Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.36665

Research on Digital Transformation Alleviating Financing Constraints and Promoting Green Innovation Through Semantic Measurement and Causal Forests of Large Language Models

Hongyi Xue

Business School, University of Jinan, Jinan, China
18506492226@163.com

Abstract. Digital transformation transforms an enterprise's green innovation capabilities, requiring the measurement and identification of genuine organizational changes and analyzing how capital constraints shape the marginal returns of technological investment. This paper, based on 18,642 annual observations of Chinese A-share listed companies, uses an open-source Chinese large language model to construct a digital transformation index. With financing constraints as the transmission variable, and using the five-fold cross-fitting causal forest to identify the conditional average treatment effect. The results show that digital transformation significantly enhances green innovation and reduces financing constraints; the Bootstrap indirect effect is 0.044, accounting for 23.66% of the total effect. Financing constraints are the most important condition variable in heterogeneous effects, and the digital transformation effect increases from the low constraint group to the higher constraint group, with a slight decline at the highest quantile but no significant difference. The study can reveal the resource boundaries masked by the average effect while maintaining economic implications.

Keywords: Large Language Model, Digital Transformation, Financing Constraints, Green Innovation, Causal Forest

1. Introduction

The returns from green innovation projects are lagging behind, and external capital has difficulty identifying the technological value. Whether digital investment can be transformed into environmental innovation depends on information reorganization, capital availability, and organizational acceptance conditions [1-3].

Previous studies are constrained by measurement biases. The dictionary method cannot distinguish digital embellishment; large models extracting information does not equate to an accurate measurement of organizational change [4-6]; linear models compress resource differences, and average coefficients are difficult to reveal conditional returns.

DOI: 10.54254/2755-2721/2026.36665

This paper constructs digital variables through manual annotation, repeated sampling, and external cross-validation, uses causal forests to depict the distribution of treatment effects, and transforms financing constraints from control variables to transmission mechanisms, revealing that in the constraints, the high range amplifies digital returns and the extreme range forms the upper limit of transformation.

2. Theoretical analysis and research hypotheses

Digital transformation alters the way enterprises identify green opportunities. The technical attributes must be integrated into organizational actions to be effective [1]; unified environmental data and digital simulation reduce the cost of trial and error, shifting the search for green technologies towards continuous identification; data standards integrate processes, and knowledge is reorganized across departments, with expected returns exceeding adjustment costs, thus forming a net boost.

H1: Digital transformation significantly promotes green innovation by enterprises.

Green innovation is highly sensitive to financing conditions. Green patents have strong intangibility, and the separation of technical achievements and environmental benefits across periods leads to high external assessment costs. Digitalization alleviates financing constraints by enhancing the verifiability of business information. Supplementing external data with financial statements reduces the judgment errors of funding parties [2]; standardizing internal processes reduces agency costs and stabilizes research and development cash flow.

H2: Digital transformation alleviates financing constraints by improving verifiability and capital allocation efficiency, thereby promoting green innovation.

This effect varies with the intensity of constraints. In low-constraint enterprises, cash flow is abundant, and the additional financing benefits are limited; in medium-high constraint scenarios, traceable data and transparent processes can release shelved projects, and the effect increases progressively. In extremely high constraint scenarios, the debt repayment pressure erodes redundancies, new resources are used to maintain cash flow, and insufficient supporting R&D capabilities make it difficult to implement green opportunities [3, 7-9], and the effect tends to saturate or even decline.

H3: The promoting effect increases from low constraint to medium-high constraint intervals, and saturates and may decline after entering the extremely high constraint interval.

3. Research design

3.1. Sample and variables

The research subjects are Chinese A-share listed companies. The annual reports， financial information， and green patent application records identified based on the WIPO green patent list are matched. After handling missing values and implementing 1% and 99% quantile truncation for continuous variables， 18，642 enterprise annual observations are obtained. Green innovation (GI) is represented by the natural logarithm of the number of green patent applications by the enterprise plus 1. Financing constraints (FC) are measured by the absolute value of the SA index， with a larger value indicating stronger constraints. Control variables include enterprise size， debt-to-equity ratio， return on total assets， Tobin's Q， enterprise age， proportion of the largest shareholder's shareholding， proportion of independent directors， dual positions， property rights nature， and industry competition degree.

DOI: 10.54254/2755-2721/2026.36665

3.2. Semantic index of large language model

Core explanatory variables come from the management discussion and analysis text in the annual reports. First， the text is segmented by paragraphs， and then the open-source Chinese large language model evaluates digital strategic cognition， digital infrastructure， digital technology application， data element configuration， and digital organizational transformation. Embedded in the small sample instructions are the manually identified anchor paragraphs and their scores, enabling the abstract semantic scale to have comparable judgment criteria. The temperature is set to 0.2. Each paragraph is sampled 5 times and the average is taken. Let i be an enterprise and t be the year. If i contains J_it valid paragraphs in t, and the rth evaluation score of the jth paragraph on dimension d is s_itjdr, and R = 5, then the enterprise's annual index is:

$$Digital_{it} = \frac{\sum_{d=1}^{5}\sum_{j=1}^{J_{it}}\sum_{r=1}^{R} s_{itjdr}}{5J_{it}R} \tag{1}$$

Randomly select 2,000 paragraphs for independent double-labeling by two individuals, and use the intra-group correlation coefficient to test the consistency between the model and the manual judgment.

3.3. Fixed effects and transmission test

The baseline model controls for the characteristics of enterprises that do not change over time and the common year impact, and clusters the standard errors at the enterprise level:

$$GI_{it} = \beta_0 + \beta_1 Digital_{it} + \gamma' X_{it} + \mu_i + \lambda_t + \varepsilon_{it} \tag{2}$$

The transmission relationship of financing constraints is depicted by the following coupled stepwise model:

$$FC_{it} = \alpha_0 + \alpha_1 Digital_{it} + \theta' X_{it} + \mu_i + \lambda_t + u_{it} \tag{3}$$

$$GI_{it} = \delta_0 + \delta_1 Digital_{it} + \delta_2 FC_{it} + \phi' X_{it} + \mu_i + \lambda_t + v_{it} \tag{4}$$

The indirect effect is defined as α_1 × δ_2, and a confidence interval is obtained through 1,000 Bootstrap iterations. This product reflects the transmission evidence consistent with financing constraints.

3.4. Causal forest

Define Digital being higher than the annual median of the industry to which it belongs as the treatment state D = 1, otherwise D = 0. The causal forest estimates the conditional average treatment effect when given enterprise characteristics X = x within the potential outcome framework:

$$\tau(x) = E[GI_i(1) - GI_i(0) X_i = x] \tag{5}$$

The model is set with 2,000 trees, using five-fold cross-validation, with a minimum node size of 10 and a feature sampling ratio of 0.5. The hyperparameters are selected through grid search and combined with R-loss and coverage performance. The conditioning variables include financing

DOI: 10.54254/2755-2721/2026.36665

constraints, enterprise size, R&D intensity, property rights nature, industry competition degree, financial marketization degree, debt-to-asset ratio, Tobin Q and equity concentration. The causal forest can present non-parametric heterogeneity [10, 11], but machine learning does not eliminate the identification boundaries of the observed data [12].

4. Empirical results

4.1. Descriptive features

The semantic index has an intra-group correlation coefficient of 0.821 with the manually annotated group, and the coefficient of variation of multiple sampling is only 0.043, indicating high stability under repeated reasoning at low temperatures. This index has a correlation coefficient of 0.612 with the dictionary index, and is moderately correlated with digital hardware investment and software asset intensity. The LLM measure inherits the effective information in traditional dictionary indicators and also retains additional identification of context, intensity, and organizational meaning. Table 1 summarizes this evidence chain. The means of Digital, GI, and FC in the sample are 5.214, 0.684, and 3.715 respectively, and all three variables have sufficient cross-sectional differences to support panel identification.

Table 1. Semantic validity of the LLM-based digital-transformation index

Validity criterion	Estimate
Intraclass correlation with human coding	0.821***
Coefficient of variation across repeated samples	0.043
Correlation with dictionary-based index	0.612***
Correlation with digital hardware investment	0.453***
Correlation with software asset intensity	0.398***

Note: *** denotes significance at the 1% level. The human-coding sample contains 2,000 paragraphs. The following tables are the same.

4.2. Financing constraint transmission

The benchmark regression shows that the coefficient of Digital for GI is 0.186 (significant at the 1% level); When FC is used as the dependent variable, the Digital coefficient is -0.213, indicating a negative correlation between digitalization level and financing constraints. After incorporating FC into the green innovation equation, the FC coefficient was -0.207, and the Digital coefficient decreased to 0.142 but remained significant; The indirect effect is 0.044, and the Bootstrap 95% confidence interval is [0.021,0.071], accounting for 23.66% of the total effect. The results support H1 and H2.

Table 2. Digital transformation, financing constraints, and green innovation

Variable	(1) GI	(2) FC	(3) GI
Digital	0.186*** (0.042)	-0.213*** (0.055)	0.142*** (0.039)

DOI: 10.54254/2755-2721/2026.36665

Table 2. (continued)

Financing constraints			-0.207*** (0.061)
Controls	Yes	Yes	Yes
Firm fixed effects	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes
Observations	18,642	18,642	18,642
Adjusted R-squared	0.217	0.184	0.231
Indirect effect			0.044*** [0.021, 0.071]

4.3. Heterogeneity of the causal forest

In the ranking of variable importance, the weight of financing constraints is 0.283, significantly higher than that of enterprise size (0.176) and R&D intensity (0.148), making it the primary condition for the split of treatment effects. The CATE increases from 0.081 in Q1 to 0.247 in Q4, the difference between Q4 and Q1 is 0.166, and the 95% confidence interval is [0.037, 0.295]. Q5 drops to 0.213, but the difference between it and Q4 does not pass the significance test. The results support the main judgment of H3. The decline at the upper tail can only be explained as possible marginal saturation, and is not sufficient to confirm the reversal [13].

4.4. Endogeneity and robustness

Table 3 examines the stability of the test conclusions under alternative identification and variable setting. When using the industry annual Digital mean value after excluding the current enterprise as the instrumental variable, the first-stage Kleibergen-Paap F statistic is 68.43, and the second-stage coefficient is 0.224. The ATT of the matched sample is 0.157; replacing green innovation, replacing Digital, excluding the year 2020, and using the lagged Digital all result in a positive core coefficient. The mean of the coefficients obtained from 1， 000 random permutations is 0.003， and the two-tailed p-value is 0.714， reducing the possibility that the results are driven by the random grouping structure.

Table 3. Endogeneity and robustness checks

Specification	Estimate	S.E.	Diagnostic
Instrumental variable	0.224***	0.068	KP F = 68.43
Propensity-score matched ATT	0.157***	0.047	95% CI [0.065, 0.249]
Green patent grants as outcome	0.164***	0.038	Alternative outcome
Dictionary-based Digital	0.121**	0.049	Alternative treatment
Excluding observations in 2020	0.177***	0.041	Sample adjustment
Lagged Digital	0.159***	0.037	Temporal ordering
Placebo assignments	0.003	0.052	Two-sided p = 0.714

DOI: 10.54254/2755-2721/2026.36665

5. Discussion and conclusion

After controlling for financing constraints, Digital's effect on GI declined from 0.186 to 0.142 (indirect effect = 0.044, 23.66%, supporting H1/H2), while the CATE rose from 0.081 in Q1 to 0.247 in Q4 (difference = 0.166, 95% CI [0.037, 0.295]) before falling to 0.213 in Q5 with no significant difference from Q4, indicating that FC has both transmission and effect-boundary effects and that heterogeneity shows tail saturation rather than direction reversal.

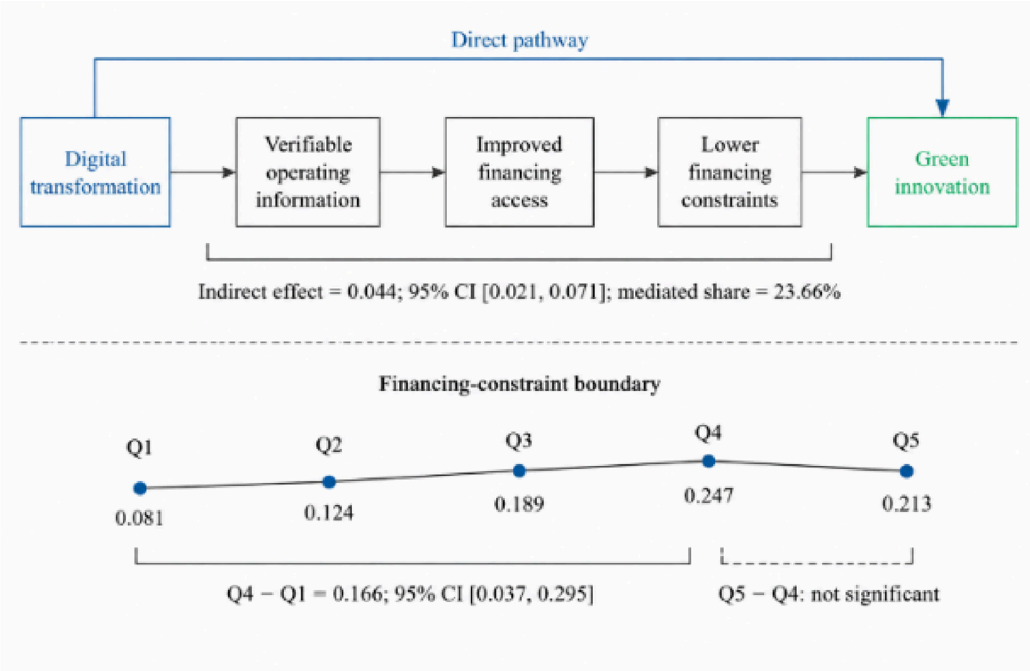

Figure 1. Evidence diagram

In terms of methods, artificial consistency was incorporated into the validity chain, suppressing the "expressive but conceptually distorted" nature of large models; causal forests depicted the distribution of conditional effects.In terms of management, enterprises should uniformly allocate data governance and green R&D budgets; financial institutions can use operational data to reduce due diligence costs while retaining independent audits; policy resources should prioritize supporting enterprises with moderate to high constraints and absorption capabilities, and provide capacity building and phased financing for enterprises with extreme constraints to avoid the accumulation of one-time subsidies.

Research shows that the semantic index of large models can stably identify the essence of enterprise digital transformation. Digital transformation significantly promotes green innovation, and the alleviation of financing constraints constitutes an important resource channel; causal forests show that the returns of green innovation increase from low constraints to medium-high ranges, and there are signs of slowdown at the tail. The focus of digital transformation evaluation should shift from investment scale to verifiable organizational transformation capabilities, and differentiated digital policies should simultaneously examine the degree of financing scarcity and absorption capacity.

DOI: 10.54254/2755-2721/2026.36665

References

[1] Sun, Z., Zhao, L., Mehrotra, A., Salam, M. A., & Yaqub, M. Z. (2024). Digital transformation and corporate green innovation: An affordance theory perspective. *Business Strategy and the Environment, 34*(1), 433-449. https://doi.org/10.1002/bse.3991

[2] Yang, Y., & Han, J. (2023). Digital transformation, financing constraints, and corporate environmental, social, and governance performance. *Corporate Social Responsibility and Environmental Management, 30*(6), 3189-3202. https://doi.org/10.1002/csr.2546

[3] Li, L., & Lin, J. (2023). Digital transformation for the sustainable development of firms: The role of green capability and green culture. *Sustainable Development, 32*(3), 1861-1875. https://doi.org/10.1002/sd.2756

[4] Bronzini, M., Nicolini, C., Lepri, B., Passerini, A., & Staiano, J. (2024). Glitter or gold? Deriving structured insights from sustainability reports via large language models. *EPJ Data Science, 13*(1). https://doi.org/10.1140/epjds/s13688-024-00481-2

[5] Kou, H., Tang, R., & Chen, N. (2025). Enterprise digitalization and ESG performance: Evidence from interpretable AI large language models. *Systems, 13*(9), 832. https://doi.org/10.3390/systems13090832

[6] Mishra, L., Dhibi, S., Kim, Y., Ramis, C. B., Gupta, S., Dolfi, M., & Staar, P. (2024). Statements: Universal information extraction from tables with large language models for ESG KPIs. *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change*, 193-214. https://doi.org/10.18653/v1/2024.climatenlp-1.15

[7] Yu, J., Xu, Y., Zhou, J., & Chen, W. (2024). Digital transformation, total factor productivity, and firm innovation investment. *Journal of Innovation & Knowledge, 9*(2), 100487. https://doi.org/10.1016/j.jik.2024.100487

[8] Wu, W., & Lin, X. (2025). Digital inclusive finance for green transformation: Insight from green innovation, industrial upgrading, and employment quality. *Journal of Innovation & Knowledge, 10*(3), 100726. https://doi.org/10.1016/j.jik.2025.100726

[9] Xu, X., & Liu, Q. (2025). Transformation of innovation in heavily polluting enterprises under resource constraints: The role of green finance policy. *International Review of Economics & Finance, 103*, 104524. https://doi.org/10.1016/j.iref.2025.104524

[10] Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association, 113*(523), 1228-1242. https://doi.org/10.1080/01621459.2017.1319839

[11] Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics, 47*(2), 1148-1178. https://doi.org/10.1214/18-AOS1709

[12] Henningsen, A., Low, G., Wuepper, D., Dalhaus, T., Storm, H., Belay, D., & Hirsch, S. (2025). Estimating causal effects with observational data: Guidelines for agricultural and applied economists. *Journal of Agricultural Economics, 77*(2), 356-382. https://doi.org/10.1111/1477-9552.70019

[13] Hu, Z., Zhang, J. R., & Chang, C.-P. (2025). AI and green credit: A new catalyst for green innovation in Chinese enterprises. *Oeconomia Copernicana, 16*(4), 1447-1489. https://doi.org/10.24136/oc.3020