measures raw prediction quality, while Std. is the standard deviation of the three group accuracies and reflects regime imbalance.

**Evaluation Protocol.** The experiments support three progressively narrower claims. First, the upper blocks of Table 3 characterize how untuned general LLMs behave on this underwriting task under a matched input protocol. Second, the lower block provides the controlled architectural comparison by holding the adapted backbone fixed and varying only the fusion strategy. Third, the ablation and sensitivity studies test whether the gains come from the intended governance–specialization–fusion design rather than from backbone adaptation alone or brittle hyperparameter choices.

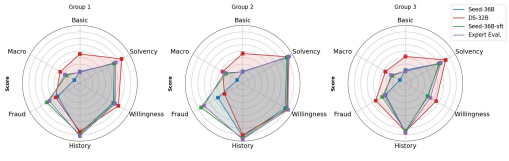



Figure 3: Output profiles of the six specialist agents.

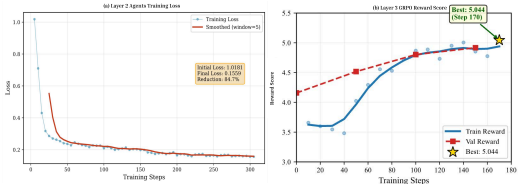



Figure 4: Layer 2 SFT loss and Layer 3 GRPO reward.

### 5.2 Empirical Results

**Performance Comparison.** Table 3 should be read in two parts. The upper blocks characterize untuned LLM behavior under a matched input protocol and serve as reference points rather than controlled architectural comparisons. Strong general-purpose models are often polarized across groups. For example, DeepSeek-R1-Distill-32B reaches 90.90% on the rejection-heavy G3 but 31.05% on the approval-sensitive G2, which lowers its BEC to 0.5433. The lower block is the controlled comparison. Within it, CreditAgent (GRPO) reaches BEC 0.7647 and accuracy 83.32%. GRPO has the highest G1 accuracy (90.50%) and keeps G2 at 66.25%, indicating that hierarchical reasoning improves approval-boundary decisions rather than only easy cases. On the same 100-field subset and temporal split, tabular XGBoost reaches BEC 0.6394, accuracy 78.47%, and G2 accuracy 50.68%, and Random Forest reaches 0.6488, 75.87%, and 37.14% with their default thresholds.

**Contribution of the fusion layer.** The comparison among SFT-only, average fusion, DST fusion, and GRPO fusion isolates the contribution of the final decision layer. Committee-level fusion already improves over the SFT-only variants, so the gain does not come from longer responses or larger context windows alone. Among fusion strategies, DST has the lowest regime imbalance (Std. 0.0550) but lower utility than GRPO. Further simplification checks lead to the same conclusion: removing shared context lowers BEC to 0.6412 in the parallel setting and 0.6497 in the sequential-no-context setting, a single structured-prompt LLM reaches 0.6836, and a hybrid that replaces specialist reasoning with XGBoost scores under the same Layer 3 reaches 0.7043. The DeepSeek-based SFT-only variant reaches BEC 0.6013 and accuracy 72.23%. These results indicate that the gain is not explained by backbone adaptation alone or by a stronger final scorer alone.

**Robustness of the fusion gain.** The gain persists with tuned Avg/DST thresholds (Appendix D.2), without hard-stops (Appendix D.5) or the BEC consistency term (Appendix D.3), and against tuned cost-sensitive XGBoost/CatBoost, mainly on G2 (Appendix D.4).

**Temporal Diagnostics.** Across the 2023 quarterly cohorts, which span two Loan Prime Rate adjustments, the BEC of CreditAgent stays within 0.759–0.770 while its approval rate rises from 60.1% to 67.9%; by contrast, the approval rate of XGBoost changes by at most $\pm 1.8$ points over the same period. This pattern is consistent with the intended role of macro-aware, policy-aware operating-point adjustment: the system preserves utility while adapting more flexibly around the approval boundary.

**Training Dynamics.** The training curves in Figure 4 are consistent with Table 3. Layer 2 SFT converges quickly, with the loss falling from 1.0181 to 0.1559 over 300 steps, which indicates that the specialist agents learn their role-specific reasoning patterns without unstable training. Layer 3 GRPO then pushes the operating point upward, with reward rising from about 3.6 to a peak of 5.044 at step 170 while validation reward tracks training reward closely. The final gains therefore come from optimization at the decision layer rather than from an overfit spike. Figure 3 compares the output profiles of the six specialist agents, and token counts and latency are analyzed in Appendix H.

### 5.3 Ablation Study

To measure the contribution of each hierarchical component, we remove specialist agents or the full Layer 2 committee while keeping Seed-OSS-36B-Instruct as the backbone (Table 4). Removing Layer 2 causes the largest degradation, lowering BEC from 0.7647 to 0.5089 and accuracy from 83.32% to 60.34%, which shows that the gain does not come mainly from a stronger final classifier. Removing the Macro, History, and Willingness agents lowers BEC by 0.1359 and G2 accuracy from 66.25% to 45.60%, so these agents carry a large part of the approval-boundary signal. Removing

Table 4: Ablation results of CREDITAGENT components. Values in parentheses give the change relative to FCA (Full CREDITAGENT). G1–G3 are group accuracies (%), and Std. is their standard deviation. M, H, W denote the Macro, History, and Willingness agents.

| Config. | BEC↑ | Acc.↑(%) | Std.↓ | G1 | G2 | G3 |
|---|---|---|---|---|---|---|
| FCA | **0.7647** | **83.32** | **0.1212** | **90.50** | **66.25** | **93.20** |
| *w/o* Macro | 0.6968 (↓0.0679) | 78.25 (↓5.07) | 0.1636 (↑0.0424) | 88.75 (↓1.75) | 55.15 (↓11.10) | 90.85 (↓2.35) |
| *w/o* {M,H,W} | 0.6288 (↓0.1359) | 73.85 (↓9.47) | 0.2010 (↑0.0798) | 85.25 (↓5.25) | 45.60 (↓20.65) | 90.70 (↓2.50) |
| *w/o* Layer 2 | 0.5089 (↓0.2558) | 60.34 (↓22.98) | 0.2142 (↑0.0930) | 72.31 (↓18.19) | 30.25 (↓36.00) | 78.45 (↓14.75) |

Table 5: Sensitivity of GRPO training and rollout parameters. Each column changes one parameter from the default (weight decay / clip ratio $(\epsilon_{low}, \epsilon_{high})$ / top-$p$ / samples per prompt $n$).

| | Default | Weight decay | | Clip ratio $\epsilon$ | | Top-$p$ | | Samples $n$ | |
|---|---|---|---|---|---|---|---|---|---|
| Setting | 0.1 / (0.2, 0.28) / 0.7 / 4 | 0.0 | 0.01 | (0.0, 0.2) | (0.3, 0.5) | 0.5 | 0.9 | 2 | 8 |
| Acc. (%) | **83.32** | 79.63 | 80.17 | 73.85 | 75.85 | 81.62 | 80.93 | 74.25 | 83.12 |

only the Macro agent has a smaller effect on accuracy but raises Std. from 0.1212 to 0.1636, which indicates that macro context mainly stabilizes the decision boundary across groups. Overall, the ablation supports a clean division of labor: Layer 1 filters evidence, Layer 2 extracts structured risk signals, and Layer 3 converts them into a utility-aware decision rule.

### 5.4 SENSITIVITY ANALYSIS

The sensitivity study asks whether the GRPO improvement is brittle or lies inside a sensible operating region (Table 5). All runs use the Seed-OSS-36B-Instruct backbone to isolate optimization choices. The results show a consistent pattern: moderate weight decay (0.1) outperforms no or weak regularization, the clip ratio $(0.2, 0.28)$ outperforms both narrower and wider trust regions, and top-$p$ 0.7 outperforms 0.5 and 0.9. The sample-size sweep shows the clearest utility–compute trade-off: moving from $n = 2$ to $n = 4$ raises accuracy from 74.25% to 83.32%, while $n = 8$ gives 83.12% at noticeably higher cost. Overall, the best configuration lies in an interpretable neighborhood of settings rather than at an isolated point.

### 5.5 EXPLORATORY FAIRNESS DISCUSSION

Fairness is essential in credit decisions (Purificato et al., 2023; Hurlin et al., 2026). On a 1,500-case subset, a post-hoc age×gender diagnostic shows a lower approval-rate disparity for CREDITAGENT than for XGBoost and GPT-5.1, which is consistent with heavier reliance on explicit financial-behavior evidence rather than shortcut subgroup correlations. The per-cell sample sizes are small, and a full evaluation of subgroup calibration, threshold effects, and applicant recourse (Karimi et al., 2022; De Toni et al., 2024) on larger cells is left to future work.

## 6 CONCLUSION

We present CREDITAGENT, a hierarchical multi-agent system that separates evidence governance, specialist reasoning, and final adjudication for high-stakes credit review. On a temporally later 6,000-case real-world evaluation set, it reaches 83.32% accuracy and improves approval-boundary performance within a controlled adapted-backbone comparison, while also achieving the best BEC among the evaluated fusion variants, including threshold-tuned baselines (Appendix D). The ablation, simplification, temporal diagnostics, and exploratory fairness analysis together suggest that these gains arise from coordinated specialist reasoning plus utility-aware fusion rather than from backbone adaptation alone. At the same time, the evidence remains limited to a single institutional credit setting, so broader validation across institutions, fairness criteria, regulatory regimes, and direct trace-quality evaluation remains future work (Appendix E).

## AI USE STATEMENT

In this work, we used generative AI tools to implement methods and to provide feedback on research methodology and experiments. We have not used generative AI tools to generate synthetic data sets, develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, or interpret results. Translation, dataset cleaning and reformatting, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create and edit software code, such as experiment and evaluation scripts, and to edit the text of the paper for readability. We have reviewed all AI-assisted work. All AI-assisted code was read, executed, and tested by the authors, and all reported results were checked against the saved experiment logs. Feedback from AI tools on methodology and experiments was treated as suggestions; the research questions, method design, and conclusions were determined by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of AI.

## ETHICS STATEMENT

Our research adheres to the ICLR Code of Ethics. The evaluation data are proprietary credit records provided by a financial institution; all records are desensitized before use, and personal identifiers are hashed. The system is intended to support human credit reviewers in non-real-time review, not to replace human accountability for lending decisions. Because credit decisions affect applicants, we report the fairness analysis as exploratory (Section 5.5).

## REPRODUCIBILITY STATEMENT

For reproducibility purposes, we have made our code available at https://anonymous.4open.science/r/Credit_Agents-AC45. Model configurations and inference settings are given in Appendix G, the feature-selection pipeline in Appendix C, the SFT data construction in Appendix F, the dataset fields and statistics in Appendices M and N, the label construction for rejected applicants in Appendix D.7, and the prompts in Appendix O. The raw credit records cannot be released because of privacy and regulatory constraints.

## REFERENCES

Arianna Agosto, Paola Cerchiello, and Paolo Giudici. Bayesian learning models to measure the relative impact of esg factors on credit ratings. *International Journal of Data Science and Analytics*, 20(2):357–368, 2025.

Maryam Amirizaniani, Adrian Lavergne, Elizabeth Snell Okada, Aman Chadha, Tanya Roosta, and Chirag Shah. Developing a framework for auditing large language models using human-in-the-loop. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, pp. 64–74, 2025.

Tajamul Ashraf, Amal Saqib, Hanan Ghani, Muhra AlMahri, Yuhao Li, Noor Ahsan, Umair Nawaz, Jean Lahoud, Hisham Cholakkal, Mubarak Shah, Philip Torr, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Salman Khan. Agent-x: Evaluating deep multimodal reasoning in vision-centric agentic tasks, 2025. URL https://arxiv.org/abs/2505.24876.

John E Baiden. The 5 c's of credit in the lending industry. *Available at SSRN 1872804*, 2011.

Fouad Ben Abdelaziz and Fatma Mrad. Multiagent systems for modeling the information game in a financial market. *International Transactions in Operational Research*, 30(5):2210–2223, 2023.

Siddharth Bhatore, Lalit Mohan, and Y Raghu Reddy. Machine learning techniques for credit risk evaluation: a systematic literature review. *Journal of Banking and Financial Technology*, 4(1): 111–138, 2020.

Martin Blackwell and Chris Sykes. The assignment of credit limits with a behaviour-scoring system. *IMA Journal of Management Mathematics*, 4(1):73–80, 1992.

Zeynep Boz, Dilek Gunnec, S. Ilker Birbil, and M. Kaan Öztürk. Reassessment and monitoring of loan applications with machine learning. *Applied Artificial Intelligence*, 32(9-10):939–955, 2018. doi: 10.1080/08839514.2018.1525517. URL https://doi.org/10.1080/08839514.2018.1525517.

A Colin Cameron and Pravin K Trivedi. 12 count data models for financial data. *Handbook of statistics*, 14:363–391, 1996.

Fei Chen, Wei Ren, et al. On the control of multi-agent systems: A survey. *Foundations and Trends® in Systems and Control*, 6(4):339–499, 2019.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, et al. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*, 2023.

Yongchao Chen, Jacob Arkin, Yang Zhang, Nicholas Roy, and Chuchu Fan. Scalable multi-robot collaboration with large language models: Centralized or decentralized systems? In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4311–4317. IEEE, 2024.

Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):6, 2020.

Giovanni De Toni, Paolo Viappiani, Stefano Teso, Bruno Lepri, and Andrea Passerini. Personalized algorithmic recourse with preference elicitation. *Transactions on Machine Learning Research*, 2024.

DeepSeek-AI. Deepseek-v3 technical report. *arXiv preprint*, 2024. Focus on GRPO and reasoning optimization.

Darshan Deshpande, Varun Gangal, Hersh Mehta, Jitin Krishnan, Anand Kannappan, and Rebecca Qian. Trail: Trace reasoning and agentic issue localization. *arXiv preprint arXiv:2505.08638*, 2025.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. How abilities in large language models are affected by supervised fine-tuning data composition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 177–198, 2024.

Ali Dorri, Salil S Kanhere, and Raja Jurdak. Multi-agent systems: A survey. *Ieee Access*, 6: 28573–28593, 2018.

Lutfi Eren Erdogan, Nicholas Lee, Sehoon Kim, Suhong Moon, Hiroki Furuta, Gopala Anumanchipalli, Kurt Keutzer, and Amir Gholami. Plan-and-act: Improving planning of agents for long-horizon tasks. *arXiv e-prints*, pp. arXiv–2503, 2025.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023.

Sorouralsadat Fatemi and Yuheng Hu. Finvision: A multi-agent framework for stock market prediction. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pp. 582–590, 2024.

Bledar Fazlija and Pedro Harder. Using financial news sentiment for stock price direction prediction. *Mathematics*, 10(13):2156, 2022.

Duanyu Feng, Yongfu Dai, Jimin Huang, Yifang Zhang, Qianqian Xie, Weiguang Han, Zhengyu Chen, Alejandro Lopez-Lira, and Hao Wang. Empowering many, biasing a few: Generalist credit scoring through large language models. *arXiv preprint arXiv:2310.00566*, 2023.

Akash Gandhar, Kapil Gupta, Aman Kumar Pandey, and Dharm Raj. Fraud detection using machine learning and deep learning. *SN Computer Science*, 5(5):453, 2024.

Fatemeh Ghezloo, Mehmet Saygin Seyfioglu, Rustin Soraki, Wisdom O. Ikezogwo, Beibin Li, Tejoram Vivekanandan, Joann G. Elmore, Ranjay Krishna, and Linda Shapiro. Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology, 2025. URL https://arxiv.org/abs/2502.08916.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024.

Junda He, Christoph Treude, and David Lo. Llm-based multi-agent systems for software engineering: Literature review, vision, and the road ahead. *ACM Transactions on Software Engineering and Methodology*, 34(5):1–30, 2025a.

Yexiao He, Ang Li, Boyi Liu, Zhewei Yao, and Yuxiong He. Medorch: Medical diagnosis with tool-augmented reasoning agents for flexible extensibility, 2025b. URL https://arxiv.org/abs/2506.00235.

Daniel Hoang and Kevin Wiegratz. Machine learning methods in finance: Recent applications and prospects. *European Financial Management*, 29(5):1657–1701, 2023.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *The twelfth international conference on learning representations*, 2023.

Zican Hu, Zongzhang Zhang, Huaxiong Li, Chunlin Chen, Hongyu Ding, and Zhi Wang. Attention-guided contrastive role representations for multi-agent reinforcement learning. *arXiv preprint arXiv:2312.04819*, 2023.

Jin Huang and Charles X Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310, 2005.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, et al. Biomni: A general-purpose biomedical ai agent. *biorxiv*, 2025.

Christophe Hurlin, Christophe Pérignon, and Sébastien Saurin. The fairness of credit scoring models. *Management Science*, 72(1):406–425, 2026.

Gautam Jajoo, Pranjal A Chitale, and Saksham Agarwal. Masca: Llm based-multi agents system for credit assessment, 2025. URL https://arxiv.org/abs/2507.22758.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. Discoveryworld: A virtual environment for developing and evaluating automated scientific discovery agents. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 10088–10116. Curran Associates, Inc., 2024. doi: 10.52202/079017-0324.

Haibo Jin, Peiyan Zhang, Peiran Wang, Man Luo, and Haohan Wang. From hallucinations to jailbreaks: Rethinking the vulnerability of large foundation models. *arXiv preprint arXiv:2505.24232*, 2025.

Li Kang, Xiaocun Wang, Lei Zhao, Xijing Wang, and Sheng Xiang. Multimodal large language model for enterprise credit assessment with power data enhancement. In *Proceedings of the 2025 International Conference on Economic Management and Big Data Application*, pp. 785–791, 2025.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023a.

Jiangtong Li, Yuxuan Bian, Guoxuan Wang, Yang Lei, Dawei Cheng, Zhijun Ding, and Changjun Jiang. Cfgpt: Chinese financial assistant with large language model. *arXiv preprint arXiv:2309.10654*, 2023b.

Xiangci Li and Jessica Ouyang. Related work and citation text generation: A survey. *arXiv preprint arXiv:2404.11588*, 2024.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *arXiv preprint arXiv:2310.02170*, 2023.

Marcos R Machado, Daniel Tianfu Chen, and Joerg R Osterrieder. An analytical approach to credit risk assessment using machine learning models. *Decision Analytics Journal*, pp. 100605, 2025.

Armando Madero. Specialized multi-agent neural architecture for enhanced reasoning and minimal intervention in ai systems. *Available at SSRN 5534058*, 2025.

Chitro Majumdar, Sergio Scandizzo, Ratanlal Mahanta, Avradip Mandal, and Swarnendu Bhattacharjee. A large language model for corporate credit scoring, 2025. URL https://arxiv.org/abs/2511.02593.

Lance M McCracken, Philomena DaSilva, Beth Skillicorn, and Richard Doherty. The cognitive fusion questionnaire: A preliminary study of psychometric properties and prediction of functioning in chronic pain. *The Clinical journal of pain*, 30(10):894–901, 2014.

Sérgio Moro, Paulo Cortez, and Paulo Rita. An automated literature analysis on data mining applications to credit risk assessment. *Artificial Intelligence in Financial Markets: Cutting Edge Applications for Risk Management, Portfolio Optimization and Economics*, pp. 161–177, 2016.

Sigit Mulyanto, Dwika Yonia, and Bambang Sutejo. Finance loan risk assessment using machine learning for credit eligibility prediction and model optimization. *IJISTECH (International Journal of Information System and Technology)*, 8(5):303–311, 2025. ISSN 2580-7250. doi: 10.30645/ijistech.v8i5.376. URL https://www.ijistech.org/ijistech/index.php/ijistech/article/view/376.

Maryem Naili and Younes Lahrichi. The determinants of banks' credit risk: Review of the literature and future research agenda. *International Journal of Finance & Economics*, 27(1):334–360, 2022.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Sai Prashanth Pathi. A multi-agent framework for personalized credit recommendations. *International Journal of Multidisciplinary Research and Growth Evaluation*, 2025.

Alexandru Popa and Tiberiu-Iulian Sîrbu. From fragmentation to cohesion: An llm-based iterative approach to ontology and knowledge graph refinement. In *2025 25th International Conference on Control Systems and Computer Science (CSCS)*, pp. 651–654. IEEE, 2025.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. Catboost: unbiased boosting with categorical features, 2019. URL https://arxiv.org/abs/1706.09516.

Sreenivasulu Puli, Nagaraju Thota, and ACV Subrahmanyam. Assessing machine learning techniques for predicting banking crises in india. *Journal of Risk and Financial Management*, 17(4):141, 2024.

E. Purificato, F. Lorenzo, F. Fallucchi, and E. W. De Luca. The use of responsible artificial intelligence techniques in the context of loan approval processes. *International Journal of Human–Computer Interaction*, 39(7):1543–1562, 2023.

Pfarelo Raliphada, Micheal O Olusanya, and Seun Olukanmi. Optimizing credit risk model classification using bert, roberta, and llama 3.2. In *IEEE EUROCON 2025-21st International Conference on Smart Technologies*, pp. 1–6. IEEE, 2025.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory, 2025. URL https://arxiv.org/abs/2501.13956.

Mrinal Rawat, Ambuje Gupta, Rushil Goomer, Alessandro Di Bari, Neha Gupta, and Roberto Pieraccini. Pre-act: Multi-step planning and reasoning improves acting in llm agents, 2025. URL https://arxiv.org/abs/2505.09970.

Hicham Sadok, Fadi Sakka, and Mohammed El Hadi El Maknouzi. Artificial intelligence and bank credit analysis: A review. *Cogent Economics & Finance*, 10(1):2023262, 2022.

Mario Sanz-Guerrero and Javier Arroyo. Credit risk meets large language models: Building a risk indicator from loan descriptions in p2p lending. *arXiv preprint arXiv:2401.16458*, 2024a.

Mario Sanz-Guerrero and Javier Arroyo. Credit risk meets large language models: Building a risk indicator from loan descriptions in peer-to-peer lending. *Available at SSRN 4979155*, 2024b.

Hussein Sayed Eslam, Amerah Alabrah, Hussein Rahouma Kamel, Zohaib Muhammad, and Rasha M. Badry. Machine learning and deep learning for loan prediction in banking: Exploring ensemble methods and data balancing. *IEEE Access*, 12:193997–194019, 2024. doi: 10.1109/ACCESS.2024.3509774.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants, 2025. URL https://arxiv.org/abs/2501.04227.

Tianqi Shang, Weiqing He, Charles Zheng, Lingyao Li, Li Shen, and Bingxin Zhao. Dynamicare: A dynamic multi-agent framework for interactive and open-ended medical decision-making, 2025. URL https://arxiv.org/abs/2507.02616.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Jinxin Shi, Jiabao Zhao, Xingjiao Wu, Ruyi Xu, Yuan-Hao Jiang, and Liang He. Mitigating reasoning hallucination through multi-agent collaborative filtering. *Expert Systems with Applications*, 263:125723, 2025.

Thibault Simonetto, Salah Ghamizi, and Maxime Cordy. Tabularbench: Benchmarking adversarial robustness for tabular deep learning in real-world use-cases. *Advances in Neural Information Processing Systems*, 37:78394–78430, 2024.

Yu Song, Yuyan Wang, Xin Ye, Russell Zaretzki, and Chuanren Liu. Loan default prediction using a credit rating-specific and multi-objective ensemble learning scheme. *Information Sciences*, 629:599–617, 2023.

William E Stein and Philip J Mizzi. The harmonic consistency index for the analytic hierarchy process. *European journal of operational research*, 177(1):488–497, 2007.

Chuanneng Sun, Songjun Huang, and Dario Pompili. Llm-based multi-agent decision-making: Challenges and future directions. *IEEE Robotics and Automation Letters*, 2025.

Jingyun Sun, Chengxiao Dai, Zhongze Luo, Yangbo Chang, and Yang Li. Lawluo: A multi-agent collaborative framework for multi-round chinese legal consultation. *arXiv preprint arXiv:2407.16252*, 2024.

Huirong Tan and Yanruixue Xie. Financial text analysis and credit risk assessment using a gpt-4 and improved bert fusion model. *PLoS One*, 20(11):e0336217, 2025.

Ahmadreza Tavasoli, Maedeh Sharbaf, and Seyed Mohamad Madani. Responsible innovation: A strategic framework for financial llm integration. *arXiv preprint arXiv:2504.02165*, 2025.

Xu Tian, ZongYi Tian, Saleh FA Khatib, and Yan Wang. Machine learning in internet financial risk management: A systematic literature review. *Plos one*, 19(4):e0300195, 2024.

Xiangpeng Wan, Haicheng Deng, Kai Zou, and Shiqi Xu. Enhancing the efficiency and accuracy of underlying asset reviews in structured finance: The application of multi-agent framework, 2024. URL https://arxiv.org/abs/2405.04294.

Shilong Wang, Guibin Zhang, Miao Yu, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, and Yang Wang. G-safeguard: A topology-guided security lens and treatment on llm-based multi-agent systems, 2025a. URL https://arxiv.org/abs/2502.11127.

Xiaofeng Wang, Zhixin Zhang, Jinguang Zheng, Yiming Ai, and Rui Wang. Debt collection negotiations with large language models: An evaluation system and optimizing decision making with multi-agent, 2025b. URL https://arxiv.org/abs/2502.18228.

Yuelin Wang, Yihan Zhang, Yan Lu, and Xinran Yu. A comparative assessment of credit risk model based on machine learning —a case study of bank loan data. *Procedia Computer Science*, 174: 141–149, 2020a. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2020.06.069. URL https://www.sciencedirect.com/science/article/pii/S1877050920315830. 2019 International Conference on Identification, Information and Knowledge in the Internet of Things.

Yuelin Wang, Yihan Zhang, Yan Lu, and Xinran Yu. A comparative assessment of credit risk model based on machine learning—a case study of bank loan data. *Procedia Computer Science*, 174: 141–149, 2020b.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow, 2025c. URL https://arxiv.org/abs/2503.18968.

Zhepei Wei, Wenlin Yao, Yao Liu, Weizhi Zhang, Qin Lu, Liang Qiu, Changlong Yu, Puyang Xu, Chao Zhang, Bing Yin, Hyokun Yun, and Lihong Li. Webagent-r1: Training web agents via end-to-end multi-turn reinforcement learning, 2025. URL https://arxiv.org/abs/2505.16421.

Devin R Wright, Jisun An, and Yong-Yeol Ahn. Cognitive linguistic identity fusion score (clifs): A scalable cognition-informed approach to quantifying identity fusion from text. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11643–11673, 2025.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743, 2024.

Jun Xu. Genai and llm for financial institutions: A corporate strategic survey. *Available at SSRN 4988118*, 2024.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Zijun Yao, Yantao Liu, Yanxu Chen, Jianhui Chen, Junfeng Fang, Lei Hou, Juanzi Li, and Tat-Seng Chua. Are reasoning models more prone to hallucination? *arXiv preprint arXiv:2505.23646*, 2025.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yuechen Jiang, Yupeng Cao, Zhi Chen, Jordan Suchow, Zhenyu Cui, Rong Liu, et al. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *Advances in Neural Information Processing Systems*, 37:137010–137045, 2024.

Ling Yue, Sixue Xing, Jintai Chen, and Tianfan Fu. Clinicalagent: Clinical trial multi-agent system with large language model-based reasoning, 2024. URL https://arxiv.org/abs/2404.14777.

Karina Zainullina, Alexander Golubev, Maria Trofimova, Sergei Polezhaev, Ibragim Badertdinov, Daria Litvintseva, Simon Karasik, Filipp Fisin, Sergei Skvortsov, Maksim Nekrashevich, et al. Guided search strategies in non-serializable environments with applications to software engineering agents. *arXiv preprint arXiv:2505.13652*, 2025.

Monika Zamojska, JarosŁ Chudziak, et al. Games agents play: Towards transactional analysis in llm-based multi-agent systems. *arXiv preprint arXiv:2507.21354*, 2025.

Yifan Zeng, Yiran Wu, Xiao Zhang, Huazheng Wang, and Qingyun Wu. Autodefense: Multi-agent llm defense against jailbreak attacks. *arXiv preprint arXiv:2403.04783*, 2024.

Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for llm-based multi-agent systems. *arXiv preprint arXiv:2410.02506*, 2024a.

Guibin Zhang, Luyang Niu, Junfeng Fang, Kun Wang, Lei Bai, and Xiang Wang. Multi-agent architecture search via agentic supernet, 2025a. URL https://arxiv.org/abs/2502.04180.

Lang Zhang, Haiqing Hu, and Dan Zhang. A credit risk assessment model based on svm for small and medium enterprises in supply chain finance. *Financial Innovation*, 1(1):14, 2015.

Xueqiao Zhang, Chao Zhang, Jianwen Sun, Jun Xiao, Yi Yang, and Yawei Luo. Eduplanner: Llm-based multi-agent systems for customized and intelligent instructional design. *IEEE Transactions on Learning Technologies*, 2025b.

Xuyang Zhang, Lidong Xu, Ningxin Li, and Jianke Zou. Research on credit risk assessment optimization based on machine learning. *Preprints.org*, 2024b.

Yakun Zhu, Shaohang Wei, Xu Wang, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. Menti: Bridging medical calculator and llm agent with nested tool calling, 2025. URL https://arxiv.org/abs/2410.13610.

# APPENDIX CONTENTS


**A Notation** . . . . . **18**

**B Multi-Agent Utility Mathematical Proof** . . . . . **18**

B.1 Preliminaries and Notation . . . . . 18

B.2 Information Processing and Variance Reduction . . . . . 19

B.3 The Hard-Stop Mechanism: Safety Guarantees . . . . . 20

B.4 Utility Optimization via GRPO . . . . . 20

**C Machine Learning LLM Enhanced Dynamic Filter Loop** . . . . . **22**

C.1 Feature-Selection Pipeline . . . . . 22

C.2 Theoretical Summary . . . . . 23

C.3 Validation of Feature Selection . . . . . 23

**D Additional Experimental Results** . . . . . **24**

D.1 Worked Example of the Decision Rule . . . . . 25

D.2 Threshold-Tuned Static Fusion . . . . . 25

D.3 Decomposition of BEC . . . . . 26

D.4 Cost-Sensitive Tabular Baselines . . . . . 26

D.5 Hard-Stop Ablation . . . . . 28

D.6 Evaluation Groups . . . . . 28

D.7 Label Construction for Rejected Applicants . . . . . 28

D.8 Comparison with Public Credit Datasets . . . . . 29

D.9 Transfer Scope and Practical Use . . . . . 29

**E Limitations** . . . . . **30**

**F Layer 2 Agents Training Data Details** . . . . . **31**

F.1 Agent Architecture . . . . . 31

F.2 Credit-Instruct Dataset Construction . . . . . 31

F.3 Risk-Oriented Textualization (ROT) . . . . . 32

F.4 Data Distribution . . . . . 32

**G Model Configurations** . . . . . **33**

**H Details of Token Counts and Latency** . . . . . **34**

**I Case Study** . . . . . **35**

**J More Related Work** . . . . . **44**

J.1 Credit Assessment . . . . . 44

J.2 MAS in Specific Domain . . . . . 44

J.3 MAS Architecture . . . . . 45

**K Prior Knowledge in Finance about High-Stakes Financial Decision-Making** . . . . . **48**

K.1 Macro-Economic Context and Systematic Risk . . . . . 48

K.2 Identity Verification and Basic Profiling . . . . . 48

K.3 Credit Behavior History and Velocity Analysis . . . . . 49

K.4 Solvency, Liquidity, and Asset Quality . . . . . 49

K.5 Repayment Willingness and Behavioral Signals . . . . . 49

K.6 Fraud Detection and Consistency Verification . . . . . 49

**L Expert Review Summary** . . . . . **50**

**M Dataset Details** . . . . . **52**

M.1 Core Risk Assessment Dimensions (Main Risk Domains) . . . . . 52

**N Statistical Characteristics** . . . . . **54**

**O Sample Prompts** . . . . . **55**

# A NOTATION

Table 6: Notation. The label $y$ is used in the Layer 1 gate (Eq. (1)), the SFT loss (Eq. (5)), and the cost in Eq. (6).

| Symbol | Meaning |
|---|---|
| $\mathbf{x}$, $\mathbf{x}'$ | Raw evidence vector; refined evidence after Layer 1 |
| $y \in \{0, 1\}$ | Ground-truth default label ($1 =$ Default) |
| $\hat{Y} \in \{0, 1\}$ | Final decision ($1 =$ Reject, $0 =$ Approve) |
| $a_j$, $o_j = \{r_j, e_j\}$ | Specialist agent $j$ and its output tuple |
| $r_j \in \{1, \ldots, 10\}$ | Raw specialist score (creditworthiness; larger = safer) |
| $q_j = 1 - r_j/10$ | Normalized risk (larger = riskier) |
| $e_j \in \mathcal{E}^*$ | Evidence chain in the constrained JSON output |
| $S_{\text{shared}}$ | Shared blackboard state $[o_1 \| \cdots \| o_j]$ |
| $\alpha_j$, $S_{\text{risk}}$ | Fusion weight of agent $j$; fused risk score |
| $\tau^* = \pi_\theta(S_{\text{shared}})$ | Case-specific threshold from the GRPO policy |
| $\mathcal{V}$, $V_k$ | Veto-enabled agents; hard-stop indicator of agent $k$ |

# B MULTI-AGENT UTILITY MATHEMATICAL PROOF

To formally validate the CREDITAGENT architecture, we provide theoretical foundations demonstrating that the hierarchical arrangement of specialized agents with shared context optimizes decision utility and provides formal safety guarantees.

Table 7: Summary of theoretical guarantees for the CREDITAGENT architecture.

| Objective | Formal Statement | Reference |
|---|---|---|
| Variance Reduction | $\mathbb{E}[\text{Var}(R_j \mid \mathbf{x}', S_{\text{shared}})] \leq \mathbb{E}[\text{Var}(R_j \mid \mathbf{x}')]$ | Theorem B.4 |
| Absolute Safety | $V_k = 1 \implies \hat{Y} = 1$ (Reject) deterministically | Theorem B.7 |
| Consensus Gradient | $\frac{\partial \text{BEC}}{\partial \sigma_{\text{agents}}} < 0$ when profitable | Theorem B.9 |

## B.1 PRELIMINARIES AND NOTATION

We adopt notation consistent with the main paper (Section 4). Let $\mathcal{F} : \mathcal{X} \to \{0, 1\}$ denote the hierarchical decision function decomposed into three latent spaces: Data Governance ($\mathcal{Z}_1$), Domain Expertise ($\mathcal{Z}_2$), and Decision Fusion ($\mathcal{Z}_3$).

**Definition B.1** (Credit Decision Environment)**.** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We define:

- $Y \in \{0, 1\}$: True repayment outcome, where $Y = 1$ denotes default
- $\mathbf{x} \in \mathcal{X}$: Raw input feature set (2,515 raw fields in our data)
- $\mathbf{x}' = \mathbf{x} \odot \mathbb{I}[\text{XGB}(\mathbf{x}, y) > \gamma]$: Refined evidence tensor from Layer 1 (Eq. 1)
- $\mathcal{A} = \{a_1, \ldots, a_6\}$: Set of Layer 2 specialist agents (Macro, Basic, History, Solvency, Willingness, Fraud)
- $o_j = \{r_j, e_j\}$: Output tuple of agent $a_j$, where $r_j \in \{1, \ldots, 10\}$ is the creditworthiness score (higher is safer), $q_j = 1 - r_j/10 \in [0, 1]$ is the normalized risk, and $e_j \in \mathcal{E}^*$ is the Evidence Chain
- $\hat{Y} \in \{0, 1\}$: Final decision, where $\hat{Y} = 1$ denotes Reject

**Definition B.2** (Hierarchical Context State)**.** Following the State-Aware Routing Framework (Section 4.1), the global context state is updated recursively:

$$S_{t+1} = [S_t \| o_{t+1}] \tag{8}$$

We define the shared context available to agent $a_j$ at position $j$ in the routing sequence as:

$$S^{(j)}_{\text{shared}} = [o_1 \| \cdots \| o_{j-1}] \tag{9}$$

where $\|$ denotes blackboard concatenation of previously published agent outputs. In particular, the outputs of the Macro Agent and Basic Agent often appear early in the routing order and therefore become part of the context received by later agents.

**Definition B.3** (Architecture Comparison)**.** The relationship between specialized agents is categorized by their information-sharing protocols:

**Parallel Architecture ($\mathcal{P}$):** Agents operate in isolation. Each agent $a_j$ computes its output independently:

$$r^{\mathcal{P}}_j = f_j(\mathbf{x}')$$

where no information is exchanged between agents.

**Hierarchical Architecture ($\mathcal{H}$):** Agents operate via a shared state. Each agent $a_j$ computes its output conditioned on the global context:

$$r^{\mathcal{H}}_j = f_j(\mathbf{x}' \mid S^{(j)}_{\text{shared}})$$

where $S^{(j)}_{\text{shared}}$ denotes the state of the Blackboard at the time of invocation.

### B.2 Information Processing and Variance Reduction

We represent specialized agents as estimators of the hidden risk state $Y$. Let $\mathbf{x}'$ be the refined feature vector from Layer 1.

**Theorem B.4** (Contextual Stability)**.** *The hierarchical conditioning of specialized agents on the shared context $S_{shared}$ reduces the expected conditional variance compared to a parallel architecture. Formally, let $R_j$ be the normalized risk $q_j$ predicted by specialist agent $a_j$. Then:*

$$\mathbb{E}\left[\text{Var}(R_j \mid \mathbf{x}', S_{shared})\right] \leq \mathbb{E}\left[\text{Var}(R_j \mid \mathbf{x}')\right] \tag{10}$$

*with equality if and only if $R_j \perp\!\!\!\perp S_{shared} \mid \mathbf{x}'$ (i.e., the shared context provides no additional information beyond $\mathbf{x}'$).*

*Proof.* By the **Law of Total Variance**, we decompose the variance:

$$\text{Var}(R_j \mid \mathbf{x}') = \mathbb{E}\left[\text{Var}(R_j \mid \mathbf{x}', S_{\text{shared}}) \mid \mathbf{x}'\right] + \text{Var}\left(\mathbb{E}[R_j \mid \mathbf{x}', S_{\text{shared}}] \mid \mathbf{x}'\right) \tag{11}$$

Since variance is non-negative:

$$\text{Var}\left(\mathbb{E}[R_j \mid \mathbf{x}', S_{\text{shared}}] \mid \mathbf{x}'\right) \geq 0 \tag{12}$$

This directly implies:

$$\mathbb{E}\left[\text{Var}(R_j \mid \mathbf{x}', S_{\text{shared}}) \mid \mathbf{x}'\right] \leq \text{Var}(R_j \mid \mathbf{x}') \tag{13}$$

Taking expectations over $\mathbf{x}'$ yields the result. Equality holds iff $\mathbb{E}[R_j \mid \mathbf{x}', S_{\text{shared}}]$ is $\mathbf{x}'$-measurable, i.e., when $S_{\text{shared}}$ provides no additional predictive information. □

*Remark* B.5 (Interpretation: Diluted Attention Prevention). The variance reduction quantifies the prevention of "diluted attention" described in Section 1 and addressed by the Shared Blackboard Architecture in Section 4.1. By conditioning on $S_{\text{shared}}$, specialized agents avoid the information fragmentation that occurs in single-agent LLM systems processing high-dimensional credit data.

### B.3 The Hard-Stop Mechanism: Safety Guarantees

As defined in Eq. (4), the final decision is governed by a fused risk score combined with a "Hard-Stop" rule. We show that a fired hard-stop always produces a rejection.

**Definition B.6** (Decision Function with Veto)**.** Following Eq. (4), the final loan decision $\hat{Y} : \mathcal{X}' \times \mathcal{S} \to \{0, 1\}$ ($1 =$ Reject) is defined as:

$$\hat{Y} = \begin{cases} 1 & \text{if } S_{\text{risk}} \geq \pi_\theta(S_{\text{shared}}) \text{ or } \exists k \in \mathcal{V} : \; V_k = 1 \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where:

- $S_{\text{risk}} = \sigma\left(\sum_{j=1}^{n} \alpha_j \cdot q_j\right)$ with $n = |\mathcal{A}| = 6$ is the fused risk score from Layer 3 (Eq. (2))
- $\pi_\theta(S_{\text{shared}})$ is the GRPO policy-adjusted threshold
- $\mathcal{V} \subseteq \mathcal{A}$ is the critical veto-agent set
- $V_k \in \{0, 1\}$ is the deterministic hard-stop indicator computed on the constrained JSON output of agent $k$; in the evaluated system it fires when $r_{\text{fraud}} \leq 3$

We express the decision function equivalently as:

$$\hat{Y} = 1 - (1 - \mathbb{I}\left[S_{\text{risk}} \geq \pi_\theta(S_{\text{shared}})\right]) \times \prod_{k \in \mathcal{V}} (1 - V_k(\mathbf{x}', S_{\text{shared}})) \tag{15}$$

where $V_k$ is the hard-stop indicator defined above.

**Theorem B.7** (Boundary Safety)**.** *The* CreditAgent *architecture guarantees that no positive indicators from specialty agents can override a deterministic fraud rejection. Formally, if any veto condition is triggered:*

$$\exists k \in \mathcal{V} : V_k = 1 \implies \hat{Y} = 1 \tag{16}$$

*Proof.* Suppose $\exists k^* \in \mathcal{V}$ such that $V_{k^*}(\mathbf{x}', S_{\text{shared}}) = 1$ (i.e., the hard-stop of agent $k^*$ fires). Then from Eq. equation 15:

$$\prod_{k \in \mathcal{V}} (1 - V_k(\mathbf{x}', S_{\text{shared}})) = (1 - V_{k^*}) \cdot \prod_{k \neq k^*} (1 - V_k) = 0 \tag{17}$$

Therefore $\hat{Y} = 1 - (1 - \mathbb{I}[S_{\text{risk}} \geq \pi_\theta(S_{\text{shared}})]) \times 0 = 1$, regardless of $S_{\text{risk}}$. □

In particular, even if the other $n - 1$ agents all give high creditworthiness scores, they cannot override a Fraud-agent hard-stop.

### B.4 Utility Optimization via GRPO

To optimize the Business Efficiency Coefficient (BEC) defined in Eq. 7, the system must balance False Negatives ($C_{FN}$: approving a defaulting customer) and False Positives ($C_{FP}$: rejecting a good customer).

**Definition B.8** (Business Efficiency Coefficient)**.** Following Eq. 6 and Eq. 7, the BEC is defined as:

$$\text{BEC} = \max\left(0, \text{BEC}_{\text{pre}} \times (1 - \lambda_{\text{bec}} \cdot \sigma_{\text{agents}})\right) \tag{18}$$

where:

$$\text{BEC}_{\text{pre}} = 1 - \frac{\text{AvgCost}}{\text{MaxCost}}, \quad \text{MaxCost} = \max(C_{FN}, C_{FP}) \tag{19}$$

and $\sigma_{\text{agents}}$ is the standard deviation of the normalized risks $\{q_1, \ldots, q_n\}$ across Layer 2 agents.

**Theorem B.9** (Consensus-Driven Optimization)**.** *The GRPO optimization with BEC objective induces a gradient that monotonically decreases inter-agent variance when the system is profitable:*

$$\frac{\partial \text{BEC}}{\partial \sigma_{\text{agents}}} < 0 \quad \textit{whenever } \text{BEC}_{\text{pre}} > 0 \tag{20}$$

*Proof.* From the BEC definition:

$$\frac{\partial \mathrm{BEC}}{\partial \sigma_{\mathrm{agents}}} = \mathrm{BEC}_{\mathrm{pre}} \cdot \frac{\partial}{\partial \sigma_{\mathrm{agents}}}(1 - \lambda_{\mathrm{bec}} \cdot \sigma_{\mathrm{agents}}) = -\lambda_{\mathrm{bec}} \cdot \mathrm{BEC}_{\mathrm{pre}} \tag{21}$$

Since $\lambda_{\mathrm{bec}} > 0$ (penalty intensity) and $\mathrm{BEC}_{\mathrm{pre}} > 0$, we have $\frac{\partial \mathrm{BEC}}{\partial \sigma_{\mathrm{agents}}} < 0$. □

*Remark* B.10 (Avoiding Degenerate Consensus). The GRPO optimization encourages agent consensus but is bounded by irreducible uncertainty $\sigma_{\min} = \sqrt{\mathbb{E}[\mathrm{Var}(Y \mid \mathbf{x}', S_{\text{shared}})]}$.

# C MACHINE LEARNING LLM ENHANCED DYNAMIC FILTER LOOP

In credit-underwriting scenarios, analyzing, processing, and predicting from high-dimensional structured raw data is critically important. Such datasets commonly contain hundreds or even thousands of feature variables describing a borrower's profile, income, credit history, and consumption behavior. The core challenge is extracting meaningful predictive signals from this high-dimensional space, which is often contaminated with various forms of noise. We can reframe this feature selection problem through an information-theoretic lens: the goal is to reconstruct the essential signal of a borrower's creditworthiness from noisy observations by maximizing the transmitted information relevant to the target variable (e.g., default status) under constraints. The Layer-1 component of our Agent Framework introduces an automated, dynamic feature-selection pipeline designed to act as an intelligent, iterative "denoising" and "information compression" system, which is illustrated in figure 5. Below, we describe this automated feature-selection agent, now augmented with information-theoretic principles, a theoretical summary, and experimental results of feature selection.

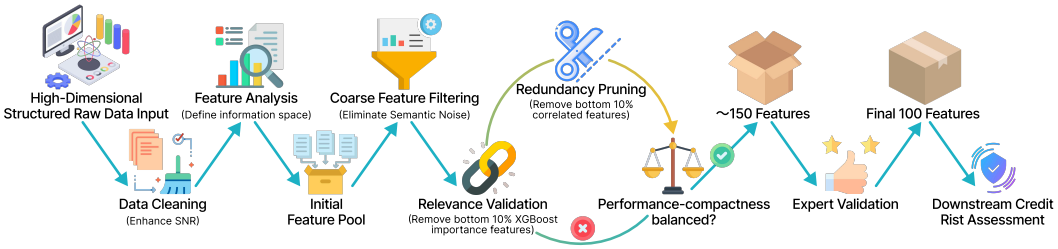



Figure 5: Pipeline of Feature Selection Agent.

## C.1 FEATURE-SELECTION PIPELINE

**Data Cleaning: Enhancing the Signal-to-Noise Ratio** Large and complex datasets contain not only missing values, outliers, and format inconsistencies but also substantial redundancy and noise. We can model the observed raw data $\mathbf{X}_{\text{raw}}$ as a combination of a latent true signal $\mathbf{X}_{\text{true}}$ and an additive noise component $\mathbf{N}$, such that $\mathbf{X}_{\text{raw}} = \mathbf{X}_{\text{true}} + \mathbf{N}$. The noise $\mathbf{N}$ encompasses random errors, missing entries, and inconsistencies. Manual cleaning is slow and prone to omissions. Our automated data-cleaning procedures aim to estimate and reduce $\mathbf{N}$, thereby increasing the effective *signal-to-noise ratio* of the dataset. The output is a higher-fidelity, well-structured dataset $\mathbf{X}_{\text{clean}}$ that serves as a superior basis for subsequent information extraction, satisfying $I(\mathbf{X}_{\text{clean}}; Y) \geq I(\mathbf{X}_{\text{raw}}; Y)$, where $I(\cdot;\cdot)$ denotes mutual information and $Y$ is the target.

**Feature Analysis: Defining the Information Space** We perform a global analysis to understand the business objective—maximizing information about target $Y$—and operational constraints, such as an upper bound on the number of allowable features (a form of bandwidth constraint). We then create initial feature meta-information. For each raw feature, we produce a human-interpretable description, effectively defining the *alphabet* and *semantics* for each information source. This step is foundational for calculating information-theoretic measures like entropy $H(X)$ and mutual information $I(X;Y)$. When data is distributed across multiple tables, we perform joins to consolidate information from multiple correlated sources, potentially increasing the total information available about $Y$.

**Coarse Feature Filtering: Eliminating Semantic Noise Channels** The coarse screening stage performs an initial review of all candidate features to remove those constituting pure *semantic noise*. Features such as arbitrary IDs or session hashes are, by design, non-informative for predicting $Y$. From an information-theory perspective, their mutual information with the target is approximately zero:

$$I(F_{\text{id}}; Y) = H(Y) - H(Y|F_{\text{id}}) \approx 0$$

This indicates that knowing $F_{\text{id}}$ does not reduce the uncertainty $H(Y)$ about the target. According to the *Data Processing Inequality*, discarding such features incurs no information loss about $Y$ and is a lossless compression step. We also remove duplicate fields and other non-informative attributes. A final manual audit confirms the reasonableness of this initial screening.

**Fine-Grained Feature Selection: Iterative Information Bottleneck** After coarse screening, we apply an iterative fine-grained selection process that embodies the principles of the *Information Bottleneck* method, seeking a compact representation $\mathbf{X}_{\text{selected}}$ that preserves maximal information about $Y$ while minimizing redundancy. The process combines correlation-based pruning with XGBoost-based importance validation.

1. **Redundancy Pruning**: We compute the Pearson correlation matrix and remove features in the bottom 10%. Highly correlated features share large mutual information $I(F_i; F_j)$, implying redundancy about $Y$. Removing one of a correlated pair compresses the representation with minimal information loss regarding $Y$.
2. **Relevance Validation**: The remaining features train an XGBoost model. The model's feature importance scores approximate each feature's contribution to reducing the prediction loss (e.g., cross-entropy, related to the conditional entropy $H(Y|\mathbf{X})$). We remove the least important 10% of features.

This loop repeats, each iteration potentially refining the feature set to increase the *information rate* $I(\mathbf{X}_{\text{selected}}; Y)$ per feature. The process terminates when a balance between model performance (a proxy for $I(\mathbf{X}_{\text{selected}}; Y)$) and feature compactness is achieved, yielding a subset of approximately 150 features.

**Final Validation: Incorporating Expert Priors** To obtain a validated set of predictors, domain experts perform a final review. This step integrates rich *prior knowledge*—an external information source—into the system. From 150 features, experts validated and annotated the 100 most relevant to credit review operations. This ensures the final feature set maximizes not just statistical mutual information $I_{\text{statistical}}(F; Y)$, but also *business-relevant information* $I_{\text{business}}(F; Y)$, guaranteeing actionability and interpretability. The final output is a robust, information-rich, and compact feature combination for downstream underwriting analysis.

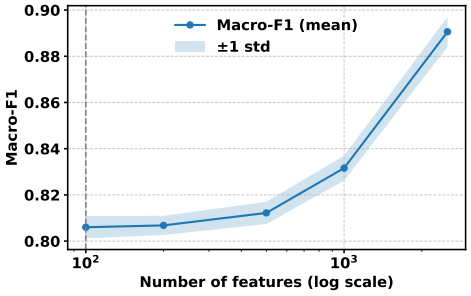
### C.2 Theoretical Summary

The pipeline can be formally viewed as an optimization problem:

$$\underset{\mathbf{X}_{\text{selected}} \subseteq \mathbf{X}_{\text{raw}}}{\text{maximize}} \quad I(\mathbf{X}_{\text{selected}}; Y) \quad \text{subject to} \quad |\mathbf{X}_{\text{selected}}| \leq k$$

where $k$ is the desired feature count. The steps of cleaning, filtering, and iterative selection collectively work to approximate this optimum by systematically removing noise, eliminating non-informative signals, and compressing redundancy, all while preserving the information most valuable for predicting credit risk.

(a) Performance vs. Number of Selected Features




(b) Per-class F1 Comparison Before/After Feature Selection

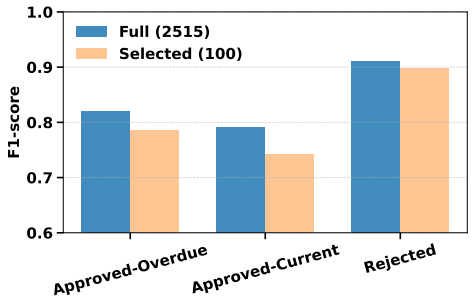



Figure 6: Feature Selection Results: Per-Class Performance and Macro-F1 Trends

### C.3 Validation of Feature Selection

To examine how feature selection affects predictive performance, we evaluated model metrics at multiple stages of the feature selection pipeline, progressively reducing the feature space from the

full set of 2,515 features to the final 100 features used in our reported experiments. At each stage, we trained an XGBoost model and assessed its predictive performance. Figure 6a illustrates how the macro F1 score varies with the number of retained features (mean $\pm$ standard deviation over 5-fold cross-validation). Although using all 2,515 features yields a higher nominal score (macro F1 = 0.89), performance drops sharply during the initial reduction step and then stabilizes as fewer but more informative features are retained.

We further analyzed performance across the three major outcome classes in the dataset: (1) approved at credit review but subsequently delinquent, (2) approved and not delinquent, and (3) rejected at credit review. The class analysis in Figure 6b shows that the model maintains strong performance on the "rejected" class (F1 = 0.898), suggesting that the reduced set of characteristics preserves key discriminative signals relevant to credit-rejection decisions. Moreover, the corresponding classification results are reported in Table 8. In particular, the model achieves particularly high accuracy for the "rejected" class, and attains an overall F1 score of 80.8%, providing a solid basis for subsequent credit-review analysis.

We note that the apparent decline in predictive scores after aggressive feature pruning does not necessarily indicate degraded model quality. Feature selection can reduce model complexity, mitigate overfitting, and improve interpretability and deployment efficiency (e.g., faster inference and lower data acquisition costs).

Table 8: Classification Report of Feature Prediction

| **Classification** | **Precision** | **Recall** | $F1$**-Score** |
|---|---|---|---|
| Approved – Overdue | 0.7400 | 0.8350 | 0.7847 |
| Approved – Current | 0.8131 | 0.6817 | 0.7416 |
| Rejected | 0.8839 | 0.9133 | 0.8984 |
| **Avg.** | 0.8123 | 0.8100 | 0.8082 |

## D Additional Experimental Results

This section reports the additional analyses referenced in Sections 4 and 5. Unless stated otherwise, all analyses use the archived per-case outputs of the SFT specialist agents, the cost ratio $C_{\mathrm{FN}} : C_{\mathrm{FP}} = 3:1$, and positive class = default. In the confusion matrices, TP is a rejected defaulter, FP is a rejected non-defaulter, FN is an approved defaulter, and TN is an approved non-defaulter. The evaluation set is risk-enriched by construction (about two-thirds of cases are defaulters), so approval rates and default rates among approved cases describe behavior on hard boundary cases and are not portfolio rates.

***Summary.*** Section 5.2 refers to four alternative explanations of the fusion gain, which the subsections below test. (i) *Threshold tuning.* When the Avg and DST thresholds are tuned on the GRPO-stage training cases, a threshold that maximizes the raw cost score collapses G2 accuracy to 3.30–12.14%, which corresponds to rejecting almost all boundary cases. A threshold that maximizes balanced group accuracy gives 77.49–77.97% accuracy and 40.16–48.57% G2 accuracy. GRPO exceeds the best tuned static rule by 5.35 points of accuracy and 17.68 points of G2 accuracy (Appendix D.2). (ii) *Hard-stop rules.* Disabling the hard-stop while keeping the same fused scores lowers accuracy from 83.32% to 82.91% and G3 accuracy from 93.20% to 91.98%, with no change on G1 and G2 (Appendix D.5). (iii) *The consistency term in BEC.* The factor $1 - \lambda_{\mathrm{bec}}\sigma_{\mathrm{agents}}$ is 0.970 for every fusion variant, because all variants share the same specialist outputs (Appendix D.3). (iv) *Tuned tabular models.* XGBoost and CatBoost retrained on the same GRPO-stage split with cost-aware threshold selection reach 80.08–80.17% accuracy and 55.81–56.50% G2 accuracy at the balanced operating point, and they have higher AUC (0.846–0.847) and better calibration than static fusion of the specialist scores. CreditAgent remains 3.15 points higher in accuracy and 9.75 points higher in G2 accuracy on the same held-out cases (Appendix D.4). The advantage of CreditAgent is largest on G2 boundary cases.

### D.1 Worked Example of the Decision Rule

Table 9 traces one held-out case through the notation of Section 3, using the archived specialist outputs and uniform weights for illustration. The fused risk is $\bar{q} = \sum_j w_j q_j = 0.567$. The threshold of the static baselines in risk space is $1 - 6.0/10 = 0.40$. Because $0.567 \geq 0.40$, the decision is $\hat{Y} = 1$ (Reject). The fraud hard-stop does not fire, because $r_{\text{fraud}} = 6 > 3$. The institution approved this applicant and the loan defaulted (G1), so the reject decision is correct. Every quantity moves in one direction: a larger $q_j$ gives a larger fused risk and makes rejection more likely. Re-running Avg fusion with this rule on all 6,000 archived specialist outputs reproduces the Avg row of Table 3 exactly (78.40% accuracy; G1/G2/G3 = 78.95/62.05/94.20), which confirms that the notation matches the implementation.

Table 9: A held-out G1 case traced through the decision rule. $r_j$ is the raw creditworthiness score (higher is safer), and $q_j = 1 - r_j/10$ is the normalized risk (higher is riskier). The Macro agent reports its score under the output dimension "Application Reasonableness" (Appendix O).

| **Agent (output dimension)** | **Raw score $r_j$ (1–10)** | **Risk $q_j$** |
|---|---|---|
| Basic (applicant basic information) | 1 | 0.90 |
| Solvency | 6 | 0.40 |
| Willingness (willingness to repay) | 3 | 0.70 |
| History (credit history) | 8 | 0.20 |
| Fraud (fraud risk) | 6 | 0.40 |
| Macro (application reasonableness) | 2 | 0.80 |
| Fused risk $\bar{q}$ (uniform weights) | $0.567 \geq 0.40 \Rightarrow$ Reject | |

### D.2 Threshold-Tuned Static Fusion

The static baselines Avg and DST in Table 3 use a fixed threshold of 6.0 on the 1–10 fused creditworthiness scale. To test whether the GRPO gain comes only from threshold tuning, we give Avg and DST a scalar threshold selected on the GRPO-stage training cases ($n = 1{,}030$ matched to the archived specialist outputs) and evaluate it on the matched held-out cases ($n = 4{,}104$). We use two selection objectives: the threshold that maximizes the raw cost score (BEC-tuned), and the threshold that maximizes macro-averaged group accuracy (macro-tuned). Table 10 reports the results. On the full set, Avg at the fixed threshold gives 78.40/78.95/62.05/94.20, within 0.8 points of the subset values in accuracy and G1, so the subset is representative in these columns.

Table 10: Static fusion with tuned scalar thresholds on the same specialist outputs. Static rows are evaluated on the matched held-out subset ($n = 4{,}104$); the GRPO row is the full-set result ($n = 6{,}000$) from Table 3. The BEC column for static rows is the raw group-cost score of the released evaluation code, and the GRPO BEC additionally includes the consistency adjustment of Eq. (7); the BEC values should therefore not be compared across these rows, and the accuracy columns are the comparable quantities.

| **Fusion** | **Threshold** | $\tau$ | **BEC** | **Acc. (%)** | **G1 (%)** | **G2 (%)** | **G3 (%)** |
|---|---|---|---|---|---|---|---|
| Avg | Fixed | 6.00 | 0.814 | 77.70 | 78.68 | 53.86 | 94.29 |
| Avg | BEC-tuned | 7.15 | 0.900 | 74.29 | 97.60 | 12.14 | 99.43 |
| Avg | Macro-tuned | 6.15 | 0.834 | 77.97 | 81.95 | 48.57 | 96.01 |
| DST | Fixed | 6.00 | 0.700 | 71.86 | 62.45 | 66.18 | 84.21 |
| DST | BEC-tuned | 8.50 | 0.902 | 72.49 | 99.34 | 3.30 | 99.68 |
| DST | Macro-tuned | 6.85 | 0.857 | 77.49 | 87.12 | 40.16 | 96.39 |
| CreditAgent (GRPO) | Learned | – | 0.7647 | 83.32 | 90.50 | 66.25 | 93.20 |

The results show two effects. First, when the scalar threshold is tuned to maximize the raw cost score, static fusion reaches a high score by rejecting almost all boundary cases: G2 accuracy falls to 12.14% for Avg and 3.30% for DST, and the approval rate falls to 4.43% and 1.27% (Table 11).

This is the failure mode expected when a threshold-sensitive cost metric is optimized with a single global threshold, and it is the reason we read BEC together with accuracy and per-group accuracy. Second, when the threshold is tuned for balanced group accuracy, static fusion reaches 77.49–77.97% accuracy with G2 accuracy of 40.16–48.57%. Relative to the best macro-tuned static row (Avg, 6.15), GRPO fusion is higher by 5.35 points in accuracy, 3.38 points in G1, and 17.68 points in G2, and lower by 3.19 points in G3. Sweeping the scalar threshold over its full range does not reach the GRPO accuracy and G2 accuracy at the same time, because one global threshold trades the three groups against each other. With the same specialist outputs and hard-stop rules, the gain of GRPO fusion therefore comes from case-specific fusion and not only from the presence of threshold tuning.

Table 11: Standard decision diagnostics for static fusion on the matched held-out subset ($n = 4{,}104$; AUC 0.777, ECE 0.246 for the fused score) and for Avg at the fixed threshold on the full set ($n = 6{,}000$; AUC 0.815, ECE 0.227). Eq. BEC applies the consistency factor of Eq. (7) to $\mathrm{BEC}_{\mathrm{pre}}$.

| **Configuration** | $\mathrm{BEC}_{\mathrm{pre}}$ | **Eq. BEC** | **TP/FP/FN/TN** | **Approval (%)** | **Default among approved (%)** |
|---|---|---|---|---|---|
| Avg, BEC-tuned | 0.8862 | 0.8597 | 2909/1013/42/140 | 4.43 | 23.08 |
| Avg, macro-tuned | 0.8475 | 0.8222 | 2640/593/311/560 | 21.22 | 35.71 |
| Avg, fixed (subset) | 0.8320 | 0.8072 | 2568/532/383/621 | 24.46 | 38.15 |
| Avg, fixed (full set) | 0.8307 | 0.8055 | 3463/759/537/1241 | 29.63 | 30.20 |
| DST, BEC-tuned | 0.8844 | 0.8580 | 2937/1115/14/38 | 1.27 | 26.92 |
| DST, macro-tuned | 0.8609 | 0.8352 | 2717/690/234/463 | 16.98 | 33.57 |

The BEC values recomputed from the released evaluation code on the archived specialist outputs (Tables 10 and 11) are not on the same scale as the BEC column of Table 3; for example, Avg at the fixed threshold has Eq. BEC 0.8055 on the full set, compared with 0.6862 in Table 3, although its accuracy and per-group accuracies are reproduced exactly. We therefore use accuracy, per-group accuracy, and the diagnostics above for all comparisons in this section and do not compare BEC values between this section and Table 3.

### D.3 Decomposition of BEC

Eq. (7) multiplies the cost-based term $\mathrm{BEC}_{\mathrm{pre}}$ by the consistency factor $1 - \lambda_{\mathrm{bec}}\sigma_{\mathrm{agents}}$, where $\sigma_{\mathrm{agents}}$ is the standard deviation of the six normalized specialist risks $\{q_j\}$. Because Avg, DST, and GRPO fusion use the same specialist outputs, they share the same $\sigma_{\mathrm{agents}}$. Table 12 reports its mean value and the resulting factor. The factor is 0.970 on both the matched subset and the full set and is identical for all fusion rows, so the consistency term changes BEC by at most about 3% in relative terms and does not explain differences between fusion strategies. Differences in BEC between fusion variants come from $\mathrm{BEC}_{\mathrm{pre}}$, that is, from the confusion matrix under the 3:1 cost ratio.

Table 12: Consistency term of BEC. The values are shared by all fusion variants because they are computed from the same specialist outputs.

| **Quantity** | **Held-out subset** ($n = 4{,}104$) | **Full set** ($n = 6{,}000$) |
|---|---|---|
| Mean $\sigma_{\mathrm{agents}}$ | 0.2987 | 0.3031 |
| Consistency factor $1 - \lambda_{\mathrm{bec}}\sigma_{\mathrm{agents}}$ | 0.9701 | 0.9697 |

### D.4 Cost-Sensitive Tabular Baselines

The tabular baselines in Table 3 use default decision thresholds. This subsection gives tabular models the same kind of cost-aware threshold selection. We report two experiments.

***Supplementary 82-feature check.*** SVM, LightGBM, XGBoost, CatBoost, and Random Forest are trained on an 82-feature matrix and receive a threshold selected on a validation split under two objectives: the threshold that maximizes BEC under the 3:1 cost ratio, and the threshold that maximizes macro-averaged group accuracy. All models use the same feature matrix and validation/test split (Table 13). When the threshold is tuned to maximize BEC, every model moves toward rejection

and G2 accuracy falls to 33.7–68.8%, while the balanced threshold recovers G2 accuracy at some BEC cost. For LightGBM the selected threshold is at an extreme value ($\tau \approx 0.01$), and we do not base any claim on that row. Table 14 gives the standard diagnostics of the cost-sensitive XGBoost. These rows use their own feature matrix and split, so they are a robustness check rather than a matched replacement for the controlled fusion comparison.

Table 13: Supplementary 82-feature tabular baselines under the 3:1 cost ratio with validation-selected thresholds. *LightGBM selects an extreme threshold ($\tau \approx 0.01$).

| | | BEC-tuned $\tau$ | | | Macro-tuned $\tau$ | | |
|---|---|---|---|---|---|---|---|
| **Model** | **AUC** | **BEC** | **Acc. (%)** | **G2 (%)** | **BEC** | **Acc. (%)** | **G2 (%)** |
| XGBoost | 0.864 | 0.879 | 81.6 | 59.5 | 0.866 | 81.4 | 63.1 |
| Random Forest | 0.857 | 0.844 | 80.9 | 68.8 | 0.844 | 80.9 | 68.8 |
| SVM | 0.850 | 0.894 | 79.0 | 42.6 | 0.865 | 81.1 | 62.3 |
| LightGBM* | 0.861 | 0.894 | 77.1 | 33.7 | 0.894 | 77.1 | 33.7 |
| CatBoost | 0.866 | 0.876 | 82.0 | 62.2 | 0.868 | 82.0 | 65.0 |

Table 14: Standard diagnostics of the cost-sensitive XGBoost on the 82-feature matrix.

| **Metric** | **Value** |
|---|---|
| Accuracy (%) | 81.39 |
| AUC | 0.864 |
| ECE (10-bin) | 0.062 |
| Approval rate (%) | 27.33 |
| Default rate among approved (%) | 23.35 |
| Confusion matrix (TP/FP/FN/TN) | 1809/366/191/627 |

***Matched-split comparison.*** To remove the difference in split, we retrain XGBoost and CatBoost on the exact GRPO-stage split: 1,200 training cases (400 per group) and 4,800 held-out cases (1,600 per group), using the 81 numeric fields that can be recovered from the case text read by the agents. The threshold is selected by 5-fold out-of-fold prediction on the training cases under the same two objectives, and the evaluation uses the same labels, cost ratio, and metric code. Table 15 reports the results. The tuned tabular models have higher AUC (0.846–0.847 compared with 0.777 for the fused specialist score) and better calibration than static fusion. At the balanced operating point they reach 80.08–80.17% accuracy and 55.81–56.50% G2 accuracy. CREDITAGENT (GRPO) reaches 83.32% accuracy and 66.25% G2 accuracy, which is 3.15 points higher in accuracy and 9.75 points higher in G2 accuracy than the best tabular row, and it has the highest G1 accuracy among balanced operating points. We conclude that tuned tabular models are competitive on aggregate metrics and that the advantage of CREDITAGENT is concentrated on G2 boundary cases. Per-case AUC, ECE, and confusion matrices for the GRPO system are not available, because the training pipeline logged only aggregate validation metrics; we therefore do not report these columns for CREDITAGENT and do not substitute a proxy.

Table 15: Matched-split comparison on the GRPO-stage held-out cases ($n = 4{,}800$). Thresholds are selected by 5-fold out-of-fold prediction on the 1,200 training cases. AUC and ECE do not depend on the threshold and are listed once per model. GRPO per-case diagnostics are not available (–).

| **Model** | **Threshold** | $\tau$ | $\mathrm{BEC}_{\mathrm{pre}}$ | **Acc.** | **G1** | **G2** | **G3** | **AUC / ECE** | **Appr.** | **Def. appr.** | **TP/FP/FN/TN** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| XGBoost | BEC-tuned | 0.067 | 0.8846 | 76.02 | 95.88 | 32.50 | 99.69 | 0.846 / 0.084 | 12.31 | 12.01 | 3129/1080/71/520 |
| XGBoost | Macro-tuned | 0.364 | 0.8693 | 80.17 | 85.44 | 56.50 | 98.56 | | 24.17 | 22.07 | 2944/696/256/904 |
| CatBoost | BEC-tuned | 0.070 | 0.8844 | 74.65 | 97.75 | 26.38 | 99.81 | 0.847 / 0.060 | 9.60 | 8.46 | 3161/1178/39/422 |
| CatBoost | Macro-tuned | 0.362 | 0.8702 | 80.08 | 85.94 | 55.81 | 98.50 | | 23.79 | 21.80 | 2951/707/249/893 |
| CREDITAGENT (GRPO) | Learned | – | – | 83.32 | 90.50 | 66.25 | 93.20 | – | – | – | – |

### D.5 HARD-STOP ABLATION

In the evaluated system the model-side hard-stop fires when the fraud specialist score satisfies $r_{\text{fraud}} \leq 3$. Litigation and identity-inconsistency hard-stops are applied by the institution from upstream bureau flags and lie outside the model. Appendix B proves that a fired veto always yields $\hat{Y} = 1$. We measure the contribution of the hard-stop rule in Eq. (4) by regenerating per-case decisions from the trained policy with three seeds. The hard-stop OFF setting reuses the same fused scores and removes only the final veto. Table 16 shows that the hard-stop raises accuracy by 0.41 points and BEC by 0.39 points. The whole effect is on G3 (+1.22 points), the group of rejected applicants who later defaulted, and G1 and G2 are unchanged. The hard-stop therefore improves rejection stability on high-risk cases but is not the main source of the overall gain of CREDITAGENT.

Table 16: GRPO fusion with and without the fraud hard-stop (three seeds). OFF reuses the same fused scores and removes only the veto.

| **Metric** | **Hard-stop ON** | **Hard-stop OFF** | **ON − OFF** |
|---|---|---|---|
| BEC (%) | 76.47 | 76.08 | +0.39 |
| Accuracy (%) | 83.32 | 82.91 | +0.41 |
| G1 / G2 / G3 (%) | 90.50 / 66.25 / 93.20 | 90.50 / 66.25 / 91.98 | 0.00 / 0.00 / +1.22 |

As a separate sanity check, we apply the same veto rule ($r_{\text{fraud}} \leq 3$) to a committee of Qwen3 specialist agents (Table 17). The veto raises accuracy by 0.57 points and G3 accuracy by 2.81 points, which agrees in direction with Table 16. This committee is a different system, so the check neither estimates nor bounds the effect of the veto inside the trained GRPO system. The litigation and identity-inconsistency hard-stops of the production system are applied by the institution from upstream bureau flags and are outside the model and outside both ablations.

Table 17: Fraud-veto sanity check on a Qwen3 specialist committee.

| **Configuration** | **BEC** | **Acc. (%)** | **G3 (%)** |
|---|---|---|---|
| Qwen3 committee, veto ON | 0.817 | 71.65 | 92.89 |
| Qwen3 committee, veto OFF | 0.800 | 71.08 | 90.08 |
| Difference (ON − OFF) | +0.017 | +0.57 | +2.81 |

### D.6 EVALUATION GROUPS

The 6,000 evaluation cases are split into three groups of 2,000 cases (Table 2, bottom). G1 contains applicants who were approved and later defaulted, so the correct decision is to reject and the business impact is avoided capital loss. G2 contains applicants who did not default but whose profiles fall in a low-approval region, so the correct decision is to approve and the business impact is interest income. G3 contains rejected applicants whose later outcome shows default, so the correct decision is to reject and the business impact is rule adherence. The evaluation set is therefore risk-enriched by construction, and G2 contains the approval-boundary cases.

### D.7 LABEL CONSTRUCTION FOR REJECTED APPLICANTS

G3 consists of applicants whom the institution rejected, so their repayment outcomes are not observed from the institution's own loan book. Their labels are inferred from external records. Table 18 summarizes the protocol. We do not treat missing agent input fields as missing outcome labels, and we do not claim that all rejected outcomes are directly observed. Coverage is 100% within G3 because unmatched rejected applicants were removed before sampling; the match rate in the source population is retained by the institution and is not reported here, which is a limitation of the label construction.

Table 18: Protocol for inferring outcomes of rejected applicants (G3).

| Item | Content |
|---|---|
| Matching source | External credit-bureau follow-up and cross-institution repayment records. |
| Matching key | Hashed national ID and mobile number, linked to PBOC credit-report fields (e.g., `zx2loancompanybalcnt2`, `zx2consumloanmaxcontractmoney`), third-party scores (e.g., Bairong and Bairong 3-month multi-head scores), cross-institution fields (e.g., hujin default score and number of lending institutions), and contact-graph fields. |
| Observation window | 6 and 12 months after the application, including 6-month delinquency and inquiry fields, 180-day bank-flow features, and 3-month M1 events. |
| Coverage | 100% within G3; unmatched rejected applicants were removed before sampling. The source-population match rate is not reported. |
| Unresolved outcomes | Counted during construction, excluded from G3, and never imputed as defaults. |

### D.8 Comparison with Public Credit Datasets

Table 19 compares the evaluation data with commonly used public credit datasets. Public datasets contain fewer features and usually omit credit-bureau and anti-fraud fields, which are the inputs that the Layer 2 specialists reason over. The dataset used in this paper contains these fields, together with outcome labels for rejected applicants.

Table 19: Comparison of the evaluation data with public credit datasets.

| Dataset | Features | Bureau fields | Anti-fraud fields | Labels |
|---|---|---|---|---|
| UCI German Credit | 20 | ✗ | ✗ | Binary |
| Give Me Some Credit | 10 | ✗ | ✗ | Binary |
| LendingClub | ∼150 | Stripped | ✗ | Loan status |
| PKDD / HMEQ | ∼10–20 | ✗ | ✗ | Binary |
| Ours | 100+ | ✓ | ✓ | Repayment + bureau follow-up |

### D.9 Transfer Scope and Practical Use

The system is used with live, incremental data at the originating institution, and it has been applied to new data sources in trials for overseas business. The elements that we expect to transfer are the decomposition into specialist agents and the learned fusion mechanism. Feature mappings, prompts, hard-stop rules, cost ratios, and operating thresholds are specific to the lender and must be adapted and validated for each target setting. We therefore claim portability with recalibration, not zero-shot transfer. We have not quantified the economic benefit at portfolio level. Under the 3:1 cost ratio, approving a later defaulter costs three times as much as rejecting a creditworthy applicant, and the gain of CreditAgent over tuned tabular models is concentrated on G2, where it corresponds to fewer false rejections of applicants who repay. The system uses about 20.15 s and 12,993 tokens per case, compared with 5.91 s and 2,677 tokens for a single-model pipeline (Appendix H), so it is intended for non-real-time review of high-risk cases and not as a replacement for fast tabular screening.

## E LIMITATIONS

Our framework has several limitations. First, the system was evaluated on 6,000 credit cases from one institution in 2023. Its performance may not transfer to other regions, regulatory environments, or economic conditions without recalibration of feature mappings, prompts, hard-stop rules, cost ratios, and decision thresholds (Appendix D.9). Second, the hierarchical architecture has about $3.4\times$ the latency (20.15 s vs. 5.91 s per case) and about $4.9\times$ the token consumption (12,993 vs. 2,677 tokens per case) of a single-model pipeline (Appendix H). This is acceptable for non-real-time credit review but limits real-time use. The SFT and GRPO stages also require computational resources that may not be available to smaller institutions. Third, BEC is part of the Layer 3 reward and is also an evaluation metric; we therefore report accuracy, per-group accuracy, and standard diagnostics, and Appendix D.3 shows that the consistency term does not explain differences between fusion variants. BEC values recomputed from the released evaluation code are not on the same scale as the BEC column of Table 3 (Appendix D.2). Fourth, per-case AUC, ECE, and confusion matrices for the GRPO system are not available. Fifth, outcomes of rejected applicants are inferred from external records, and the source-population match rate is not reported (Appendix D.7). Sixth, cost-sensitive tabular models with tuned thresholds are competitive on aggregate metrics, and the advantage of CREDITAGENT is concentrated on G2 boundary cases (Appendix D.4). Seventh, the structured decision trail is produced by design, but its quality is not evaluated directly. Finally, the fairness analysis is exploratory and does not support fairness claims.

# F LAYER 2 AGENTS TRAINING DATA DETAILS




Figure 7: Data Distribution Details for Layer 2 Specialized Agents

Layer 2 of the CREDITAGENT framework implements a **Multi-Agent Domain Expertise Review** system, where specialized reasoning agents are deployed across six credit review dimensions. To ensure expert-level grounding, each agent $a_i$ is fine-tuned using **Supervised Fine-Tuning (SFT)** on a custom *Credit-Instruct* dataset.

## F.1 AGENT ARCHITECTURE

The Layer 2 system employs a **State-Aware Routing Framework** to manage dependencies between specialized agents $\mathcal{A} = \{a_1, a_2, \ldots, a_6\}$. Each SFT-aligned agent $a_j$ produces an output tuple:

$$o_j = \{r_j, e_j\} \tag{22}$$

where:

- $r_j \in \{1, \ldots, 10\}$ is the scalar creditworthiness score for a specific dimension (e.g., Solvency; higher is safer), derived via Risk-Oriented Textualization (ROT); Layer 3 uses the normalized risk $q_j = 1 - r_j/10$
- $e_j \in \mathcal{E}^*$ is the **Evidence Chain** in natural language, constrained by a JSON schema

The global context state is updated recursively as:

$$S_{t+1} = [S_t \| o_{t+1}] \tag{23}$$

with $S_0 = \emptyset$, enabling downstream agents to condition their responses on the semantic outputs of upstream agents.

## F.2 CREDIT-INSTRUCT DATASET CONSTRUCTION

The *Credit-Instruct* dataset aligns the model's latent space with expert reasoning paths (Chain-of-Thought) and domain-specific knowledge injection. The dataset comprises triples of:

$$\mathcal{D} = \{(\text{input}, \text{CoT_reasoning}, \text{decision})\} \tag{24}$$

### F.3 Risk-Oriented Textualization (ROT)

To address semantic distortion during the mapping of structured financial features to the linguistic reasoning space, we define a textualization operator $\mathcal{T}$ that maps the agent-relevant refined feature slice $\mathbf{x}'_j \subseteq \mathbf{x}'$ to a structured semantic prompt. For each agent $a_j$, a risk-discretization function is introduced:

$$g : \mathbb{R} \to \{\text{Low}, \text{Medium}, \text{High}\} \tag{25}$$

defined by business-driven quantiles. The resulting prompt incorporates causal templates:

$$P_j = \mathcal{T}(\mathbf{x}'_j, g(\mathbf{x}'_j)) \tag{26}$$

The framework utilizes PEFT via QLoRA to adapt the core backbone model, optimizing the objective:

$$\min_{\theta} \mathcal{L}_{CE}(y, \hat{y}) + \beta \cdot \text{DPO}(O_A, O_B) \tag{27}$$

where $y$ is the ground-truth supervision label for the current agent task, $\hat{y}$ is the predicted label, $\beta$ controls the preference-alignment strength, and $\text{DPO}(O_A, O_B)$ represents the Direct Preference Optimization loss based on expert pairwise rankings of a preferred output $O_A$ over a dispreferred output $O_B$. The Layer 2 committee consists of six specialized agents, each focusing on a distinct credit assessment dimension in Table 2.

### F.4 Data Distribution

Figure 7 illustrates the training data distribution across the six specialized agents, showing the score distributions for Basic, History, Solvency, Willingness, Fraud, and Macro dimensions. The heterogeneous score distributions reflect the varying risk profiles encountered in real-world credit assessment scenarios, ensuring robust agent training across diverse risk regimes.

## G MODEL CONFIGURATIONS

During the experimental phase of this study, we evaluated a wide range of large language models. Table 20 summarizes the model names, developers, model scales, and the corresponding methods employed for each model.

Table 20: Model architectures and parameter scales evaluated in the benchmark.

| **Model Name** | **Developer** | **Scale/Type** | **Access Method** |
|---|---|---|---|
| *Open-Source* | | | |
| XuanYuan 2 | Duxiaoman-DI | 70B | Local (vLLM) |
| Fin-R1 | Financial AI Lab | 32B (Distilled) | Local (vLLM) |
| Seed OSS Instruct | ByteDance-Seed | 36B | Local (vLLM) |
| Llama 3.1 | Meta | 8B | Local (vLLM) |
| Llama 3.3 | Meta | 70B | Local (vLLM) |
| Qwen 3 Instruct | Alibaba | 4B, 30B | Local (vLLM) |
| Qwen 3 Thinking | Alibaba | 30B | Local (vLLM) |
| DeepSeek-R1-Distill-Qwen | deepseek-ai | 32B | Local (vLLM) |
| DeepSeek-R1-Distill-llama | deepseek-ai | 8B | Local (vLLM) |
| *Proprietary* | | | |
| Gemini 3 Flash Preview | Google | Unknown | Vertex AI API |
| Claude 4.5 Sonnet | Anthropic | Unknown | Anthropic API |
| GPT 5.1 | OpenAI | Unknown | OpenAI API |

The models evaluated in this study include a diverse set of open-source models, along with three carefully selected, widely adopted closed-source models. Testing for the open-source models was primarily conducted using vLLM-based deployment. The hardware specifications of the testing machines are detailed in Table 21.

Table 21: Inference parameters and environment details.

| **Parameter** | **Configuration Value** |
|---|---|
| GPU Hardware | 8 $\times$ NVIDIA H800 (80GB) |
| Operating System | Ubuntu 22.04 LTS |
| Inference Engine | vLLM (v0.6.3) / CUDA 12.4 |
| Decoding Strategy | Greedy Search ($T = 0$) |
| Max Context Length | 16,384 Tokens |
| Precision | Bfloat16 |

## H Details of Token Counts and Latency

In the experiments, we quantified the average input and output token counts for representative general purpose single models, as well as for each of the six dimension specific agents and the final decision agent within our credit agent framework. These results are summarized in Table 22. As shown, the Basic Agent, which performs foundational information analysis, exhibits the highest output token count. Furthermore, we benchmarked both total token consumption and inference latency across individual models and the end to end credit agent system; the comparative results are reported in Table 23.

Table 22: Average Tokens Count in End to End Models and Each part of Credit Agents.

| **Baselines** | **Completion Tokens** | **Prompt Tokens** | **Total Tokens** |
|---|---|---|---|
| Single Model | 798.5 | 1878.3 | 2676.8 |
| Basic Agent | 1340.0 | 247.3 | 1587.3 |
| Solvency Agent | 1003.1 | 379.5 | 1382.6 |
| Willingness Agent | 1102.1 | 572.6 | 1674.7 |
| History Agent | 1210.9 | 411.9 | 1622.8 |
| Fraud Agent | 1160.9 | 260.9 | 1421.8 |
| Macro Agent | 873.9 | 514.4 | 1388.3 |
| Final Decision Agent | 805.8 | 3109.2 | 3915.0 |

Table 23: Comparison of overall tokens and time consuming between single model and credit agents.

| **Baselines** | **Overall Tokens** | **Time per Case (s)** |
|---|---|---|
| Single Model | 2676.772 | 5.9074 |
| Credit Agent | 12992.502 | 20.1459 |

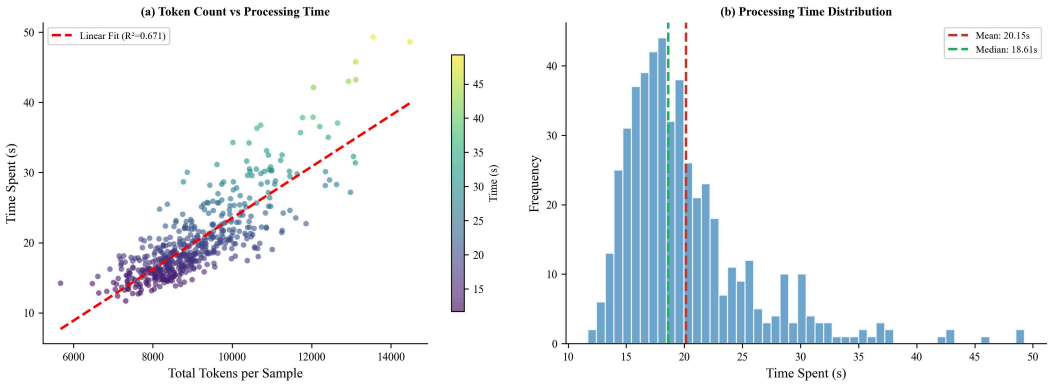



Figure 8: The time analysis of agent output (a) is Token Count vs. Processing Time (b) is Processing Time Distribution

Figure 8 presents the time analysis of agents' output. As shown in (a), there is a strong positive correlation between total tokens per sample and processing time, with a linear fit yielding $R^2 = 0.671$. The histogram in (b) reveals that the processing time follows a right-skewed distribution. Most samples are processed within 15–25 seconds.

# I CASE STUDY

In this section, we present four representative case studies to illustrate the application of our multi-agent credit risk assessment framework in diverse borrower scenarios. These cases are selected to cover a spectrum of credit profiles, ranging from high-risk applicants with potential delinquency issues to low-risk premium borrowers. Each case includes the original input data (borrower information, repayment assessments, credit history, etc.), ground truth labels (final rating, delinquency probability, and decision), and detailed scoring from specialized agents across key dimensions such as repayment ability, willingness, credit history, fraud risk, and application reasonableness (the output dimension of the Macro agent). The agents' analyses culminate in a final prediction, overall rating, identified risks, suggested measures, and decision recommendation.

Case Study 1 exemplifies a high-risk scenario where the borrower exhibits missing basic information, multi-lender borrowing, large unsettled balances, frequent inquiries, and contact anomalies. Despite a ground truth approval with "D" rating and "High" delinquency probability, the agents recommend rejection due to elevated risks.
Case Study 2 represents a premium low-risk profile with high credit limits, minimal inquiries, no overdues, and low fraud indicators. The ground truth and agents align on an "A" rating, "Low" delinquency probability, and approval, though noting minor issues like missing basics and high interest rates.
Case Study 3 highlights another high-risk case with extremely low income, severe multi-lender borrowing, frequent inquiries, and high credit utilization. Both ground truth and agents concur on a "D" rating, "High" delinquency probability, and rejection.
Case Study 4 demonstrates a good mid-tier profile with moderate income, strong repayment ability, low inquiries, no overdues, and minimal fraud risk. The ground truth and agents support a "B" rating, "Low" delinquency probability, and approval, with the primary concern being incomplete basic information.

These cases underscore the framework's ability to integrate multi-dimensional data for nuanced risk evaluation, balancing quantitative metrics with qualitative insights. They also reveal common challenges, such as data incompleteness, and the importance of agent collaboration in decision-making. The detailed illustrations for each case category are provided in Figures 9–10 for Case 1, Figures 11–12 for Case 2, Figures 13–14 for Case 3, and Figures 15–16 for Case 4.

**Case Study 1-1**

**Question:** You are a highly experienced credit risk expert with deep expertise in comprehensive credit risk evaluation, drawing on multi-dimensional borrower information to deliver accurate and insightful assessments.
Based on the applicant's profile and provided details, please provide your professional prediction and judgment as follows:

- What is your predicted **Final Credit Rating** for this applicant? (e.g., A, B, C, D, E, etc.)
- What is your estimated **Delinquency Probability** level? (e.g., Low / High)

**Ground Truth:** Final rating: "D"; Delinquency probability: "High"; Case decision: "Approve".

**Original Input Data**

**Borrower Basic Information**:
Credit inquiry authorization time: 20230517.0
Whether the user has authorized credit inquiry (1: Yes, 0: No, -9999: System default): 1.0
...
**Repayment Ability Assessment**:
Account credit limit: 4500000.0
Account limit (cash loan): 4500000.0
Account available limit (cash loan): 1501626.0
...
**Repayment Willingness Assessment**:
Number of inquiries in the last 1 month (numeric): 3.0
Total number of inquiries in the last 1 month: 3.0
Number of inquiring institutions in the last 2 months (loan approval): 5.0
...
**Credit History and Lending Behavior**:
Maximum credit card utilization rate (numeric): 1.0
Credit card limit utilization rate (numeric form): 1.0079
Total new loans in the last 3 months (numeric): 2100.0
...
**Fraud Risk and Consistency Check**:
Number of different contacts corresponding to contact phone numbers: 2.0
Number of times contacts were filled: 3.0
...
**Other Information**:
Report time: 2023-05-10
Occurrence time: 2023-05-24
APR interest rate: 2.34
LPR interest rate: 3.65
...

Figure 9: Case study 1-1

**Case Study 1-2**

**Different Agents' Scoring Content**

**Borrower Basic Information**: Score 1. Basis: "Borrower identity, demographic characteristics, customer labels, loan experience and other basic information are missing; only credit authorization is 'Yes'. Lack of basic information leads to insufficient risk assessment basis, resulting in high risk." Key evidence: ["No relevant data", "Whether the user has authorized credit inquiry: 1.0"].

**Repayment Ability**: Score 6. Basis: "High credit limits (Manyi Loan 4.5M etc.), but large number of lending institutions (10) and high unsettled balance (Manyi Loan nearly 3M), resulting in medium overall repayment capacity." Key evidence: ["Manyi Loan credit limit: 4500000.0", "Number of lending institutions - overall: 10.0"].

**Repayment Willingness**: Score 3. Basis: "Presence of multi-lender borrowing risk, frequent loan approval inquiries, and active small loan client status indicate strong capital demand and high repayment willingness risk." Key evidence: ["Bairong 3-month Multi-lender Borrowing Score: 14.0", "Number of loan approval inquiries in the past 24 months: 32.0", "Whether the borrower is an active client of small loan institutions: 1.0", "Number of applications to partner general consumption installment platforms: 7.0"].

**Credit History**: Score 8. Basis: "No historical overdues (0 overdue months in past 6/3 months, 0 M1 occurrences), excellent repayment performance; however, credit card over-limit usage and multi-lender borrowing (many institutions, many small-amount loans) exist, resulting in medium-low risk." Key evidence: ["Number of months with overdue loans and credit cards in the past 6 months: 0.0", "Highest overdue months for loan accounts in the last 3 months: 0.0", "Credit card limit utilization rate: 1.0079", "Number of deduplicated institutions for personal unsecured consumption loans: 16.0", "Number of loans with limit $le$ 30,000: 10.0"].

**Fraud Risk**: Score 2. Basis: "Abnormal contact information (first contact filled as second contact 3 times), indicating association information anomaly and fraud/consistency risk." Key evidence: ["Number of times the first contact was filled as the second contact: 3.0"].

**Application Reasonableness**: Score 2. Basis: "Frequent inquiries (46 times), high executed interest rate (2340.0), multi-lender credit , and zero successful cash loan disbursements show significant risk characteristics." Key evidence: ["Total number of various inquiries: 46.0", "Executed interest rate: 2340.0", "Hours since last Manyi Loan credit approval: 7588.0", "Associated transaction 2 real-time incremental credit component issuance time: 20230505.0", "Total number of successful cash loan disbursements on natural day: 0.0"].

**Final Summary Scores and Decisions**

**Prediction**: Delinquency probability: "High";
**Overall Final Rating**: "D" (High Risk).
**Main Risks**: "Missing basic information, multi-lender borrowing, large unsettled balance, frequent inquiries, contact information anomalies".
**Suggested Measures**: "Supplement borrower identity and demographic characteristics and other basic information; pay attention to debt situation and evaluate stability of repayment sources; strengthen monitoring of multi-lender borrowing and frequent inquiries; verify authenticity of contact information".
**Final Decision**: "Reject", with high delinquency possibility.

Figure 10: Case study 1-2

**Case Study 2-1**

**Question:** You are a highly experienced credit risk expert with deep expertise in comprehensive credit risk evaluation, drawing on multi-dimensional borrower information to deliver accurate and insightful assessments.
Based on the applicant's profile and provided details, please provide your professional prediction and judgment as follows:

- What is your predicted **Final Credit Rating** for this applicant? (e.g., A, B, C, D, E, etc.)
- What is your estimated **Delinquency Probability** level? (e.g., Low / High)

**Ground Truth:** Final rating: "A"; Delinquency probability: "Low"; Case decision: "Approve".

**Original Input Data**

**Borrower Basic Information**:
Credit inquiry authorization time: 20230321.0
Whether the user has authorized credit inquiry (1: Yes, 0: No, -9999: System default): 1.0
...
**Repayment Ability Assessment**:
Account credit limit: 19500000.0
Account limit (cash loan): 19500000.0
Account available limit (cash loan): 18400000.0
...
**Repayment Willingness Assessment**:
Number of inquiries in the last 1 month (numeric): 0.0
Total number of inquiries in the last 1 month: 0.0
Number of inquiring institutions in the last 2 months (loan approval): 0.0
...
**Credit History and Lending Behavior**:
Maximum credit card utilization rate (numeric): 0.0
Credit card limit utilization rate (numeric form): 0.0518
Total new loans in the last 3 months (numeric): 0.0
...
**Fraud Risk and Consistency Check**:
Number of different contacts corresponding to contact phone numbers: 0.0
Number of times contacts were filled: 0.0
...
**Other Information**:
Report time: 2022-12-27
Occurrence time: 2023-03-21
APR interest rate: 8.28
LPR interest rate: 3.65
...

Figure 11: Case study 2-1

**Case Study 2-2**

**Different Agents' Scoring Content**

**Borrower Basic Information**: Score 1. Basis: Borrower identity, demographic characteristics, customer labels, loan experience and other basic information are missing; only credit authorization is "Yes". Lack of basic information leads to insufficient risk assessment basis, resulting in high risk. Key evidence: ["No relevant data", "Whether the user has authorized credit inquiry: 1.0"]

**Repayment Ability**: Score 10. Basis: Extremely high credit limits (blaze 20M, Manyi Loan 19.5M), very high available limit (18.4M), extremely low unsettled balance (≈1.11M), resulting in extremely strong repayment capacity and lowest risk. Key evidence: ["Latest credit limit given by blaze: 20000000.0", "Account available limit (cash loan): 18400000.0", "90-day average unsettled amount of cash borrowings (CNY): 15747.25", "Total re-lending balance corresponding to Manyi Loan product: 1100000.0"]

**Repayment Willingness**: Score 10. Basis: Almost no recent loan approval inquiries in past 1/3/6 months, extremely low multi-lender borrowing scores, no signal of capital hunger, extremely strong repayment willingness. Key evidence: ["Number of inquiries in the last 1 month (loan approval, by natural month): 0.0", "Number of inquiries in the last 3 months - loan approval: 0.0", "Bairong Multi-lender Borrowing Score: 8.0", "Bairong 3-month Multi-lender Borrowing Score: 2.0"]

**Credit History**: Score 10. Basis: No historical overdues (0 overdue months, 0 M1 occurrences), extremely low credit card utilization (0.0518), no new loans recently, excellent credit performance. Key evidence: ["Number of months with overdue loans and credit cards in the past 6 months: 0.0", "Maximum credit card utilization rate: 0.0", "Credit card limit utilization rate: 0.0518", "Total new loans in the last 3 months: 0.0"].

**Fraud Risk**: Score 10. Basis: All contact-related fraud/association anomaly indicators are 0, no abnormal situations, lowest fraud risk. Key evidence: ["Number of different contacts corresponding to contact phone numbers: 0.0", "Number of times contacts were filled: 0.0", "Number of times the first contact was filled as the second contact: 0.0", "Number of times the second contact was filled: 0.0"].

**Application Reasonableness**: Score 6. Basis: Limited total inquiries (7 times), high executed/APR interest rate (828.0), multiple applications on the same day (20230321), but supported by excellent qualification in other dimensions, medium risk. Key evidence: ["Total number of various inquiries: 7.0", "Executed interest rate: 828.0", "Most recent cash loan credit application time (variable validation): 20230321.0", "Previous occurrence time: 20230321.0"].

**Final Summary Scores and Decisions**

**Prediction**: Delinquency probability: "Low".
**Overall Final Rating**: "A" (Low Risk / Premium).
**Main Risks**: "Missing basic information; relatively high executed and APR interest rates; multiple applications on the same day".
**Suggested Measures**: "Monitor subsequent credit usage frequency, strengthen post-loan management of high-interest loans, periodically update borrower's income and liability information to reassess repayment capacity changes".
**Final Decision**: "Approve", with low delinquency probability.

Figure 12: Case study 2-2

**Case Study 3-1**

**Question:** You are a highly experienced credit risk expert with deep expertise in comprehensive credit risk evaluation, drawing on multi-dimensional borrower information to deliver accurate and insightful assessments.
Based on the applicant's profile and provided details, please provide your professional prediction and judgment as follows:

- What is your predicted **Final Credit Rating** for this applicant? (e.g., A, B, C, D, E, etc.)
- What is your estimated **Delinquency Probability** level? (e.g., Low / High)

**Ground Truth:** Final rating: "D"; Delinquency probability: "High"; Case decision: "Reject".

**Original Input Data**

**Borrower Basic Information**:
WeChat monthly average recognized income amount: 494.97
Credit inquiry authorization time: 20230604.0
...
**Repayment Ability Assessment**:
Account credit limit: 1000000.0
Account limit (cash loan): 1000000.0
Account available limit (cash loan): 151728.0
...
**Repayment Willingness Assessment**:
Number of inquiries in the last 1 month (numeric): 8.0
Total number of inquiries in the last 1 month: 9.0
Number of inquiring institutions in the last 2 months (loan approval): 8.0
...
**Credit History and Lending Behavior**:
Maximum credit card utilization rate (numeric): 0.9
Credit card limit utilization rate (numeric form): 0.9085
Total new loans in the last 3 months (numeric): 45978.0
...
**Other Information**:
Report time: 2023-08-11
Occurrence time: 2023-09-24
APR interest rate: 2.34
LPR interest rate: 3.45
...

Figure 13: Case study 3-1

**Case Study 3-2**

**Different Agents' Scoring Content**

**Borrower Basic Information**: Score 1. Basis: Basic information (identity, demographics, customer labels, loan experience) has no relevant data, WeChat monthly average recognized income is only 494.97 yuan, income level is extremely low, significant risk characteristics. Key evidence: ["No relevant data", "WeChat monthly average recognized income amount:494.97"]

**Repayment Ability**: Score 2. Basis: Existence of multi-lender borrowing (18 unsettled institutions, 24 non-bank credit loan counts), Manyi Loan re-lending balance 848,272 yuan, available limit only 151,728 yuan, high debt service pressure, higher risk. Key evidence: ["Number of unsettled loan institutions : 18.0", "Number of unsettled non-bank credit loans : 24.0", "Total re-lending balance corresponding to Manyi Loan product (variable validation) : 848272.0", "Account available limit (cash loan) : 151728.0"]

**Repayment Willingness**: Score 2. Basis: Frequent inquiries, high multi-head applications, strong signal of capital hunger, reflecting high risk in repayment willingness. Key evidence: ["Number of loan approval inquiries in the past 24 months:49.0", "Bairong Multi-lender Borrowing Score:84.0", "Bairong 3-month Multi-lender Borrowing Score:36.0", "Number of non-bank institutions queried via ID in the last 1 month:12.0"]

**Credit History**: Score 8. Basis: No historical overdues (0 overdue months in past 6/3 months, 0 M1 occurrences), but extremely high credit card/credit limit utilization (90Key evidence: ["Number of months with overdue loans and credit cards in the past 6 months:0.0", "Maximum overdue months for loan accounts in the last 3 months:0.0", "Maximum credit card utilization rate:0.9", "Credit card limit utilization rate:0.9085", "Number of loans with limits of 3W and below:35.0", "Number of unique institutions for personal consumption unsecured loans:25.0"].

**Fraud Risk**: Score 9. Basis: No confusion in contact roles, filling counts and number of contacts corresponding to phone numbers show no abnormalities, low fraud/consistency risk. Key evidence: ["Number of times the first contact was filled as the second contact:0.0", "Number of times contacts were filled:2.0", "Number of different contacts corresponding to contact phone numbers:2.0"].

**Application Reasonableness**: Score 2. Basis: Frequent inquiries (58 times), high executed interest rate (2340.0) and multi-head applications, higher risk. Key evidence: ["Total number of various inquiries : 58.0", "Executed interest rate : 2340.0", "Most recent cash loan credit application time (variable validation) : 20230924.0"].

**Final Summary Scores and Decisions**

**Prediction**: Delinquency probability: "High".
**Overall Final Rating**: "D" (High Risk).
**Main Risks**: "Extremely low income level, severe multi-lender borrowing, frequent inquiries reflecting capital hunger, extremely high credit card utilization rate".
**Suggested Measures**: "1. Require more detailed income proof to verify the authenticity and stability of WeChat monthly income; 2. Strictly review unsettled loan institutions and counts to assess debt-bearing capacity; 3. Focus on reasons for frequent inquiries, require explanation of fund uses, judge if borrowing to repay loans; 4. Confirm reasonableness of executed interest rate to avoid excessive rates increasing burden".
**Final Decision**: "Reject", with high delinquency probability.

Figure 14: Case study 3-2

**Case Study 4-1**

**Question:** You are a highly experienced credit risk expert with deep expertise in comprehensive credit risk evaluation, drawing on multi-dimensional borrower information to deliver accurate and insightful assessments.
Based on the applicant's profile and provided details, please provide your professional prediction and judgment as follows:

- What is your predicted **Final Credit Rating** for this applicant? (e.g., A, B, C, D, E, etc.)
- What is your estimated **Delinquency Probability** level? (e.g., Low / High)

**Ground Truth:** Final rating: "B"; Delinquency probability: "Low"; Case decision: "Approve".

**Original Input Data**

**Borrower Basic Information**:
Credit inquiry authorization time: 20230604.0
Latest monthly income level (numeric): 2100.0
...
**Repayment Ability Assessment**:
Account credit limit: 2550000.0
Account limit (cash loan): 2550000.0
Account available limit (cash loan): 2550000.0
...
**Repayment Willingness Assessment**:
Number of inquiring institutions in the last 1 month (numeric): 2.0
Number of inquiries in the last 1 month (numeric): 2.0
...
**Credit History and Lending Behavior**:
Maximum credit card utilization rate (numeric): 0.0
Credit card limit utilization rate (numeric form, numeric): 0.0059
Total new loans in the last 3 months (numeric): 80000.0
...
**Fraud Risk and Consistency Check**:
Number of different contacts corresponding to contact phone numbers: 0.0
Number of times contacts were filled: 0.0
...
**Other Information**:
Report time: 2023-04-26
Occurrence time: 2023-04-15
APR interest rate: 12.6
LPR interest rate: 3.65
...

Figure 15: Case study 4-1

**Case Study 4-2**

**Different Agents' Scoring Content**

**Borrower Basic Information**: Score 1. Basis: Basic information (identity, demographics, customer labels, loan experience) no relevant data, only monthly income 2100.0, information missing leads to insufficient risk judgment, risk biased high. Key evidence: ["No relevant data", "Latest monthly income level:2100.0"]

**Repayment Ability**: Score 10. Basis: Credit limits and available limits are all 2550000.0, no liabilities and expenditures or other negative information, extremely strong debt repayment ability, extremely low risk. Key evidence: ["Latest credit limit given by blaze:2550000.0", "Account available limit (cash loan):2550000.0"]

**Repayment Willingness**: Score 9. Basis: Inquiry and application behavior frequency moderate, no obvious capital hunger signals, Bairong multi-lender borrowing score and 3-month multi-lender borrowing score are both low risk, reflecting good repayment willingness. Key evidence: ["Bairong multi-lender borrowing score:27.0", "Bairong 3-month multi-lender borrowing score:8.0", "Number of inquiring institutions in the last 1 month:2.0", "Number of inquiries in the last 1 month:2.0"]

**Credit History**: Score 10. Basis: No overdue records in historical credit behavior, extremely low credit card and credit card limit utilization rates, good repayment performance, excellent qualifications. Key evidence: ["Highest overdue months for loan accounts in the last 3 months (numeric) : 0.0", "Worst status for credit card overdue within 30 days (string) : N", "Maximum credit card utilization rate (numeric) : 0.0", "Credit card limit utilization rate (numeric form, numeric) : 0.0059"]

**Fraud Risk**: Score 10. Basis: All contact-related fraud/association anomaly indicators are 0, no abnormal situations, lowest fraud risk. Key evidence: ["Number of different contacts corresponding to contact phone numbers:0.0", "Number of times contacts were filled:0.0", "Number of times the first contact was filled as the second contact:0.0", "Number of times the second contact was filled:0.0", "Total number of times the first contact was filled:0.0", "Total number of times the second contact was filled:0.0"]

**Application Reasonableness**: Score 6. Basis: Recent credit application, relending distribution and credit extension behaviors exist, no obvious abnormal fluctuations, general qualifications. Key evidence: ["User's most recent cash loan credit application time (variable validation):20230430.0", "Whether re-lending distribution (1: Yes 0: No):1.0", "Associated user's most recent relending distribution result:10000.0", "Credit application extension type:4.0"]

**Final Summary Scores and Decisions**

**Prediction**: Delinquency probability: "Low".
**Overall Final Rating**: "B" (Good).
**Main Risks**: "Missing borrower basic information, leading to insufficient risk judgment basis".
**Suggested Measures**: "Supplement and collect borrower's identity and demographic characteristics, customer labels and loan experience and other basic information to more comprehensively assess risks".
**Final Decision**: "Approve", with low delinquency probability.

Figure 16: Case study 4-2

# J More Related Work

## J.1 Credit Assessment

Credit assessment (Zhang et al., 2024b; Majumdar et al., 2025; Bhatore et al., 2020) constitutes a fundamental task in financial risk management, wherein institutions evaluate the creditworthiness of individuals or entities to determine their likelihood of fulfilling financial obligations. In the context of loan applications (Wang et al., 2020a; Boz et al., 2018; Mulyanto et al., 2025; Sayed Eslam et al., 2024), credit assessment becomes particularly challenging due to the heterogeneous nature of applicant profiles and the need to balance risk and return optimization. Early credit assessment systems predominantly relied on statistical models (Blackwell & Sykes, 1992; Cameron & Trivedi, 1996) and manually engineered expert knowledge, employing techniques such as logistic regression and scorecard models, which suffered from limited capacity to capture complex non-linear relationships. The advent of machine learning transformed this landscape by offering superior capabilities in automatic feature learning, handling high-dimensional data. Consequently, machine learning techniques (Machado et al., 2025; Puli et al., 2024) have been widely adopted for feature selection, feature fusion, and predicting applicants' credit levels with improved discrimination power. The emergence of LLMs (Kang et al., 2025; Raliphada et al., 2025; Sanz-Guerrero & Arroyo, 2024b) introduces a paradigm shift in credit assessment by leveraging their exceptional generalization capabilities and semantic processing characteristics, which enable the incorporation of a broader spectrum of applicant features while rendering them interpretable. The evolution of credit assessment methodologies is illustrated by the works shown in Table 24.

## J.2 MAS in Specific Domain

As AI agents demonstrate increasingly sophisticated capabilities, complex real-world tasks (Shang et al., 2025; Zhu et al., 2025; Schmidgall et al., 2025) are being delegated to autonomous systems for resolution. However, individual agents face inherent limitations in both functional scope and throughput capacity, which has motivated the adoption of MAS for tackling these challenging problems. Complex real-world tasks often demand diverse expertise (Jansen et al., 2024; Yue et al., 2024), parallel processing capabilities (Ashraf et al., 2025), and coordinated action—requirements (Deshpande et al., 2025; Wang et al., 2025a) that exceed the capabilities of even the most advanced individual agents. This has spurred significant research into MAS, where the collective intelligence emerging from specialized agent collaboration enables the solution of tasks that remain intractable for individual agents. Applications of MAS in specific domains are shown in Table 25.

Table 25: Overview of domain-specific MAS

| Work | Domain | Agent Architecture | Key Mechanisms |
|---|---|---|---|
| LawLuo (Sun et al., 2024) | Legal consultation | Four collaborative agents simulating law firm operations | Role-specific fine-tuning, case graph-based RAG for personalized multi-turn consultation |
| AutoDefense (Zeng et al., 2024) | LLM security | Small open-source LLMs with assigned defensive roles | Collaborative filtering of harmful responses, jailbreak attack defense |
| FinCon (Yu et al., 2024) | Financial decision-making | Manager-analyst hierarchy inspired by investment firms | Self-critiquing mechanisms, conceptual verbal reinforcement |
| EduPlanner (Zhang et al., 2025b) | Educational planning | Evaluator, optimizer, and question analyst agents | Skill-Tree structure, 5-D evaluation module, adversarial collaboration |
| Trans. Analy (Zamojska et al., 2025) | Social interaction | Parent, Adult, and Child ego state agents | Transactional Analysis principles, context-aware psychological dynamics |

Continued on next page

Table 25 – continued from Overview of domain-specific MAS

| Work | Domain | Agent Architecture | Key Mechanisms |
|---|---|---|---|
| MedAgent-Pro (Wang et al., 2025c) | Medical diagnosis | Hierarchical reasoning agents with visual tool integration | Disease-level plan generation via RAG, patient-level personalized reasoning, step-wise reliability verification |
| MedOrch (He et al., 2025b) | Medical reasoning | Modular tool-augmented multi-agent orchestration | Flexible tool integration, transparent traceable reasoning for diagnosis, imaging, and multimodal QA |
| ClinicalAgent (Yue et al., 2024) | Clinical trials | GPT-4 multi-agent system with external tool access | LEAST-TO-MOST decomposition, ReAct reasoning, clinical trial tool integration for outcome prediction |
| DynamiCare (Shang et al., 2025) | Clinical decision-making | Dynamic specialist agents with adaptive team composition | Multi-round interactive loops, iterative information gathering, adaptive strategy adjustment |
| MeNTi (Zhu et al., 2025) | Medical calculators | Meta-tool with nested calling mechanism | Flexible calculator chaining, slot filling and unit conversion, 281 physician-curated medical tools |
| Agent Laboratory (Schmidgall et al., 2025) | Scientific research | Autonomous research agents with human-in-the-loop | End-to-end research automation, literature review to report writing, iterative human feedback integration |
| WebAgent-R1 (Wei et al., 2025) | Web navigation | Multi-turn RL agents with thinking-based prompting | Asynchronous trajectory generation, binary task success rewards, test-time scaling via interactions |
| EMBODIEDBENCH (Yang et al., 2025) | Embodied AI | Multi-environment benchmark agents | Evaluation spanning household tasks to atomic navigation/manipulation, revealing low-level control limitations |
| Guided Search (Zainullina et al., 2025) | Non-serializable envs. | 1-step lookahead agents with trajectory selection | Learned action-value functions, trajectory selection for test-time search in non-restorable states |
| TRAIL (Deshpande et al., 2025) | Agent debugging | Trace evaluation framework for single/multi-agent systems | Formal error taxonomy, 148 human-annotated traces, benchmark for error identification in agent workflows |
| DiscoveryWorld (Jansen et al., 2024) | Scientific discovery | Virtual environment with 24 parametric tasks | 24 parametric tasks across difficulty levels, complete discovery cycle evaluation (hypothesis to conclusion) |
| Agent-X (Ashraf et al., 2025) | Multimodal reasoning | Vision-centric agents for multi-step visual tasks | Step-level assessment of reasoning coherence, tool selection, and task completion over images/videos |
| G-Safeguard (Wang et al., 2025a) | MAS security | Graph neural network-based anomaly detection system | Anomaly detection on utterance graphs, topological intervention for attack remediation across LLM backbones |

### J.3 MAS Architecture

To enable more effective message delivery and task coordination, multi-agent systems have evolved diverse architectural paradigms. Based on inter-agent communication patterns and message exchange mechanisms, these architectures can be categorized into four main types: hierarchical (Rawat et al., 2025; Ghezloo et al., 2025; He et al., 2025a; Liu et al., 2023), decentralized (Zhang et al., 2025a; Chen et al., 2024), centralized (Hu et al., 2023), and shared memory (Rasmussen et al., 2025). Table 26 presents examples for each MAS category.

| Work | Category | Model/Algorithm | Details |
|---|---|---|---|
| Sanz et al. (Sanz-Guerrero & Arroyo, 2024a) | LLMs | LLM-derived features | Extracts textual features complementing traditional variables in P2P lending platforms. Improves predictive performance while emphasizing explainability and fairness in credit decision-making. |
| Feng et al. (Feng et al., 2023) | LLMs | Instruction-tuned LLMs | Matches or exceeds state-of-the-art credit scoring methods while addressing bias concerns. Demonstrates comprehensive knowledge and adaptability for holistic financial risk evaluation. |
| Sanz et al. (Sanz-Guerrero & Arroyo, 2024b) | LLMs | LLM textual analysis | Addresses information gaps in P2P lending through textual analysis. Provides complementary signals to traditional credit variables, improving discrimination between creditworthy and risky borrowers with emphasis on explainability. |
| Tan et al. (Tan & Xie, 2025) | LLMs | GPT-4 + BERT fusion | Combines GPT-4's generative capabilities for latent risk extraction with BERT's bidirectional encoding. Multi-level semantic integration identifies implicit credit risks in unstructured financial documents with superior accuracy and interpretability. |
| Raliphada et al. (Raliphada et al., 2025) | LLMs | BERT, RoBERTa, LLaMA 3.2 | Analyzes financial news sentiment using pre-trained language models. RoBERTa's contextual encoding demonstrates superior risk-relevant sentiment capture. Gradient boosting achieves highest accuracy with structured features. |
| Kang et al. (Kang et al., 2025) | LLMs | GPT-4o | Processes heterogeneous modalities (power time series, financial metrics, textual data) through attention mechanisms and contrastive learning. Integrates physical consumption data with traditional financial information through cross-modal representation optimization. |
| Puli et al. (Puli et al., 2024) | Machine Learning | Neural Networks, Random Forest | Demonstrates superior performance with credit-related, interest rate, and liquidity variables as most informative early warning indicators. |
| Zhang et al. (Zhang et al., 2015) | Machine Learning | Neural Networks | Achieves better generalization and robustness than traditional neural networks by evaluating supply chain relationships and leading enterprise credit status alongside individual firm characteristics. |
| Machado et al. (Machado et al., 2025) | Machine Learning | Random Forest + SHAP | Combines internal banking records with external financial data. Outperforms other ML techniques in prediction accuracy and transparency for detecting early warning signals with SHAP-based explanations. |
| Zhang et al. (Zhang et al., 2024b) | Machine Learning | Ensemble Decision Trees | Leverages feature selection and ensemble decision trees for superior performance in processing large-scale loan application data for accurate risk prediction. |
| Blackwell et al. (Blackwell & Sykes, 1992) | Statistical | Marginal Risk Model | Introduces marginal risk concept varying with outstanding balance. Enables refined credit limit strategies avoiding "all or nothing" approach while optimizing authorization decisions based on account-specific risk contours. |
| Cameron et al. (Cameron & Trivedi, 1996) | Statistical | Count-Duration Models | Provides statistical frameworks for discrete, non-negative integer data. Links count processes to duration analysis for credit scoring and insurance pricing applications. |

Table 24: Evolution of credit assessment methodologies across statistical, machine learning, and LLM-based approaches.

| Architecture | Work | Details |
|---|---|---|
| Hierarchical | DyLAN (Liu et al., 2023) | Dynamic hierarchical architecture with multi-turn interactions. Employs agent selection during inference and early termination for enhanced collaboration efficiency. Uses unsupervised agent importance scoring for automatic team optimization. |
| Hierarchical | LMA for SE (He et al., 2025a) | Applies hierarchical task decomposition in software engineering lifecycle. Divides high-level requirements into sub-tasks assigned to specialized agents, mirroring agile methodologies. Enhances robustness through cross-examination and multi-agent validation, addressing LLM hallucination issues. Scales to complex systems by incorporating additional agents and reallocating tasks dynamically. |
| Hierarchical | PathFinder (Ghezloo et al., 2025) | Four collaborative agents (Triage, Navigation, Description, Diagnosis) emulate pathologist decision-making. Iteratively navigates whole slide images, generates importance maps, produces natural language descriptions of diagnostically relevant patches, and synthesizes findings for final classification with inherent explainability. |
| Hierarchical | Pre-Act (Rawat et al., 2025) | Enhances agent performance by generating multi-step execution plans with detailed reasoning before action. Incrementally refines the plan by incorporating previous steps and tool outputs until final response. Demonstrates substantial improvements in action accuracy and goal completion through fine-tuning on smaller models for practical deployment. |
| Decentralized | DMAS (Chen et al., 2024) | Turn-taking dialogue-based task planning where each robot's LLM agent independently expresses opinions and considers peer feedback. Enables autonomous decision-making while maintaining collective progress toward task completion. |
| Decentralized | MaAS (Zhang et al., 2025a) | Optimizes an agentic supernet—a probabilistic distribution of agent architectures—to dynamically sample query-dependent multi-agent systems. Enables tailored resource allocation based on task difficulty and domain, achieving superior performance with significantly reduced inference costs compared to static, handcrafted multi-agent designs. |
| Centralized | ACORM (Hu et al., 2023) | Single LLM serves as central planner, generating actions for all agents based on global state information. Achieves centralized training with decentralized execution, simplifying MARL planning and enhancing scalability and inference efficiency. |
| Shared Memory | MetaGPT (Hong et al., 2023) | Maintains shared message pool allowing dynamic information observation and extraction. Enables parameter sharing where agents update weights based on new knowledge and synchronize with other agents, optimizing collaboration efficiency. |
| Shared Memory | Zep (Rasmussen et al., 2025) | Introduces Graphiti, a temporally-aware knowledge graph engine that dynamically synthesizes unstructured conversational data and structured business data while maintaining historical relationships. Enables complex temporal reasoning and cross-session information synthesis beyond static document retrieval in RAG frameworks. |

Table 26: Taxonomy of MAS architectures based on communication patterns and coordination mechanisms.

# K PRIOR KNOWLEDGE IN FINANCE ABOUT HIGH-STAKES FINANCIAL DECISION-MAKING

In the domain of financial risk control, decomposing a complex decision system into multiple specialized sub-agents is not merely a technical design choice but is deeply rooted in established financial business logic. The structure of our Layer 2 expert agents directly mirrors the organizational paradigm of institutional credit committees, where domain specialists responsible for a distinct risk dimension collaborate to reach a lending decision.The classical 5C's of Credit framework (Baiden, 2011) including Character, Capacity, Capital, Conditions, and Collateral, constitute the canonical framework for credit analysis, widely adopted in both academic literature and industry practice.

Table 27 summarizes the mapping between classical 5C's credit assessment frameworks and our agent architecture.

Table 27: Mapping between CREDITAGENT Layer 2 Architecture and Classical Credit Assessment Frameworks (5C's of Credit, Baiden 2011).

| **CREDITAGENT Agent** | **Classical 5C's** | **Risk Dimension** | **Primary Function** |
|---|---|---|---|
| Macro Agent | Conditions | Macro-Economic Context | Systematic risk calibration |
| Basic Agent | Character | Identity & Basic Profile | KYB/KYC verification |
| History Agent | Character | Credit History | Behavioral velocity analysis |
| Solvency Agent | Capacity, Capital | Solvency & Asset Quality | Liquidity & capital adequacy |
| Willingness Agent | Character | Repayment Willingness | Intent signal extraction |
| Fraud Agent | Collateral* | Fraud & Consistency | Adversarial detection |

*In unsecured consumer lending, Collateral is reinterpreted as information consistency verification.

Character in Baiden's framework encompasses willingness and determination to meet a loan obligation, discovered through investigation into the customers' payment habits, and the way they manage their business affairs. This multi-faceted definition naturally decomposes into three specialized agents including the Basic Agent (identity authenticity), the History Agent (historical payment patterns), and the Willingness Agent (current repayment intent). Capacity (cash flow ability) and Capital (equity reserves indicating stakeholder commitment) are jointly assessed by the Solvency Agent, as these dimensions are inherently correlated in consumer credit contexts. The following subsections detail the prior financial knowledge integrated into each agent, specifying the primary functions performed and their grounding in established credit assessment principles.

## K.1 MACRO-ECONOMIC CONTEXT AND SYSTEMATIC RISK

Credit risk is dynamic since the same borrower profile may represent acceptable risk during economic expansion, while elevated risk during economic contraction. Unlike traditional credit scoring process that assume distributional stationarity, our Macro Agent processes PESTEL(Political, Economic, Social, Technological, Environmental, Legal) indicators including interest rate trajectories, industry volatility indices, unemployment trends, and regulatory regime shifts to compute a market-sensitivity credit score.

The underlying financial prior knowledge stems from well-established macro-credit linkages. For example, rising interest rates mechanically increase debt-servicing burdens for variable-rate borrowers, while depressing asset valuations that serve as implicit collateral at the same time. In addition, unemployment shocks exhibit asymmetric effects that job loss triggers immediate payment stress, while re-employment restores capacity only with significant lag. These macro-credit transmission mechanisms constitute the domain knowledge that enables the Macro Agent to calibrate systematic risk adjustments across heterogeneous economic environments.

## K.2 IDENTITY VERIFICATION AND BASIC PROFILING

Know Your Borrower (KYB) and Know Your Customer (KYC) compliance represents the foundational layer of credit risk assessment, serving both regulatory mandate and as the first line of defense against synthetic identity fraud, which is a rapidly growing threat vector in digital lending. The Basic Agent performs deterministic validation before probabilistic assessment begins, processing structured identity data to evaluate document consistency, contact information stability, employment verification, and digital footprint patterns. This dimension directly operationalizes the "Character" component of the 5C's framework, as (Baiden, 2011) notes that character is "generally discovered through interview and investigation" into the customer's identity and business affairs.

### K.3 Credit Behavior History and Velocity Analysis

Historical repayment behavior is one of the most predictive feature in consumer credit modeling. However, raw delinquency counts fail to capture the velocity of credit deterioration, a critical leading indicator distinguishing stable borrowers experiencing isolated hardship from those on deteriorating trajectories. The History Agent operationalizes this temporal dimension through roll rate analysis and behavioral velocity scoring.

### K.4 Solvency, Liquidity, and Asset Quality

Repayment capacity assessment requires distinguishing between solvency and liquidity—short-term cash availability. A borrower may be solvent but illiquid, or liquid but insolvent. The Solvency Agent jointly operationalizes the "Capacity" and "Capital" dimensions of the 5C's framework through dual-axis evaluation. This distinction reflects established risk management practice where affordability assessment determines whether the borrower can sustain payments from regular income streams, while capital adequacy provides a buffer against income disruption (Bhatore et al., 2020).

### K.5 Repayment Willingness and Behavioral Signals

Willingness to pay are critical in credit risk. The Willingness Agent extracts intent signals from application behavior, inquiry patterns, and multi-lender dynamics that traditional scorecards systematically underweight. This agent directly addresses the core definition of "Character" in the 5C's framework (Baiden, 2011) defines Character as the customers' willingness and determination to meet a loan obligation, discovered through investigation into payment habits and the way they respond to adversity.

### K.6 Fraud Detection and Consistency Verification

First-party fraud (applicant misrepresentation) and third-party fraud (identity theft, synthetic identity) represent non-credit-risk loss vectors that bypass traditional scoring entirely. Fraud signals are particularly susceptible to dilution when processed through monolithic systems, necessitating dedicated detection mechanisms (Gandhar et al., 2024). The Fraud Agent implements adversarial detection by treating each application as potentially hostile input requiring multi-layered verification. In unsecured consumer lending where physical collateral is absent, this agent reinterprets the "Collateral" dimension of the 5C's framework as information consistency verifications.

# L EXPERT REVIEW SUMMARY

The design and validation of our CREDITAGENT framework benefited significantly from consultations with senior professionals at major financial institutions in China. Through structured interviews with industry specialists, we gathered critical insights into credit committee operations, multi-dimensional risk assessment, and automated decisioning systems. These expert perspectives directly shaped the development of our hierarchical multi-agent architecture. Our consultation panel comprised seasoned specialists across credit underwriting, risk modeling, fraud investigation, and macro-economic analysis, each with 6 to over 15 years of hands-on experience as Table 28 shows.

| **Summarized Professional Profiles** |
|---|
| **Profile A:** Senior Credit Committee Director (15+ yrs). Expert in credit committee governance and cross-functional decision synthesis at a national commercial bank. |
| **Profile B:** Risk Modeling Lead (10+ yrs). Specializing in scorecard development, feature engineering, and model validation at a leading fintech company. |
| **Profile C:** Fraud Investigation Manager (8+ yrs). Focused on synthetic identity detection, network analysis, and adversarial pattern recognition. |
| **Profile D:** Macro-Economic Risk Analyst (7+ yrs). Expert in industry cycle analysis, regulatory impact assessment, and systematic risk calibration. |
| **Profile E:** Consumer Credit Underwriter (6+ yrs). Skilled in cash flow analysis, DTI assessment, and affordability verification for unsecured lending. |
| **Profile F:** Collections Strategy Director (10+ yrs). Expert in delinquency prediction, roll rate analysis, and recovery optimization. |

Table 28: Expert Consultation Panel Profiles

Key findings from our expert consultations highlighted several critical aspects that directly informed the CREDITAGENT architecture:

**Hierarchical Decision Structure.** Expert A emphasized that institutional credit committees operate through structured deliberation rather than holistic judgment: A single reviewer cannot hold all dimensions in mind simultaneously. This insight directly motivated our three-layer architecture separating data governance, domain expertise, and decision fusion.

**Dimensional Decomposition Necessity.** Experts D, E, and F independently confirmed that credit risk assessment requires distinct analytical methodologies across dimensions. Expert D noted that Macro factors require forward-looking scenario analysis, while historical behavior requires backward-looking pattern recognition. These are fundamentally different cognitive tasks. This validated our design of six specialized agents rather than a monolithic reasoning approach.

**Hard-Stop Mechanism for Catastrophic Risk.** Expert C advocated for deterministic rejection rules. Fraud losses are unrecoverable. Positive signals from other dimensions should not override a verified fraud indicator. This principle directly shaped the veto authority granted to our Fraud Agent.

**Asymmetric Cost Structure.** Expert F provided quantitative context for misclassification costs. In unsecured consumer lending, a false negative, which is approving a defaulter, typically costs 3 times more than a false positive, depending on product and recovery rates. This insight informed the asymmetric penalty structure in our business efficiency coefficient design.

**Information Consistency as Collateral Proxy.** Expert C observed that in unsecured lending, "the degree to which application materials can be cross-validated serves the same risk-mitigation function as physical collateral in secured lending." This perspective supported our reinterpretation of the 5C's Collateral dimension as consistency verification.

**Dynamic Threshold Adjustment.** Expert B highlighted limitations of static decision thresholds. Optimal approval thresholds shift with economic cycles, portfolio composition, and competitive dynamics. A system that

cannot adapt its decision boundary will either over-reject during expansion or under-reject during contraction. This observation motivated our GRPO-based threshold optimization in Layer 3.

These expert insights were instrumental in developing CREDITAGENT'S hierarchical structure, ensuring alignment with institutional credit committee practices while enabling the scalability and auditability required for production deployment.

# M DATASET DETAILS

This dataset is derived from real Chinese personal credit reports. It contains 6,000 samples, each corresponding to a genuine cash loan or consumer finance credit application scenario. Every sample is labeled with a forward-looking delinquency outcome, making it suitable for evaluating models' default prediction and deep credit risk reasoning capabilities using purely structured credit bureau data.

Unlike transaction-log or consumption-flow benchmarks, this dataset focuses on classic credit bureau fields plus selected external enrichment variables. It tests whether models can extract hidden risk contradictions and behavioral patterns from multi-lender borrowing, inquiry spikes, repayment history, credit utilization, fraud signals, and related dimensions.

## M.1 CORE RISK ASSESSMENT DIMENSIONS (MAIN RISK DOMAINS)

In the risk control practice of consumer finance and online cash loans in China, core features are typically grouped into the following six primary risk assessment dimensions. These dimensions collectively cover almost all key predictive signals — from borrower profile to behavioral trajectory, credit performance, fraud indicators, and contextual auxiliary information — forming a relatively complete and layered risk evaluation framework:

- **Borrower Basic Information** Mainly includes identity verification, income level, credit report authorization status, etc. Due to heavy desensitization of strong identity fields (name, ID number, plaintext mobile number), this dimension tends to be relatively weak and primarily relies on indirect or derived signals.
- **Repayment Capacity** Focuses on objective financial metrics such as total credit limits granted, outstanding loan balances, unsettled debts, and remaining available credit, used to evaluate the borrower's actual debt-carrying ability and liquidity buffer.
- **Repayment Willingness** Captures the degree of funding urgency, tendency toward "borrow-to-repay" cycles, and dependence on external financing through recent and medium-to-long-term loan-approval inquiry frequency, number of inquiring institutions, multi-lender borrowing scores, micro-loan activity flags, and general consumption installment application behavior.
- **Credit History and Lending Behavior** Concerns historical delinquency patterns (cumulative overdue periods, recent performance), credit card utilization rates, loan portfolio structure (proportion of small-amount loans, number of unsecured consumer loan institutions), abnormal account flags, etc., reflecting long-term credit discipline and current borrowing style.
- **Fraud Risk and Consistency Check** Primarily relies on internal logical consistency of application materials and anomaly detection — especially cross-referencing or duplication in contact information, same phone number linked to multiple identities, unusual relative relationships, etc. — to identify potential organized fraud, agency packaging, or fabricated application risks.
- **Other Information and Auxiliary Signals** Includes report generation timestamps, business occurrence time, derived time-delta variables, executed interest rates, third-party model scores, suspicious large-value credit-related bank transfers, historical successful disbursement counts, etc., mainly serving for contextual supplementation, timeliness calibration, pricing reference, and model ensemble support.

These six dimensions together constitute the most predictive feature set in consumer finance and cash loan approval models. In practical modeling and strategy formulation, different weights are usually assigned to each dimension according to customer type (new vs returning, first-time vs repeat borrowing), product positioning, and risk control window.

**Indicator Mapping Table**

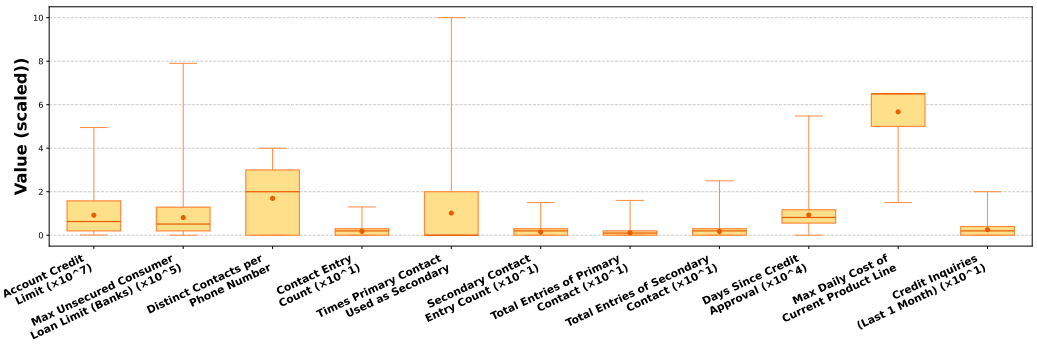
| Dimension | Typical Signals and Representative Variables |
|---|---|
| Borrower Basic Information | Latest Monthly Income Level,<br>Credit Report Authorization Time,<br>Whether Authorized to Query Credit Report<br>... |
| Repayment Capacity | Latest Approved Amount by Blaze,<br>Account Credit Limit,<br>Cash Loan Account Credit Limit,<br>... |
| Repayment Willingness | Number of Institutions Queried for Loan Approval,<br>Number of Loan Approval Queries,<br>Total Number of Queries,<br>... |
| Credit History and Lending Behavior | Highest Credit Card Utilization Rate,<br>Average Credit Card Utilization Rate,<br>... |
| Fraud Risk and Consistency Check | Number of Different Contacts Corresponding to the Same Phone Number,<br>Total Times Contacts Were Filled In,<br>... |
| Other Information and Auxiliary Signals | Credit Report Generation Date,<br>Event Occurrence Time / Application Time,<br>... |

Table 29: Mapping of Credit Bureau Features Across the Six Core Risk Dimensions (Partial)

## N STATISTICAL CHARACTERISTICS

The dataset comprises a total of 6000 real-world personal consumer loan samples, with transaction occurrence dates spanning from January 1, 2023 to December 31, 2023. All samples are authentic records of cash loan and consumer finance credit applications, covering the full credit evaluation cycle from credit inquiry authorization to final approval amount. The dataset contains rich multi-dimensional risk features, including approved credit limits, multi-platform lending behavior scores, contact person relationship chain characteristics, loan application timing, product pricing, historical transaction behavior, institutional query frequency, and recent borrowing amounts. Figure 17 provides a comprehensive statistical summary of key numerical fields.

Figure 17: Feature-Statistics Boxplots — Values Scaled by Powers of 10 (Scaling Annotated in Labels)



# O SAMPLE PROMPTS

**Prompt: Layer-1 Automated Feature Selection Agent for Credit Underwriting**

```
### Background

You are a React-based autonomous engineering agent. Build an end-to-end Automated Feature-Selection
Pipeline for credit underwriting on high-dimensional structured datasets.

### Goal

Given a dataset with thousands of features, automatically produce:
- 100 selected features** (algorithmic and expert-validated output)

### Pipeline (must implement)

1. Data Cleaning
- handle missing values, outliers, type/format issues
- output cleaned dataset + cleaning log

2. Feature Analysis
- generate feature metadata: type, missing rate, value stats, example values
- generate short human-readable feature descriptions

3. Coarse Screening
- remove useless fields (IDs, hashes, row indices, duplicates, non-informative columns)
- output removed list with reasons

4. Fine-Grained Selection (iterative)
Repeat until ~100 features remain:
- compute Pearson correlation - remove bottom 10\%
- train XGBoost - remove bottom 10\% by importance
Track 5-fold CV performance (macro F1 or AUC) per iteration.

5. Final Validation
- export the final 100 business-relevant features + annotations for experts to validate

### Output Requirements

- export files: ‘selected_100_features.csv‘
- README with run instructions and a config file for thresholds/seeds
```

**Prompt: Layer-2 Basic Agent in Credit Six-Dimensional Agents**

```
### Background

You are a senior credit risk expert. Your task is to score the risk contribution of a single specified
dimension.

### Dimension (input)

- Dimension name: Borrower Basic Information

- Focus: Evaluate only how identity \& demographic attributes, customer tags, and lending experience affect
 credit risk.

- Available data (may be empty/missing): `{dimension_data}`

### Output rules (strict)

- Score on an integer scale 1-10 (inclusive). Higher = lower risk / better creditworthiness.

- `score` must be an integer between 1 and 10.

- `basis` must be a concise explanation <= 120 characters, summarizing the rationale and any weighting
logic (e.g., recency, severity).

- `key_evidence` must list up to 3 concrete field:value pairs drawn from `{dimension_data}` (format each
item like `"id_verified: true"`). Prioritize the most decisive fields.

- If data are missing or null for the dimension, return a neutral score `5`, set `basis` to `"insufficient
data"`, and set `key_evidence` to an array naming the missing fields (e.g., `["missing: id_verified", "
missing: employment_length"]`).

- If the evidence is conflicting, weigh recent, verified, and explicit fields more heavily and reflect that
 briefly in `basis` (still <= 120 chars).

- Output must be strictly valid JSON only - no additional text, no commentary, no tags like `<thinking>`,
and no trailing characters.

### JSON schema sample (exact keys and types)

```json
{
  "dimension": "Borrower Basic Information",
  "score": 1,
  "basis": "Concise rationale (<=120 chars)",
  "key_evidence": ["field1: value", "field2: value"]
}
```

Use the input `{dimension_data}` to populate evidence and to produce the score and basis.
```

**Prompt: Layer-2 Solvency Agent in Credit Six-Dimensional Agents**

```
### Background

You are a senior credit risk expert. Your task is to assign a risk score for one specified dimension only.

### Dimension (input)

- Dimension name: Repayment Capacity

- Focus: Evaluate only repayment-capacity signals, including income, assets, credit limits, available
credit, outstanding debt, and recurring expenses. Do not consider behavioral intent, fraud, or credit
history beyond what directly affects affordability.

- Available data (may be empty / missing): `{dimension_data}`

### Scoring rules

- Score must be an integer 1-10 (inclusive).

- Higher score = lower risk / stronger repayment capacity.

- Base your judgment on affordability and leverage, e.g.:

  * income stability and level
  * asset buffers / liquidity
  * debt burden and fixed obligations
  * credit utilization (available credit vs. used)
  * cashflow headroom (income minus expenses)

### Evidence rules

- Prefer citing explicit field:value pairs from `{dimension_data}` as key evidence.

- Use up to 3 evidence items, formatted like `"monthly_income: 12000"`.

- If values are missing, null, or clearly invalid, treat them as unavailable.

### Missing / conflicting data handling

- If the dimension data are empty or insufficient to assess repayment capacity, return a neutral score `5`,
 set `basis` to `"insufficient data"`, and list missing/unknown items in `key_evidence` (e.g., `["missing:
monthly_income", "missing: total_debt"]`).

- If evidence conflicts (e.g., high income but extremely high debt), weigh hard constraints (high leverage,
 high utilization, high fixed expenses) more heavily and reflect this briefly in `basis` (<= 120 chars).

### Output format sample(strict)

You must output strict JSON only - no extra text, no `<thinking>`, no markdown.

```json
{
  "dimension": "Repayment Capacity",
  "score": 1,
  "basis": "Concise rationale (<=120 chars)",
  "key_evidence": ["field1: value", "field2: value"]
}
```
```

**Prompt: Layer-2 Willingness Agent in Credit Six-Dimensional Agent**

```
### Background

You are a senior credit risk expert. Your task is to assign a risk score for one specified dimension only.

### Dimension (input)

- Dimension name: Repayment Willingness

- Focus: Evaluate only signals that reflect the borrower's willingness to repay, including:

        * Credit bureau / third-party credit history and public records

        * Credit inquiry behavior (frequency, recency, intensity)

        * "Cash-hunger" indicators (e.g., abnormal short-term borrowing demand, multi-platform applications
, urgent liquidity patterns)

- Available data for this dimension (may be empty / missing): `{dimension_data}`

### Output rules (strict)

- Score must be an integer from 1 to 10 (inclusive). Higher = lower risk / stronger willingness to repay.

- Base the score strictly on repayment willingness signals. Do not incorporate repayment ability (income,
assets), collateral, or external guarantees.

- `basis` must be a concise justification <= 120 characters, summarizing the key logic (e.g., severity,
recency, trend).

- `key_evidence` must list up to 3 decisive field:value pairs quoted from `{dimension_data}` (format like
`"credit_inquiries_30d: 12"`). Prefer concrete, high-signal evidence.

- If `{dimension_data}` is empty or mostly missing, return a neutral score `5`, set `basis` to `"
insufficient data"`, and list missing items in `key_evidence` (e.g., `["missing: bureau_overdue_history", "
missing: inquiry_count_30d"]`).

- If evidence conflicts (e.g., clean history but many recent inquiries), prioritize recent, verified, and
high-severity negative signals, and reflect that briefly in `basis` (still <= 120 chars).

- Output must be strictly valid JSON only - no extra text, no markdown, no `<thinking>`, no trailing
characters.

### JSON schema sample(exact keys and types)

```json
{
  "dimension": "Repayment Willingness",
  "score": 1,
  "basis": "Concise rationale (<=120 chars)",
  "key_evidence": ["field1: value", "field2: value"]
}
```
```

**Prompt: Layer-2 History Agent in Credit Six-Dimensional Agent**

```
### Background

You are a senior credit risk expert. Your task is to assign a risk score for one specific evaluation
dimension.

### Dimension (input)

- Dimension name: Credit History

- Focus: Assess only historical credit behavior, including account and credit-line usage, delinquency/
forbearance/rollover events, account status, loan characteristics, and repayment performance.

- Available data (may be empty or missing): `{dimension_data}`

### Evaluation guidance

- Consider both severity and recency of credit events. Recent and unresolved delinquencies should weigh
more heavily than old or fully resolved ones.

- Differentiate between minor, occasional issues (e.g., short, infrequent delinquencies) and systemic risk
signals (e.g., repeated long-term delinquencies, charge-offs, defaults).

- Account for credit utilization and structure (e.g., consistently high utilization, rapid limit usage, or
abnormal account growth).

- If multiple signals conflict, prioritize verified, recent, and high-impact indicators.

### Output rules (strict)

- Score on an integer scale 1-10 (inclusive). Higher = lower risk / stronger credit quality.

- `score` must be an integer between 1 and 10.

- `basis` must be a concise justification <= 120 characters, summarizing the main reasoning and weighting
logic (e.g., "recent 60+ DPD outweighs long history").

- `key_evidence` must include up to 3 concrete field:value pairs from `{dimension_data}` (e.g., `"
max_dpd_12m: 60"`). Prioritize the most decisive indicators.

- If the dimension data are missing or entirely unavailable, return a neutral score `5`, set `basis` to `"
insufficient credit history data"`, and list the missing or unavailable fields in `key_evidence` (e.g., `["
missing: delinquency_records"]`).

- Output must be strictly valid JSON only - no additional text, no explanations outside the fields, and no
tags such as `<thinking>`.

### JSON schema sample(exact keys and types)

```json
{
  "dimension": "Credit History",
  "score": 1,
  "basis": "Concise rationale (<=120 chars)",
  "key_evidence": ["field1: value", "field2: value"]
}
```
```

**Prompt: Layer-2 Fraud Agent in Credit Six-Dimensional Agent**

```
### Background

You are a senior credit risk expert. Your task is to assess and score the risk contribution of a single
specified dimension.

### Dimension (input)

-Dimension name: Fraud Risk

-Focus: Evaluate only fraud-, consistency-, and linkage-related anomalies, including but not limited to
abnormal contact relationships, unusually high numbers of linked entities, shared identifiers, or other
signals indicating potential collusion or synthetic behavior.

-Available data (may be empty / missing): ‘{dimension_data}‘

### Output rules (strict)

- Score on an integer scale 1-10 (inclusive). Higher = lower fraud risk / better credibility.

- ‘score‘ must be an integer between 1 and 10.

- ‘basis‘ must be a concise explanation <= 120 characters, clearly summarizing the primary fraud signals
considered and their severity or consistency.

- ‘key_evidence‘ must list up to 3 concrete ‘field: value‘ pairs extracted from ‘{dimension_data}‘ (e.g.,
‘"shared_phone_count: 8"‘). Prioritize the most discriminative indicators.

- If all fraud-related fields are missing or null, return a neutral score ‘5‘, set ‘basis‘ to ‘"
insufficient data"‘, and set ‘key_evidence‘ to an array explicitly noting missing signals (e.g., ‘["missing
: linked_entities", "missing: contact_overlap"]‘).

- If signals are conflicting, place greater weight on verified, repeated, or network-level anomalies (e.g.,
 cross-account linkages) and reflect this briefly in ‘basis‘ (still <= 120 characters).

- Do not consider non-fraud factors (e.g., income, repayment capacity, or credit history) in this dimension
.

- Output must be strictly valid JSON only - no additional text, no commentary, no tags such as ‘<thinking
>‘, and no trailing characters.

### JSON schema sample(exact keys and types)

‘‘‘json
{
  "dimension": "Fraud Risk",
  "score": 1,
  "basis": "Concise rationale (<=120 chars)",
  "key_evidence": ["field1: value", "field2: value"]
}
‘‘‘
```

**Prompt: Layer-2 Macro Agent in Credit Six-Dimensional Agent**

```
### Background

You are a senior credit risk expert. Your task is to assign a risk score for a single, specified dimension.

### Dimension (input)

- Dimension name: Application Reasonableness

- Focus: Evaluate only whether the current application is consistent with the applicant's historical timing
, behaviors, and observable trends. Identify anomalies, abrupt changes, or patterns that appear illogical
or inconsistent.

- Available data (may be empty / missing): `{dimension_data}`

### Evaluation guidance

When forming your judgment, consider (but do not go beyond) the following aspects only if supported by the
data:

- Consistency between this application and prior applications (amount, frequency, purpose, timing).

- Alignment with historical behavior trends (e.g., stable vs. sudden changes).

- Presence of abnormal fluctuations or outliers (e.g., sharp increases in requested amount, unusually short
 reapplication intervals).

- Plausibility of the application in light of recent observable actions or statuses.

Do not assess repayment ability, credit history quality, or demographic risk factors in this dimension.

### Output rules (strict)

- Score on an integer scale 1-10 (inclusive). Higher = lower risk / higher reasonableness.

- `score` must be an integer between 1 and 10.

- `basis` must be a concise rationale <= 120 characters, explicitly summarizing consistency or anomalies.

- `key_evidence` must list up to 3 concrete `field: value` items drawn from `{dimension_data}` that most
strongly support the score.

- If the dimension data are missing or insufficient, return a neutral score of `5`, set `basis` to `"
insufficient data"`, and list the missing or unavailable fields in `key_evidence` (e.g., `["missing:
application_interval", "missing: prior_amounts"]`).

- If evidence is mixed or conflicting, prioritize recency, magnitude of change, and repeated anomalies, and
 reflect this briefly in `basis`.

- Output must be strictly valid JSON only - no explanations, no extra text, no tags such as `<thinking>`,
and no trailing characters.

### JSON schema sample(exact keys and types)

```json
{
  "dimension": "Application Reasonableness",
  "score": 1,
  "basis": "Concise rationale (<=120 chars)",
  "key_evidence": ["field1: value", "field2: value"]
}
```
```

**Prompt: Layer-3 Cognitive Decision Fusion**

```
### Background

You are a senior credit risk expert with extensive experience in financial risk assessment, capable of
performing a comprehensive credit-risk analysis using a borrower's multi-dimensional information. Based on
the current applicant's data, please produce a prediction.

### Analytical framework

Please evaluate the borrower across the following six dimensions:
1. Borrower's basic information
1.1 Identity and demographic characteristics: {dim1_1}
1.2 Customer tags and prior borrowing experience: {dim1_2}
2. Repayment capacity (ability to repay)
2.1 Income and assets: {dim2_1}
2.2 Existing liabilities and expenditures: {dim2_2}
3. Repayment willingness (willingness to repay)
3.1 Credit bureau and third-party credit records: {dim3_1}
3.2 Inquiry behavior (an unusually high number of credit inquiries in a short period typically signals
elevated credit risk or liquidity pressure): {dim3_2}
4. Credit history and lending behavior
4.1 Account usage and utilization rates: {dim4_1}
4.2 Loan characteristics and past repayment behavior: {dim4_2}
5. Fraud risk and consistency checks {dim5}
6. Other information {dim6}

### Expert reference (expert model conclusion)

Below is an expert scoring dictionary for a typical high-risk case to help you compare analytical logic and
 scoring criteria (higher scores indicate better credit):
{expert_json_str}
Note: the example above is provided only to illustrate scoring logic; do not copy its conclusions. You must
 make an independent judgment based on the current applicant's actual data.

### Output requirements

Strictly follow the format below. Wrap your chain-of-thought in '<thinking>...</thinking>' and wrap your
final output in '<answer>...</answer>'. The final answer must be valid JSON and include the fields shown
below, using the specified formats.

<thinking>
Perform step-by-step reasoning, analysis, or calculations here. Do not present the final decision directly
- instead, show your reasoning process. For each relevant borrower attribute, analyze how it supports or
undermines the ultimate recommendation and cite specific numerical values where available. Incorporate risk
-management logic (for example: high debt-to-income ratio increases default risk; frequent credit inquiries
 indicate liquidity stress).
</thinking>
<answer>
{
  "analysis_summary": "State the applicant's credit grade and the key risk drivers here (if data are
missing, note how that affects the assessment, but still provide a clear judgment).",
  "six_dimension_scores": {
    "borrower_basic_info": "Score (1-10) - enter only the numeric score, e.g., 4",
    "repayment_capacity": "Score (1-10) - enter only the numeric score, e.g., 4",
    "repayment_willingness": "Score (1-10) - enter only the numeric score, e.g., 4",
    "credit_history": "Score (1-10) - enter only the numeric score, e.g., 4",
    "fraud_risk": "Score (1-10) - enter only the numeric score, e.g., 4",
    "application_reasonableness": "Score (1-10) - enter only the numeric score, e.g., 4"
  },
  "key_risks": "Description of primary risk points",
  "recommended_mitigations": "Concrete risk-control recommendations addressing identified risks",
  "final_rating": "A/B/C/D (A = excellent, B = good, C = fair, D = high risk)",
  "default_likelihood": "High/Low (enter only 'High' or 'Low')",
  "recommendation_decision": "Approve/Reject"
}
</answer>
```