



Verifiability-Weighted Autonomy: A Conceptual Framework for Governing Agentic Artificial Intelligence in Business-to-Business Credit Risk Management

Shivanand R. Koppalkar

Researcher, Walsh College, United States

Corresponding author: skoppalk@mail.walshcollege.edu, us@innovate-software.com

• ORCID: <https://orcid.org/0009-0009-7843-2595>

Abstract

Firms that sell to other firms on credit now face a new design choice. Software agents built on large language models can gather evidence, draft credit memos, release held orders and chase overdue invoices. Yet there is little guidance on how much of this work an agent should do without a person. Most research on artificial intelligence in credit asks how accurate a model is, and most of it studies consumer lending. It says little about the decision rights of autonomous agents in business-to-business trade credit, where much activity sits outside consumer-protection and bank model-risk rules. This conceptual paper addresses that gap. It integrates research on trade credit, credit scoring, language-model agents, levels of automation, agency theory and model-risk governance. From this synthesis it builds the Verifiability-Weighted Autonomy framework. The central claim is simple: the autonomy granted to an agent should rise with the verifiability of the evidence behind a decision and fall with the decision's financial exposure and irreversibility. The paper defines four autonomy tiers and a transparent allocation rule. It maps twelve common trade-credit decisions onto the tiers. It proposes a Propose-Verify-Dispose architecture in which language-model agents propose, deterministic checks verify, and policy rules or people dispose. It offers seven testable propositions and four measures for future empirical work. For managers, the framework turns a vague question, "can we trust the agent?", into a concrete design rule that fits existing delegation-of-authority practice.

Keywords: *large language model agents; trade credit; delegation of authority; model risk management; human oversight*

1. Introduction

Trade credit is one of the oldest forms of finance. A supplier ships goods today and lets the buyer pay later. This simple act makes the supplier a lender. Research has long shown that trade credit is a major source of short-term finance for firms, and that suppliers often hold an information advantage over banks because they see their customers' ordering and payment behaviour first-hand [1]. Every firm that sells on account therefore runs a credit function, whether it calls it that or not. That function vets new customers, sets credit limits, decides whether to release orders that breach a limit, watches for warning signs and collects overdue money.

For decades, firms have automated parts of this work. Early models scored the risk of business failure from financial ratios [2]. Later, machine learning methods improved the accuracy of credit scoring across many



datasets [3], and new data sources such as digital footprints added predictive power [4]. These tools produce scores. People and fixed rules still make most of the decisions.

Large language models change this picture. When a language model is given tools, memory and a goal, it becomes an agent. It can read a customer's filed accounts, search the receivables ledger, write a credit memo and send an email, all in one loop [5], [6]. Several agents can also work together as a team of specialists [7]. Surveys of language models in finance show fast growth in such systems across research, trading and risk management [8], [9]. The question for credit managers is no longer only "how good is the score?" It is now "which decisions may the agent take by itself?"

The existing literature does not answer this question well. Studies of language models in credit mostly measure prediction accuracy. Some find that text processed by language models adds useful risk signals [10], [11]. Reviews find steady progress but also weak explainability and limited attention to regulation [12], [13]. Language models also produce fluent but false statements, often called hallucinations [14]. None of this tells a firm how to divide decision rights between an agent and its credit staff.

The gap is sharper in business-to-business settings. The European Union's Artificial Intelligence Act lists credit scoring as a high-risk use, but the listing refers to the creditworthiness of natural persons [15]. Credit extended by one company to another therefore falls largely outside that high-risk category. Bank model-risk guidance is strong [16], [17], but most trade credit is granted by manufacturers, distributors and service firms, not by banks. Many of these firms have no binding rule that tells them how to govern an autonomous credit agent. They need a framework they can adopt voluntarily.

This paper builds such a framework. It asks three research questions:

- RQ1: What property of a credit decision should govern how much autonomy an artificial intelligence agent receives?
- RQ2: What system architecture allows a firm to grant that autonomy safely?
- RQ3: How can the resulting framework be tested empirically?

The paper makes four contributions. First, it identifies evidence verifiability as the missing variable in decisions about agent autonomy. Accuracy tells us how often a model is right on average. Verifiability tells us whether we can check a specific output before acting on it. Second, it combines verifiability with decision exposure in a transparent allocation rule that yields four autonomy tiers. Third, it proposes a Propose–Verify–Dispose architecture that separates the power to suggest an action from the power to execute it. Fourth, it states seven testable propositions and four measures, so that the framework can be confirmed or rejected by future studies.

The rest of the paper is organised as follows. Section 2 reviews the relevant literature by theme. Section 3 explains the conceptual method. Section 4 presents the framework. Section 5 discusses its implications and limits. Section 6 concludes.

2. Literature Review

2.1. Trade credit and the business-to-business credit function

Trade credit differs from bank lending in three useful ways. First, the lender is also the seller, so the credit decision is tied to a sales decision. Second, the seller sees a stream of operational data, such as order size,

order timing and payment delays, that a bank does not see [1]. Third, the amounts are often small and frequent. A large supplier may make thousands of credit decisions each week, most of them routine.

In practice, credit departments control this volume through a delegation-of-authority matrix. The matrix states which role may approve which decision up to which amount. A credit analyst may set limits up to one level, a credit manager up to a higher level, and a committee above that. This structure matters for the present paper. It shows that firms already allocate decision rights by exposure. What they lack is a way to extend this matrix to artificial agents.

2.2. From scorecards to machine learning to language models

The first wave of automation used statistical scorecards. Discriminant analysis of financial ratios gave a simple way to rank firms by failure risk [2]. The second wave used machine learning. A large benchmark study found that ensemble methods often beat logistic regression in credit scoring, though the gains vary by dataset [3]. Alternative data also proved useful. Simple digital footprints, for example, carried information comparable to traditional credit bureau scores in one consumer setting [4].

Better accuracy came with a cost in transparency. Researchers responded with explanation methods that attribute a model's output to its inputs [18]. Others argued that for high-stakes decisions, firms should prefer models that are interpretable by design rather than explain black boxes after the fact [19].

The third wave uses large language models. Early studies applied them in three ways. Some used them as classifiers. One study applied few-shot classification with a generative model to peer-to-peer lending and compared it with logistic regression [20]. Some used them as feature builders. A study of peer-to-peer loan descriptions built a risk indicator from borrower text and combined it with traditional features [10]. Another fused a generative model with an encoder model to extract risk signals from financial text [11]. A 2026 systematic review of transformer-based methods in credit risk reported strong predictive results but noted that few studies included formal explainability or fairness checks [12]. A second review proposed a taxonomy of interpretability methods for language models used in credit risk assessment [13].

Two points stand out. First, almost all of this work treats the model as a predictor, not as an actor. Second, almost all of it studies consumer or peer-to-peer lending. Business-to-business trade credit, with its high volume of small operational decisions, is under-studied.

2.3. Language-model agents

An agent is a language model placed inside a loop of reasoning and action. A widely cited survey describes the core parts as a profile, memory, planning and action modules [5]. The ReAct pattern showed that interleaving reasoning steps with tool calls can improve task performance and make the trajectory easier to inspect [6]. Retrieval-augmented generation grounds the model's answers in documents fetched at run time, which can reduce some factual errors [21]. Multi-agent systems split work among specialists that plan, debate or check one another [7].

In finance, surveys report rapid adoption of these designs for research, trading, data analysis and risk management [8], [9]. They also note persistent problems. Agents can hallucinate facts, misread numbers and act on stale data [14]. A fluent but wrong credit memo is more dangerous than an obviously broken one, because it invites trust.

2.4. Levels of automation and human oversight

Human factors research offers a mature vocabulary for dividing work between people and machines. A classic model separates four functions, namely information acquisition, information analysis, decision selection and action implementation, and assigns each a level of automation from fully manual to fully automatic [22]. The model's key insight is that automation level should depend on the costs of errors and on the reliability of the automation.

Two human biases complicate this picture. Automation bias and complacency lead people to accept machine output without enough checking, especially under workload [23]. Algorithm aversion leads people to abandon an algorithm after seeing it make a single error, even when it outperforms them [24]. Notably, people are more willing to use an imperfect algorithm when they can adjust it, even slightly [25]. Human-centred artificial intelligence research argues that high automation and high human control can coexist when systems are designed for both [26]. Management research adds that automation and augmentation are not a simple either-or choice; firms must manage the tension between them over time [27].

2.5. Agency theory and delegated decision rights

Agency theory studies what happens when a principal delegates work to an agent whose actions are hard to observe [28]. One review of the theory argues that when an information system makes the agent's behaviour observable, the principal can rely on behaviour-based control rather than outcome-based control [29]. This argument transfers well to artificial agents. A language-model agent has no self-interest in the human sense. But it does have hidden errors. If the firm can observe and check the evidence behind each action, it can safely delegate more. If it cannot, it must keep more control or wait for outcomes, which in credit often arrive months later as bad debts.

2.6. Governance and regulation

Bank supervisors have long required firms to manage model risk. United States guidance defines model risk as the potential for loss from incorrect or misused models and calls for effective challenge, validation and governance [16]. The United Kingdom's Prudential Regulation Authority set out five principles for banks: model identification and risk classification, governance, development and use, independent validation, and risk mitigants [17]. The United States National Institute of Standards and Technology framework organises artificial intelligence risk work into four functions: govern, map, measure and manage [30]. The European Union's Artificial Intelligence Act places systems that evaluate the creditworthiness of natural persons in its high-risk category [15].

These instruments share a gap for the present topic. Bank guidance binds banks, not the manufacturers and distributors that grant most trade credit. The Act's credit listing targets natural persons, not companies. Both were also written mainly with predictive models in mind, not agents that take actions.

2.7. Synthesis and research gap

Table 1 summarises what each stream offers and what it misses for agentic business-to-business credit. Taken together, the streams supply the parts of an answer but not the answer. Credit research measures accuracy. Agent research builds capability. Human factors research says autonomy should depend on error costs and reliability. Agency theory says observability enables delegation. Governance research says models need tiered control. No existing work combines these into a rule for allocating decision rights to credit agents. That is the gap this paper fills.

Table 1. Literature streams relevant to agentic business-to-business credit: what each offers and what it misses.

Stream	Key sources	What it offers	What it misses for agentic business-to-business credit
Trade credit and credit scoring	[1], [2], [3]	Why suppliers lend; accurate risk scores	Treats models as scorers, not as actors
Language models in credit	[10], [11], [12], [13]	Text-based risk signals; interpretability taxonomies	Focuses on prediction, mainly in consumer and peer-to-peer lending
Language-model agents	[5], [6], [7], [9]	Designs for tool use, memory and teamwork	No rule for which actions an agent may take alone
Automation and human oversight	[22], [23], [25]	Levels of automation; known human biases	Judges reliability per system, not per decision
Agency theory	[28], [29]	Observable behaviour allows delegation	Not yet applied to artificial agents in credit
Governance and regulation	[16], [17], [30], [15]	Tiered model control; risk management functions	Binds banks or natural-person credit; written for predictive models

3. Methodology

This study is a conceptual paper. It does not collect or analyse new data. It follows established guidance for conceptual research, which identifies theory synthesis, theory adaptation, typology and model building as legitimate designs [31]. The present paper combines three of these. It synthesises five literature streams, it builds a typology of autonomy tiers, and it builds a model that links verifiability and exposure to autonomy.

The literature base was assembled through a purposive, integrative review. Integrative reviews synthesise mature and emerging literatures to produce new frameworks and perspectives [32]. They are suited to topics where research is spread across fields and methods, and where the aim is conceptual rather than a statistical summary [33]. The review proceeded in five steps.

- Step 1, scoping. The author defined five themes in advance: trade credit and credit scoring; language models in credit; language-model agents; automation and human oversight; and model-risk governance and regulation.
- Step 2, search and selection. Searches combined theme terms such as "trade credit", "credit scoring", "large language model", "agent", "levels of automation", "automation bias" and "model risk". Sources were included if they were peer-reviewed journal articles, peer-reviewed conference papers, or primary regulatory texts, and if they addressed at least one theme. Vendor material and non-reviewed web content were excluded. Foundational works were retained regardless of age because the framework builds on their constructs.
- Step 3, concept extraction. For each source, the author recorded constructs related to decision rights, error cost, evidence quality, human oversight and control.
- Step 4, framework building. The author iterated between the extracted constructs and practical knowledge of business-to-business credit operations. This abductive step is common in conceptual work. It is reported here so readers can judge its influence.

- Step 5, proposition derivation and checking. Each relationship in the framework was stated as a testable proposition. The framework was then checked for internal consistency, for clear boundary conditions, and against a worked illustration.

Because the study involved no human participants and no personal data, no ethics approval was required.

4. Results

This section presents the Verifiability-Weighted Autonomy framework. It defines the core constructs, states the allocation rule, introduces four measures, maps common decisions, describes the supporting architecture, illustrates its use and states the propositions.

4.1. Core constructs

The framework uses three constructs.

Evidence verifiability is the degree to which each material claim in an agent's output can be checked, by a deterministic procedure, against an authoritative system of record, in time to act. A material claim is any statement that would change the decision if it were false. Verifiability has three levels. High verifiability applies when claims point to system-of-record facts that can be re-queried, such as ledger balances, cash received, days past due, a filed set of accounts or an exact match on a sanctions list. Medium verifiability applies when claims rest on structured third-party data that is modelled or lagged, such as a bureau risk score or industry payment data. Low verifiability applies when claims rest on interpretation of unstructured content, such as news articles, call notes, emails or the agent's own inference about sentiment.

Verifiability is not the same as accuracy. A model can be accurate on average yet produce a specific claim that cannot be checked. A model can also be modest in accuracy yet produce claims that are easy to check. This distinction is the conceptual core of the framework. It follows directly from the agency argument that observable behaviour permits behaviour-based control [29].

Decision exposure is the potential harm if the decision is wrong. It combines three elements: the amount at risk relative to the firm's delegated limits, the irreversibility of the action, and the impact on the customer relationship or on legal standing. Releasing a small order that can be cancelled before shipment is low exposure. Writing off a debt or referring a customer to legal action is high exposure. Exposure also has three levels: low, medium and high.

Autonomy tier is the degree of independent action the firm grants to the agent for a given decision. Table 2 defines four tiers. They adapt the levels of automation model [22] to the specific workflow of a credit department.

Table 2. Autonomy tiers in the Verifiability-Weighted Autonomy framework.

Tier	Name	What the agent may do	Role of the person	Example
A0	Inform	Gather evidence and draft a summary; propose nothing binding	Decides from first principles, using the summary	Agent compiles news and filings for a credit review
A1	Recommend	Propose a specific action with reasons	Approves, edits or rejects before anything happens	Agent proposes a payment plan; officer approves it

Tier	Name	What the agent may do	Role of the person	Example
A2	Act with veto	Execute the action after a set veto window unless a person intervenes	Can reverse within the window; reviews exceptions	Agent sets a small initial limit, which takes effect after the window
A3	Act	Execute at once and log the decision	Audits a sample after the fact	Agent sends a standard overdue-payment reminder

4.2. The allocation rule

The framework's allocation rule sets the highest tier an agent may use. Let v be the verifiability level, scored 1 for low, 2 for medium and 3 for high. Let e be the exposure headroom, scored 1 for high exposure, 2 for medium and 3 for low. The maximum permitted tier is then given by Eq. (1).

$$A_{\max} = \max\{0, \min\{3, v + e - 3\}\} \quad (1)$$

The rule has three useful properties. First, it is monotonic. More verifiable evidence never reduces autonomy, and more exposure never increases it. Second, it is symmetric. Verifiability and exposure carry equal weight, so a firm can earn autonomy either by improving evidence or by limiting exposure, for example by capping order size. Third, it is conservative at the extremes. High-exposure decisions never reach full autonomy, and low-verifiability decisions never go beyond human approval. Fig. 1 shows the resulting matrix. Autonomy falls along the diagonal from the top-left cell to the bottom-right cell.

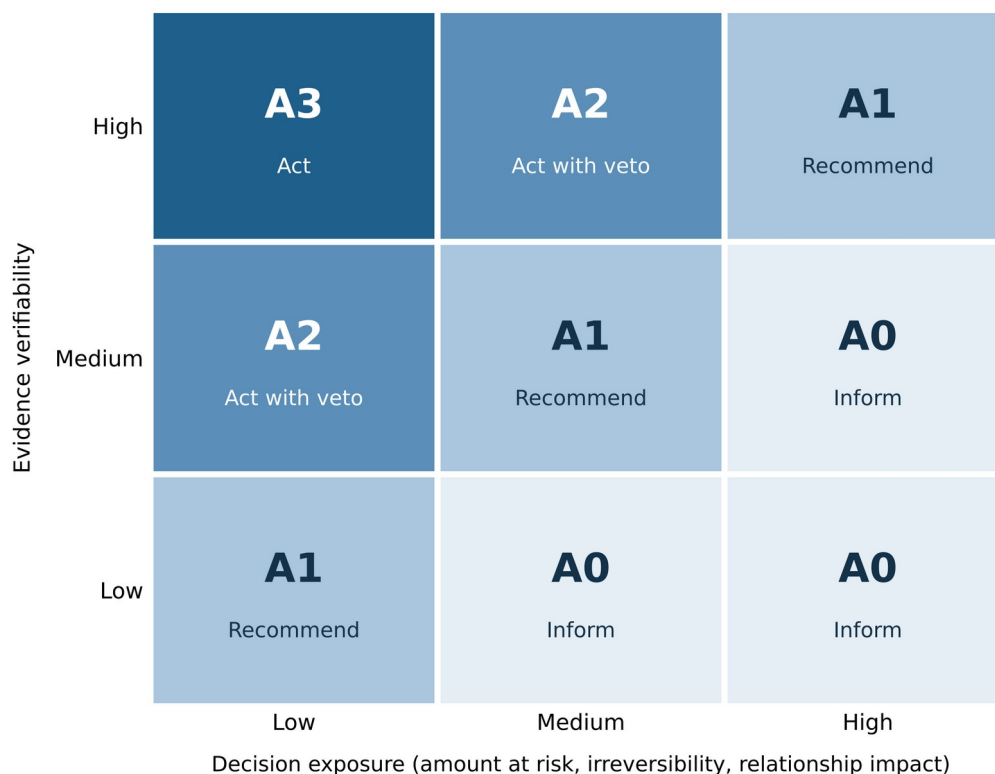


Fig. 1. Verifiability-Weighted Autonomy allocation matrix. Each cell shows the maximum autonomy tier given by Eq. (1). Autonomy falls as verifiability drops and as exposure rises.

A static rule can fail when conditions change. A sector downturn, a new fraud pattern or a change in the underlying model can raise error rates. The framework therefore adds a drift brake. Let R be the share of an

agent's decisions in a given decision type that were later reversed by a person or proved wrong, measured over a rolling window. Let θ be a tolerance set by the firm's risk appetite. The tier in force is then given by Eq. (2).

$$A_t = A_{\max} - 1 \text{ if } R > \theta; \quad A_t = A_{\max} \text{ otherwise} \quad (2)$$

The brake lowers autonomy automatically when reversals rise. Restoring the tier should require a human decision based on evidence, not an automatic reset. This asymmetry reflects the different costs of being too cautious and too permissive in credit.

4.3. Measures

Four measures make the framework operational and testable. The verification coverage ratio, defined in Eq. (3), is the share of material claims in an agent's output that the verifier could confirm against a system of record.

$$VCR = (\text{number of verified material claims}) / (\text{total number of material claims}) \quad (3)$$

The verification coverage ratio gives a run-time estimate of verifiability for each individual decision. A decision type can be rated high in verifiability, yet a particular case may lack data. In that case the ratio falls, and the policy engine can treat the case as one level lower.

The autonomy yield, defined in Eq. (4), is the share of decisions taken at a given tier that stood without later reversal.

$$AY = (\text{decisions at the tier not reversed}) / (\text{all decisions at the tier}) \quad (4)$$

The reversal rate R , used in Eq. (2), is one minus the autonomy yield for the decisions the agent took itself. Escalation precision is the share of escalations to a person in which the person changed the proposed action or added material information. Low escalation precision suggests that the firm is escalating too much and wasting staff time. Very high precision with a low escalation volume may suggest the opposite.

4.4. Mapping business-to-business credit decisions

Table 3 applies the rule to twelve common decisions across the trade-credit lifecycle. The ratings are illustrative. Each firm should calibrate them to its own limits, data and risk appetite. The table still shows a clear pattern. Routine, ledger-based actions, such as sending a standard payment reminder or releasing an order once cash has been applied, can run with full autonomy. Judgement-based actions, such as placing a hold because of a news report, stay with people even when the amount is moderate. High-exposure actions, such as large limit increases or write-offs, stay with people even when the evidence is strong.

Table 3. Illustrative mapping of twelve trade-credit decisions to autonomy tiers under Eq. (1).

No.	Decision	Main evidence	Verifiability	Exposure	Maximum tier
1	Match a new customer to its legal entity and bureau file	Company registry number; bureau identifier	High	Low	A3 Act
2	Send a standard overdue-payment reminder	Ledger shows invoice past due and no dispute flag	High	Low	A3 Act
3	Raise an early-warning flag from a change in payment pattern	Ledger payment history	High	Low	A3 Act

No.	Decision	Main evidence	Verifiability	Exposure	Maximum tier
4	Release a held order once cash is applied and the account is within limit	Bank feed; cash application; ledger	High	Low	A3 Act
5	Release an order within an approved temporary over-limit tolerance	Ledger; approved tolerance band	High	Medium	A2 Act with veto
6	Set an initial limit inside the small-account band	Bureau risk score; policy band	Medium	Low	A2 Act with veto
7	Summarise filed accounts into a draft credit memo	Filed financial statements	Medium	Low	A2 Act with veto
8	Clear a possible sanctions-list match as a false positive	List entry; customer identifiers	High	High	A1 Recommend
9	Classify a customer dispute or deduction from an email	Customer email text	Low	Low	A1 Recommend
10	Propose a payment plan for an overdue account	Ledger; customer correspondence	Medium	Medium	A1 Recommend
11	Place an account on hold because of an adverse news report	News article text	Low	Medium	A0 Inform
12	Raise a limit above the analyst band, or write off a debt	Bureau data; accounts; collection history	Medium	High	A0 Inform

Two cases deserve comment. Sanctions screening has highly verifiable evidence. However, wrongly clearing a true match carries legal exposure, so the rule caps it at recommendation level. Dispute classification from customer emails has low exposure, but its evidence is interpretive, so the rule again caps it at recommendation level. In both cases, the binding constraint differs. This is the practical value of using two dimensions rather than one.

4.5. The Propose-Verify-Dispose architecture

A rule is only useful if a system enforces it. The framework therefore specifies an architecture with one central principle: the language model may propose an action, but it never holds the authority to execute it. Execution authority sits in a deterministic policy engine that the firm controls. Fig. 2 shows the design.

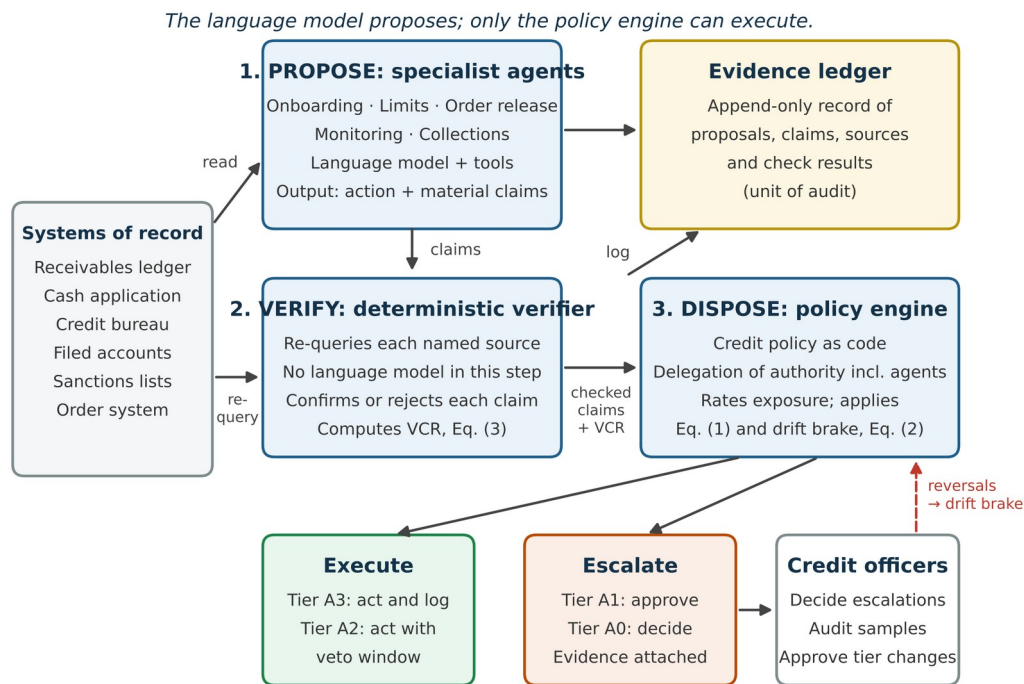


Fig. 2. Propose-Verify-Dispose architecture. Language-model agents propose actions and claims; a deterministic verifier checks each claim against systems of record; a policy engine disposes within delegated authority; reversals by credit officers feed the drift brake.

The architecture has five parts.

- **Specialist agents.** Separate agents handle onboarding, limit setting, order release, monitoring and collections. Each produces a proposed action together with a list of material claims. Each claim names its source, such as a ledger table and record, a bureau report or a filed document.
- **Evidence ledger.** Every proposal and claim is written to an append-only record with timestamps. The ledger is the unit of audit. It lets reviewers reconstruct why an action was proposed and what was checked.
- **Deterministic verifier.** The verifier re-queries each named source and confirms or rejects each claim. It does not use a language model for this step. It computes the verification coverage ratio for the proposal.
- **Policy engine.** The engine holds the firm's credit policy as code and an extended delegation-of-authority table that includes agents. It rates exposure, reads the coverage ratio, applies Eq. (1) and Eq. (2), and disposes. It either executes the action, executes it with a veto window, or routes it to a person with the evidence attached.
- **Human credit officers.** People handle escalations, review sampled autonomous decisions and approve changes to tiers. Their reversals feed the reversal rate and therefore the drift brake.

This separation has two benefits. First, it contains hallucination. A false claim fails verification and lowers the coverage ratio, which lowers the permitted tier, before any action is taken. Second, it keeps accountability clear. The firm can state, for every action, which rule permitted it and which evidence supported it. The design also reflects human-factors advice. Reviewers see which claims were verified and which were not, which is intended to reduce automation bias [23]. Credit managers can adjust tolerances and tiers, which is intended to reduce algorithm aversion [25].

4.6. Illustrative scenario

The following scenario is hypothetical. It illustrates the logic and does not report data.

A distributor's order is held because the customer's balance exceeds its credit limit. The order-release agent checks the bank feed, finds a payment received that morning, matches it to open invoices and proposes release. It makes three material claims: the payment was received, it was applied to the correct invoices, and the account is now within its limit. The verifier confirms all three against the receivables ledger. The coverage ratio is 1.0. The order value is well inside the low-exposure band, and the order can be cancelled before dispatch. Verifiability is high and exposure is low, so Eq. (1) permits tier 3. The policy engine releases the order and logs the decision.

The next day, the monitoring agent reads a news article that suggests the same customer may close a plant. It proposes placing the account on hold. Its key claim is an interpretation of news text. The verifier cannot confirm it against any system of record, so verifiability is low. A hold on an active customer has medium exposure because it can disrupt sales and the relationship. Eq. (1) yields tier 0. The policy engine routes the proposal to a credit officer, with the article and the agent's summary attached. The officer phones the customer, learns that the closure affects a single site, and decides to reduce the limit modestly rather than impose a hold.

The same customer, the same agents and the same model thus receive different autonomy for different decisions. The governing factor is not the model's average accuracy. It is whether the firm can check the specific evidence and how much harm a mistake would cause.

4.7. Propositions

Table 4 states seven propositions that follow from the framework, with suggested measures and designs for testing them. Propositions 1 to 3 concern the core logic of verifiability and the separation of proposal from execution. Propositions 4 and 5 concern human behaviour. Proposition 6 concerns dynamic control. Proposition 7 concerns governance alignment outside binding regulation.

Table 4. Testable propositions, measures and possible research designs.

No.	Proposition	Measure	Possible test design
P1	Within a decision type, the agent's reversal rate falls as the verification coverage ratio rises.	Reversal rate R; verification coverage ratio	Replay of historical receivables ledgers through simulated agents
P2	Allocating autonomy by the allocation rule yields lower loss-weighted error than allocating it by average model accuracy alone.	Loss-weighted error; autonomy yield	Counterfactual simulation comparing the two allocation policies
P3	Separating proposal from execution reduces the severity of errors caused by hallucination, without a proportional loss of throughput.	Error severity; decisions processed per hour	Controlled experiment with and without the verifier gate
P4	Showing reviewers which claims are verified and which are not reduces automation bias at tiers A1 and A2.	Rate of approving proposals that contain false claims	Randomised field experiment with credit staff

No.	Proposition	Measure	Possible test design
P5	Credit staff use and trust agents more when tiers come from an explicit rule they can adjust than from fixed vendor defaults.	Adoption rate; override behaviour; trust scales	Between-team field study
P6	A drift brake reduces tail losses during sudden changes in credit conditions, compared with static tiers.	Losses in the worst 5% of periods	Simulation over historical downturn periods
P7	Non-bank firms that adopt the framework show governance practices closer to recognised model-risk principles than firms that do not.	Coverage of recognised model-risk controls	Multi-firm survey and document analysis

5. Discussion

5.1. Theoretical implications

The framework makes three theoretical contributions. First, it adds evidence verifiability to the classic levels-of-automation model [22]. That model ties automation to reliability and error cost. Reliability, however, is a property of a system averaged over many cases. Verifiability is a property of a single case at the moment of decision. For generative agents, whose errors are fluent and case-specific [14], the case-level property matters more. The framework therefore extends the model from systems to decisions.

Second, it applies agency theory to artificial agents in a precise way. The classic argument holds that information systems which make behaviour observable allow behaviour-based control [29]. The evidence ledger and verifier are exactly such systems. The framework predicts that firms can delegate more to agents in proportion to how much of the agents' reasoning they can observe and check. This turns a general theoretical claim into a design rule.

Third, it offers a resolution to the automation–augmentation tension at the level of individual decisions [27]. A firm does not have to choose whether its credit function will be automated or augmented. Under the framework, the same function is both at once. Routine, verifiable, low-exposure decisions are automated. Interpretive or high-exposure decisions are augmented. The boundary moves as evidence improves or as reversal rates change.

5.2. Managerial implications

The framework gives credit leaders five practical steps. First, build a decision inventory. List every recurring credit decision and rate its verifiability and exposure. Table 3 offers a starting point. Second, extend the delegation-of-authority matrix to agents. Credit committees already approve human limits; they should approve agent tiers in the same way. Third, invest in verifiability. This is perhaps the most counter-intuitive implication. The quickest way to safely increase automation is often not a larger or newer model. It is better data plumbing: clean ledger links, reliable cash application, reliable entity matching and well-defined sources for each claim. Each improvement moves decisions up the verifiability axis and unlocks a higher tier. Fourth, keep exposure low where possible. Splitting large decisions into smaller, reversible steps, for example by using temporary over-limit tolerances rather than permanent limit increases, moves decisions along the exposure axis. Fifth, monitor the four measures on a regular dashboard and treat the drift brake as a standing control.

Table 5 shows how the framework's controls map onto the principles of bank model-risk management [17] and the functions of the artificial intelligence risk management framework [30]. Non-bank firms are not bound by the former. The mapping nevertheless lets them show auditors, lenders and credit insurers that their agent governance follows recognised practice.

Table 5. Mapping of framework controls to model-risk principles and artificial intelligence risk management functions.

Framework control	Model-risk principle [17]	Risk management function [30]
Decision inventory rated for verifiability and exposure	Principle 1: model identification and model risk classification	Map
Agent tiers approved within the delegation-of-authority matrix	Principle 2: governance	Govern
Evidence ledger and deterministic verifier	Principle 3: model development, implementation and use	Measure
Periodic independent review of coverage ratio, autonomy yield and sampled decisions	Principle 4: independent model validation	Measure
Veto windows, escalation routes and drift brake	Principle 5: model risk mitigants	Manage

5.3. Regulatory implications

The framework speaks to a regulatory blind spot. Business-to-business credit decisions by non-bank firms appear to sit largely outside both bank model-risk rules and the Artificial Intelligence Act's high-risk credit listing, which refers to natural persons [15]. That does not mean such decisions carry little risk. Errors can harm suppliers, customers and, through supply chains, wider markets. A voluntary, auditable framework can fill part of this gap.

Firms should also watch the edges of the business-to-business category. Sole traders are natural persons. Directors who sign personal guarantees are natural persons. An agent that assesses a guarantor's personal standing may therefore fall within the high-risk regime. The framework handles such cases by treating them as high exposure, which caps autonomy at recommendation level. Firms should still take legal advice on their specific situation, because the scope of the Act depends on facts and on guidance that continues to evolve.

5.4. Limitations

This study has five main limitations. First, it is conceptual. Its propositions have not yet been tested. Second, the three-level scales for verifiability and exposure are deliberately coarse. They make the rule easy to use but may hide important variation. Future work may develop finer scales. Third, verifiability is not truth. A system of record can contain errors, and fraudsters can forge documents that pass deterministic checks. Verification reduces one class of risk, namely agent error, but not every class. Fourth, the framework draws partly on the author's practical experience of business-to-business credit operations. This experience shaped the abductive step and may carry bias. Fifth, the capabilities of language-model agents change quickly. The specific decisions in Table 3 may move between tiers as tools improve, although the underlying rule should remain stable.

5.5. Directions for future research

The propositions suggest a clear research agenda. Researchers could replay historical receivables ledgers through simulated agents and compare outcomes under the allocation rule against outcomes under accuracy-based thresholds. They could run field experiments in which credit teams receive agent proposals with and without claim-level verification flags, and measure reviewer behaviour. Delphi studies with experienced credit managers could refine the exposure scale and test the face validity of Table 3. Multi-firm surveys could test whether firms with explicit agent tiers show stronger governance and fewer large losses. Finally, the framework could be adapted to other operational decisions where agents act on evidence, such as procurement approvals, insurance claims handling and payments.

6. Conclusion

Language-model agents are moving from producing scores to taking actions. For firms that extend trade credit, this shift raises a practical question that the accuracy-focused literature has not answered: which credit decisions may an agent take by itself? This paper answers that the autonomy granted to an agent should rise with the verifiability of the evidence behind each decision and fall with the decision's exposure. It turns that principle into four autonomy tiers, a transparent allocation rule with a drift brake, a Propose-Verify-Dispose architecture, four measures and seven testable propositions.

The framework's distinctive contribution is to move the debate from "how accurate is the model?" to "can we check this answer before we act on it?" That shift has a practical consequence. It makes data quality and verification, not model size, the main lever for safe automation. It also gives firms outside binding regulation a way to govern autonomous credit agents that is consistent with recognised model-risk principles. Future empirical work should test the propositions, refine the scales and extend the approach to other operational domains where artificial agents act on evidence.

Funding

This research received no external funding.

Conflict of Interest

The author is the founder of Innovate Software Consulting Inc Ltd, which provides enterprise analytics and technology consulting services. The firm did not fund this work, and the article does not promote any product or service. The author declares no other conflict of interest.

Data Availability Statement

No new data were created or analysed in this study. Data sharing is therefore not applicable. All sources used to build the framework are listed in the References.

References

- [1] M. A. Petersen and R. G. Rajan, "Trade credit: Theories and evidence," *Rev. Financ. Stud.*, vol. 10, no. 3, pp. 661–691, 1997, doi: 10.1093/rfs/10.3.661.
- [2] E. I. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," *J. Finance*, vol. 23, no. 4, pp. 589–609, Sep. 1968, doi: 10.1111/j.1540-6261.1968.tb00843.x.

- [3] S. Lessmann, B. Baesens, H.-V. Seow, and L. C. Thomas, "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research," *Eur. J. Oper. Res.*, vol. 247, no. 1, pp. 124–136, Nov. 2015, doi: 10.1016/j.ejor.2015.05.030.
- [4] T. Berg, V. Burg, A. Gombović, and M. Puri, "On the rise of FinTechs: Credit scoring using digital footprints," *Rev. Financ. Stud.*, vol. 33, no. 7, pp. 2845–2897, Jul. 2020, doi: 10.1093/rfs/hhz099.
- [5] L. Wang et al., "A survey on large language model based autonomous agents," *Front. Comput. Sci.*, vol. 18, no. 6, Art. no. 186345, 2024, doi: 10.1007/s11704-024-40231-1.
- [6] S. Yao et al., "ReAct: Synergizing reasoning and acting in language models," in *Proc. 11th Int. Conf. Learn. Represent. (ICLR)*, Kigali, Rwanda, 2023.
- [7] T. Guo et al., "Large language model based multi-agents: A survey of progress and challenges," in *Proc. 33rd Int. Joint Conf. Artif. Intell. (IJCAI)*, Jeju, South Korea, 2024, pp. 8048–8057, doi: 10.24963/ijcai.2024/890.
- [8] Y. Li, S. Wang, H. Ding, and H. Chen, "Large language models in finance: A survey," in *Proc. 4th ACM Int. Conf. AI Finance (ICAIF)*, New York, NY, USA, 2023, pp. 374–382, doi: 10.1145/3604237.3626869.
- [9] Y. Dong et al., "Large language model agents in finance: A survey bridging research, practice, and real-world deployment," in *Findings Assoc. Comput. Linguist.: EMNLP 2025*, Suzhou, China, 2025, pp. 17889–17907, doi: 10.18653/v1/2025.findings-emnlp.972.
- [10] M. Sanz-Guerrero and J. Arroyo, "Credit risk meets large language models: Building a risk indicator from loan descriptions in P2P lending," *Intel. Artif.*, vol. 28, no. 75, pp. 220–247, 2025.
- [11] H. Tan and Y. Xie, "Financial text analysis and credit risk assessment using a GPT-4 and improved BERT fusion model," *PLoS ONE*, vol. 20, no. 11, Art. no. e0336217, Nov. 2025, doi: 10.1371/journal.pone.0336217.
- [12] P. Raliphada, S. Olukanmi, and M. Olusanya, "Transformer-based NLP approaches for credit risk prediction: A systematic review," *Front. Artif. Intell.*, vol. 9, Art. no. 1819994, May 2026, doi: 10.3389/frai.2026.1819994.
- [13] M. Golec and M. Alabduljalil, "Interpretable LLMs for credit risk: A systematic review and taxonomy," *Expert Syst. Appl.*, vol. 306, Art. no. 130941, 2026, doi: 10.1016/j.eswa.2025.130941.
- [14] Z. Ji et al., "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, Art. no. 248, 2023, doi: 10.1145/3571730.
- [15] "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Off. J. Eur. Union*, L series, Jul. 12, 2024.
- [16] Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, "Supervisory guidance on model risk management," SR Letter 11-7, Washington, DC, USA, Apr. 2011.
- [17] Prudential Regulation Authority, "Model risk management principles for banks," Supervisory Statement SS1/23, Bank of England, London, U.K., May 2023. [Online]. Available: <https://www.bankofengland.co.uk/prudential-regulation/publication/2023/may/model-risk-management-principles-for-banks-ss>
- [18] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable machine learning in credit risk management," *Comput. Econ.*, vol. 57, no. 1, pp. 203–216, 2021, doi: 10.1007/s10614-020-10042-0.
- [19] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.
- [20] G. Babaei and P. Giudici, "GPT classifications, with application to credit lending," *Mach. Learn. Appl.*, vol. 16, Art. no. 100534, Jun. 2024, doi: 10.1016/j.mlwa.2024.100534.
- [21] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Adv. Neural Inf. Process. Syst. 33 (NeurIPS)*, 2020, pp. 9459–9474.
- [22] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 30, no. 3, pp. 286–297, May 2000, doi: 10.1109/3468.844354.
- [23] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Hum. Factors*, vol. 52, no. 3, pp. 381–410, Jun. 2010, doi: 10.1177/0018720810376055.
- [24] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err," *J. Exp. Psychol.: Gen.*, vol. 144, no. 1, pp. 114–126, 2015, doi: 10.1037/xge0000033.

- [25] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them," *Manage. Sci.*, vol. 64, no. 3, pp. 1155–1170, Mar. 2018, doi: 10.1287/mnsc.2016.2643.
- [26] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *Int. J. Hum.-Comput. Interact.*, vol. 36, no. 6, pp. 495–504, 2020, doi: 10.1080/10447318.2020.1741118.
- [27] S. Raisch and S. Krakowski, "Artificial intelligence and management: The automation-augmentation paradox," *Acad. Manage. Rev.*, vol. 46, no. 1, pp. 192–210, Jan. 2021, doi: 10.5465/amr.2018.0072.
- [28] M. C. Jensen and W. H. Meckling, "Theory of the firm: Managerial behavior, agency costs and ownership structure," *J. Financ. Econ.*, vol. 3, no. 4, pp. 305–360, Oct. 1976, doi: 10.1016/0304-405X(76)90026-X.
- [29] K. M. Eisenhardt, "Agency theory: An assessment and review," *Acad. Manage. Rev.*, vol. 14, no. 1, pp. 57–74, Jan. 1989, doi: 10.5465/amr.1989.4279003.
- [30] National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," NIST AI 100-1, Gaithersburg, MD, USA, Jan. 2023, doi: 10.6028/NIST.AI.100-1.
- [31] E. Jaakkola, "Designing conceptual articles: Four approaches," *AMS Rev.*, vol. 10, no. 1–2, pp. 18–26, 2020, doi: 10.1007/s13162-020-00161-0.
- [32] R. J. Torraco, "Writing integrative literature reviews: Guidelines and examples," *Hum. Resour. Dev. Rev.*, vol. 4, no. 3, pp. 356–367, Sep. 2005, doi: 10.1177/1534484305278283.
- [33] H. Snyder, "Literature review as a research methodology: An overview and guidelines," *J. Bus. Res.*, vol. 104, pp. 333–339, Nov. 2019, doi: 10.1016/j.jbusres.2019.07.039.