Available Online: 18 November 2026 DOI: 10.54254/3049-5768/2026.39203

Financial large language models in manufacturing supplier relationship management: applications, limits, and human-AI decision-making

Zunhe Xu

College of Humanities, Sichuan Agricultural University, Ya'an, China

3395754164@qq.com

Abstract. As Industry 4.0 is coming, Manufacturers need to evaluate suppliers with limited staff and incomplete information. Smaller suppliers may be particularly difficult to assess: their records are less standardized, and limited access to finance can obscure their capacity to fulfil orders. This paper examines whether Financial Large Language Models (FinLLMs), including FinGPT and BloombergGPT, can assist Supplier Relationship Management (SRM). It synthesizes research on quality-management workloads, supplier credit assessment, document processing, and supplier discovery, and considers the applications and limitations of the two models as representative cases. The review identifies four possible uses: extracting risk signals from text, combining textual and relational data for credit assessment, reading engineering documents, and finding suppliers through graph-based search. It also identifies limits in process-specific knowledge, numerical reasoning, confidence calibration, and the currency and coverage of training data. To clarify who should act on model outputs, the paper proposes a "boss-employee" framework organized around task division, decision authority, and calibration through human review. The framework treats AI as support for procurement staff rather than a substitute for their final judgement. It remains conceptual and requires testing in manufacturing settings.

Keywords: manufacturing, supplier relationship management, financial large language models, human-AI collaboration

1. Introduction

Manufacturing procurement teams must assess suppliers whose quality records, financial disclosures, and production capabilities are not equally visible. Supplier Relationship Management (SRM) covers qualification, performance evaluation, and risk monitoring across those relationships. Procurement accounts for more than 50% of production costs in many manufacturing firms and can reach 70% in capital-intensive sectors. Errors in supplier assessment therefore affect both production and cost. The problem is particularly acute when a buyer has only a few transactions with a supplier: there may be too little standardized information to support a routine comparison with established firms.

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Financial Large Language Models (FinLLMs), including FinGPT and BloombergGPT, have been considered for procurement and supply-chain finance because they can process large collections of text and documents. Much of the available research, however, addresses general financial tasks rather than decisions made in manufacturing SRM. It is less clear where these models can help a purchaser and where their outputs need independent checking. Consider a credit warning extracted from a report. Its use depends on whether the buyer can locate the passage and establish that it concerns the supplier under review, a test of traceability rather than fluent text generation. This paper examines those questions from the standpoint of enterprise management, procurement practice, and organizational decision-making; model training and parameter optimization are outside its scope.

The study combines a review of published research and industry reports with a comparison of the reported applications and limitations of FinGPT and BloombergGPT. It considers operational pressures in SRM, evaluates four possible applications and their constraints, and proposes a three-part "boss-employee" framework for human-AI decisions. The comparison concerns the uses and risks described in the available material, not a new head-to-head performance test. Available reports differ in task, data, and setting, so their figures cannot be read as a common ranking of the two models. For procurement teams, the distinction between assistance and decision authority matters when errors can disrupt delivery or quality. For subsequent research, the framework identifies organizational practices that could be tested rather than assuming that better model scores will improve procurement outcomes.

2. Reasons for manufacturing SRM using AI

2.1. Operational and information constraints

Manufacturing SRM faces four related constraints. Quality departments, for one, must complete documentation alongside supplier audits and corrective work. A 2026 Superkind report on German mid-sized manufacturers states that more than half of quality-management posts are difficult to fill and that preparing and checking one 8D report takes an engineer 8 to 20 hours. The estimate concerns reporting work, not the time required to resolve a defect. It nevertheless shows why assigning skilled engineers to repeated drafting can leave less capacity for site inspection or process correction.

Supplier documents pose a different problem. Krieger et al. describe a skewed distribution of invoice formats: a few major suppliers use relatively consistent layouts, while hundreds of less frequent suppliers may submit only one or two documents a year in differing formats [1, 2]. Rules designed around common templates consequently transfer poorly to unfamiliar invoices. A purchasing team still needs to match names, amounts, and terms across the documents, even where a template has never been seen before. Here the scarcity of examples for each format, rather than the total volume of invoices, explains the burden of manual checking.

Credit assessment adds a financing constraint. In a study of listed Chinese intelligent-manufacturing firms, Li et al. link limited transparency among smaller upstream suppliers to higher credit premiums and reliance on the credit standing of major downstream buyers [3]. Lenders may therefore lack the information needed to evaluate suppliers on their own records. Supplier-risk datasets also contain many financial, delivery, and quality indicators but relatively few serious defaults. A model may fit those rare cases too closely or miss them while achieving good overall accuracy [4]. For a manager, a high accuracy figure is of limited value if the system overlooks the supplier most likely to default. The financing and data problems meet at that point: limited disclosure makes the cases requiring scrutiny harder to recognize.

2.2. Where AI may fit

Automated visual inspection can take on repeated checks on standardized production lines; the evidence considered here reports defect-classification accuracy above that of an average human inspector and continuous operation [5]. That may release quality staff for investigation and supplier follow-up, though inspection is distinct from making a supplier decision. A defect flag can tell staff which component to examine; it cannot by itself establish why the defect occurred or whether the supplier should retain a contract. For documents, systems such as LayoutLM use text, page position, and visual layout together. They offer a way to handle unfamiliar invoice formats that fixed OCR-and-rule pipelines struggle to process [2]. A person can then check extracted fields against the invoice before they enter a payment or evaluation record.

Financial models address another part of the problem: Wu describes information and supervision as two channels through which they may support financing [6]. Extracting risk signals from annual reports, litigation notices, and other text can reduce the information gap between a small supplier and a lender [3, 4]. Such signals may prompt a closer look at a firm's obligations, but they do not measure repayment capacity on their own. Monitoring transaction records through a platform shared by firms and financing institutions may, in turn, reduce exclusive reliance on a core buyer's credit endorsement [6]. In this arrangement, transaction evidence helps connect a financing request to actual trade. Neither channel removes the need to check whether the underlying records are current and authentic.

Feature selection and methods that give greater attention to rare cases can help with high-dimensional, imbalanced credit datasets [4]. Their practical purpose is to improve detection of high-risk suppliers without relying solely on an overall accuracy score. A team evaluating such a system should therefore ask how often actual defaults are missed, rather than considering only the share of all cases classified correctly. Visual inspection, invoice processing, and credit assessment call for different evidence and different human checks.

3. Financial language models in manufacturing SRM

3.1. Potential contributions

The reviewed applications fall into four groups (Table 1). They serve different decisions, from collecting warning signals to identifying possible suppliers. The numerical results shown are taken from the studies and cases discussed in the draft; they are not results of new experiments in this study. Their denominators, datasets, and evaluation procedures must be checked in the underlying sources before any direct numerical comparison is made.

Table 1. Four potential contributions of FinLLMs to SRM [6, 7]

Application	SRM use	Reported example or result
Unstructured-data extraction	Supply-chain credit-risk warnings	99.5% accuracy in sampling financial indicators from PDFs
Text-and-graph credit assessment	Assessment of smaller suppliers	3-5 percentage-point improvement over a baseline on industrial credit data
Multimodal document analysis	Supplier qualification review	70-90% less time spent drafting 8D reports
Graph-assisted supplier discovery	Screening suppliers beyond established directories	Matching supplier capabilities to specific process requirements

Text-and-graph assessment is pertinent because manufacturing firms hold two kinds of evidence: audit reports and other text, alongside records of transactions and supplier relationships. A combined assessment can use information that either source alone would miss [4]. An audit note about late deliveries gains meaning when read alongside the supplier's position in a production network; a connection in the network, by contrast, does not explain the cause of a delay. Graph-assisted discovery addresses a different gap. Supplier capability may be described in exhibition material or industry-association listings rather than a searchable company database. Bringing such descriptions into a supplier graph may help a buyer search for a precise process requirement instead of relying only on keywords [7]. The gain here is potentially broader supplier coverage, not proof that a recommended supplier can meet a tolerance without technical verification.

The four rows should also be read with their boundaries in mind. A reported gain in extracting PDF figures concerns a document task; it does not establish an equivalent gain in credit decisions. Likewise, reduced time drafting an 8D report does not show that the resulting report is correct. The person approving the report still needs to confirm the defect description and the proposed corrective action. Supplier discovery raises a comparable distinction between finding a candidate and establishing its production capacity. A listing that identifies a firm able to discuss a rare process is a lead for verification, not evidence that the firm's equipment meets every requested specification. In each case, the relevant performance measure depends on what a procurement team will do with the output.

3.2. Limits in manufacturing use

FinLLMs are not designed specifically to judge production processes. Table 2 groups the limitations relevant to their use in procurement. The reported figures concern different tasks and should not be combined as though they came from a single benchmark.

Table 2. Five limitations relevant to manufacturing procurement [2, 8]

Limitation	Manufacturing implication	Reported observation
Process hallucination	A description of a non-standard process may conflict with shop-floor capability	27% hallucination rate in the cited manufacturing-capacity example
Complex numerical reasoning	Multi-step tolerance calculations may be unreliable	FinGPT trails GPT-4 by 15-22 percentage points on the reported numerical tasks
Overconfident errors	An incorrect price prediction or defect classification may appear reliable	21.6% of the reported incorrect price predictions were assigned high confidence
Outdated data	New environmental penalties or raw-material price shocks may be absent	Performance on events after the training cutoff declines
Limited coverage	Records of major suppliers outweigh those of infrequent suppliers	Accuracy for major suppliers exceeds that for long-tail suppliers by 21-28 percentage points

The third row has a direct managerial consequence. A confidence score above 90% may discourage a user from checking a recommendation, even when the process description behind it is wrong. If a purchaser accepts an unsupported claim about a supplier's manufacturing capability, the chosen firm may be unable to fulfil the order. A missed tolerance requirement can then turn a seemingly efficient shortlist into a delivery or quality problem. Joshi's discussion of hallucination in financial and business settings raises the related problem of

trusting confident but incorrect outputs when decisions are costly [8]. Confidence describes how a system presents its answer; it is not evidence that a supplier possesses the stated capability.

These risks call for controls matched to the task. A supplier's process limits need confirmation by someone with manufacturing expertise, while prices and recent penalties require up-to-date records. Credit-risk estimates should be reviewed with their underlying evidence. This division of checks also affects when AI may be used without delay: routine document triage can precede review, but a contract award should wait until material claims have been verified. A single confidence score cannot replace any of those checks.

4. Conclusion

This review considers four ways in which FinLLMs may assist manufacturing SRM: extracting risk information, combining text with supplier-relationship data, reading engineering documents, and broadening supplier search. The applications address different informational constraints, but none establishes that a model should choose suppliers without review. Reported limitations include fabricated process details, errors in multi-step calculations, high-confidence mistakes, and gaps caused by old or unevenly distributed data. The costs of these errors depend on where a model output enters the procurement workflow.

The proposed "boss-employee" framework specifies three mechanisms. Task division assigns routine, repeated information-processing tasks to AI while leaving problem definition and unusual cases with staff who know the relevant process. The division is based on the consequences of a mistake as well as the time a task takes: a short but unusual exception may demand more expertise than a lengthy routine document check. Authority boundaries keep supplier awards, negotiation strategy, and high-value contract approval under human control; low-value routine approvals may be delegated where rules and risk thresholds are explicit. Staff must know when a case crosses that threshold and who can override an automated action. Calibration requires staff to challenge doubtful outputs and supply knowledge gained in practice. The tool, in turn, should present the evidence behind a recommendation and its degree of uncertainty, allowing staff to compare the output with the final decision and correct subsequent use.

The paper is based on published literature and cases rather than observations or a dataset of Chinese manufacturers. Its proposed framework has not been tested through controlled comparison or repeated application across firms. Future studies can assess supplier evaluation using both text and relationship data in settings with many infrequent suppliers. They can also test how task assignments, approval thresholds, and evidence-review practices work when the human-AI framework is applied in a procurement team. A useful evaluation would record not only processing time but also which model suggestions staff rejected and why. Those decisions would show whether the proposed division of authority protects high-stakes purchases without slowing routine work.

References

[1] Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2020). LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20)*, 1192-1200.

[2] Krieger, F., Drews, P., & Funk, B. (2023). Automated invoice processing: Machine learning-based information extraction for long tail suppliers. *Intelligent Systems with Applications, 20*, 200285.

[3] Li, Y., Zhao, X., & Chen, M. (2025). Can supply chain finance reduce the credit risk of intelligent manufacturing enterprises? *Credit Reference*, (6), 45-53.

[4] Liu, Y., Zhang, H., & Wang, Q. (2026). Supply chain finance credit risk assessment based on graph structures enhanced by large language models. *Journal of Frontiers of Computer Science and Technology, 20*(2), 522-535.

[5] Mittal, U. (2023). AI-based evaluation system for supply chain vulnerabilities and resilience amidst external shocks: An empirical approach. *Reports in Mechanical Engineering*.

[6] Wu, Y. (2026). The dual mechanism of financial large models empowering supply chain financing for micro, small and medium-sized enterprises. *Journal of Industrial Technological Economics*(4), 23-31.

[7] Li, Y., Ko, H., & Ameri, F. (2025). Integrating graph retrieval-augmented generation with large language models for supplier discovery. *Journal of Computing and Information Science in Engineering, 25*(2), 021010.

[8] Joshi, S. (2025). Comprehensive review of AI hallucinations: Impacts and mitigation strategies for financial and business applications.