

A Review of Research on Large Language Model-Assisted Identification of Green Bond Greenium

Aili Sun*

School of Finance, Zhongnan University of Economics and Law, Wuhan 430073, China

**Corresponding author: Aili Sun.*

Abstract

Global climate change has created large-scale demand for green investment and financing, and green bonds have become an important financing instrument through which capital markets support green and low-carbon projects. Greenium reflects the market pricing of the green attributes of bonds. Its identification is a core issue in green finance research and is related to bond pricing, investment decisions, and the improvement of regulatory standards. Existing studies mainly classify bonds into green and non-green categories according to external certification labels. This dichotomous method cannot depict continuous differences in the degree of greenness, and the estimation results also lack consensus because of different measurement criteria. The application of language models in financial text analysis has continued to deepen, but it has not yet entered the research on greenium identification. This study systematically reviews the literature on greenium research and language model applications from 2016 to 2026, and summarizes the formation mechanism of greenium, traditional identification methods, and the technological evolution of language models. On this basis, this study proposes a two-stage framework of text measurement and economic identification. Language models extract green attributes from bond texts, and econometric models estimate the price effect of green attributes after controlling for traditional pricing factors. The division of labor between the two is clear, and they cannot replace each other. This framework expands the measurement of greenium from binary classification to continuous measurement, and can provide methodological reference for empirical research and regulatory practice.

Keywords

green bonds, greenium, large language models, financial text analysis

1. Introduction

Global climate change is an important public policy issue in the twenty-first century. Green bonds are a main financing instrument through which capital markets serve green and low-carbon projects. Since the European Investment Bank issued the first Climate Awareness Bond in 2007, the scale of the green bond market has continued to expand. Greenium refers to the yield difference between comparable conventional bonds and green bonds, and is the core indicator for measuring whether the environmental attributes of bonds are priced. [1] defines it as the difference in yield between comparable conventional bonds and green bonds.

[2] regards this difference as the price expression of pro-environmental preferences. Whether greenium really exists is a main question in green finance research. Existing evidence is not consistent. [3] earlier pointed out that green bonds have an issuance premium in the primary market. The systematic literature review by [4] shows that 56% of primary market studies and 70% of secondary market studies confirm the existence of greenium. For the overall premium of green bonds, [5] finds that it is only about one basis point and is not statistically significant. From the perspective of the Chinese bond market, [6] finds that the yield spread of green bonds is lower than that of ordinary bonds. [7] finds that deep green bonds have a premium of 45.4 basis points relative to non-green bonds, but this premium is not significant relative to ordinary green bonds. These differences show that whether greenium exists and how large it is depend to a large extent on sample selection, market environment, and the measurement of the degree of greenness.

In terms of measurement, external certification labels or manual classification are the main basis of traditional methods. Bonds are thereby classified into green and non-green categories, and then matching methods are used to estimate yield spreads. This dichotomous approach treats green attributes as homogeneous and makes it difficult to reflect continuous differences in the degree of greenness among different bonds. [6] finds that, compared with matched ordinary bonds, the yield spread of green bonds is lower by an average of 17 basis points. However, this estimate is based on binary classification, and it still cannot answer whether there is a price difference between deep green and light green bonds. [8] uses bond pairs issued by the same issuer on the same day and obtains a premium close to zero, showing that the matching criterion directly changes the estimation results. To make up for the limitations of the dichotomous method, text analysis provides a possible approach. [9] uses the BERT model to analyze texts on the use of proceeds in bond prospectuses and constructs a continuous measure of bond greenness. The study finds that when bond greenness increases by 10%, the annualized yield decreases by 4.86 to 8.71 basis points. However, that study still remains at the level of a general language model and does not combine text variables with econometric identification in bond pricing models. Specifically, existing research still has the following gaps. First, greenium research still lacks deep semantic mining of green information in bond issuance texts. Second, the application of language models in green finance mainly focuses on credit risk assessment and the quantification of ESG disclosure, and has not yet been combined with the core issue of greenium. Third, there is a lack of a methodological framework that can connect text measurement and economic identification. Language models can extract green attributes from prospectuses, certification documents, and environmental reports, forming explanatory variables that can be tested. However, language models alone cannot answer whether greenium exists. Text variables need to be placed in bond pricing models, and the size of the premium can be estimated only after controlling for maturity, coupon, credit risk, liquidity, and issuer characteristics.

In response to the above gaps, this study systematically reviews domestic and foreign literature on greenium research and language model applications from 2016 to 2026, and traces some necessary foundational studies. The review covers cross-cutting directions such as green bond pricing, greenium measurement, natural language processing, and financial text analysis. The full text is divided into four parts. The first explains the definition, formation mechanism, and traditional identification methods of greenium. The second summarizes the technological evolution of language models in financial text analysis. The third discusses three types of applications of language models in green finance texts. The fourth proposes a two-stage framework of text measurement and economic identification, discusses key challenges, and proposes a research agenda. The core view of this study is that language models are responsible for extracting green attributes from bond texts, while econometric models are responsible for identifying the price effect of green attributes. The two have a clear division of labor and cannot replace each other. The contributions of this study are reflected in the following aspects. At the theoretical level, this study clarifies the division of labor between language models and econometric models, and introduces text measurement into greenium research. At the practical level, this study reviews the technical routes and applicable conditions of text measurement, and can provide reference for regulators to improve certification standards and for investors to price green assets. At the level of knowledge integration, this study connects the research on greenium and the application of language models, identifies key issues and existing deficiencies in the intersecting field, and provides a list of questions and directional guidance for subsequent research.

2. Definition, Formation Mechanism, and Identification of Greenium

2.1 Concepts, Symbols, and Market Stages

Greenium refers to the yield difference between comparable conventional bonds and green bonds. Its sign convention is that greenium equals the yield of comparable ordinary bonds minus the yield of green bonds. When investors are willing to accept lower yields for green attributes, the financing cost of green bonds decreases. Their yield is lower than that of comparable conventional bonds, and greenium is positive. When green attributes are accompanied by liquidity and information disadvantages, the direction of the premium may also be reversed. [1] defines greenium in a structural pricing model as the difference in yield between comparable conventional bonds and green bonds, and analyzes the formation mechanism of green bond prices. [2] regards this difference as the price expression of pro-environmental preferences. That study argues that green bonds can be used to identify the influence of investors' non-monetary motives on bond market prices. Its theoretical root can be traced to the framework proposed by [10], in which investor preferences affect equilibrium asset prices. [11] constructs a pricing model containing non-monetary utility in a study of U.S. municipal green bonds, and predicts that green bonds will be issued at higher prices than similar ordinary bonds.

When interpreting greenium research, it is necessary to distinguish four different outcome variables. Primary market issuance spread measures the difference between the issuance pricing of green bonds and comparable bonds. Secondary market spread measures the yield difference during the life of the bond. Stock announcement effect measures the impact of issuance announcements on the issuer's stock price. Environmental performance measures the change in the issuer's environmental performance after issuance. The first two can be called greenium. The latter two belong to mechanism evidence or external validity evidence, and cannot be directly called greenium. This distinction is very important for interpreting empirical differences below.

2.2 Formation Mechanism

There is more than one formation mechanism of greenium. The first path is investor preference. Investors are willing to accept lower yields for green attributes, forming a demand-side premium. The framework of [10] provides a theoretical basis for this path. The survey of global institutional investors by [12] shows that most respondents regard climate risk as an important consideration in investment and are willing to pay costs to hedge climate risk. This provides demand-side evidence for the existence of the premium. The second path is risk factors. The studies of [13] and [12] show that carbon risk pricing and institutional investors' perception of climate risk form the macro background for the formation of the premium. The third path is information asymmetry and certification. It is difficult for investors to directly observe green attributes, and third-party certification can therefore provide a credible signal. [3] explained this mechanism earlier, and the results of [14] and [15] also support the value of certification. Once external review or certification is obtained, the premium of bonds becomes larger and the market reaction becomes stronger. [16] finds that greenium mainly appears in bonds with third-party certification, and the credibility of green claims is thereby strengthened. However, certification also has boundaries. Certification standards are not unified, and certification covers labels rather than the actual environmental benefits of the use of proceeds. [17] finds that label value fluctuates over time, and market trust in labels changes with the information environment. The fourth path is supply and demand, taxation, and liquidity. The results of [11] and [18] show that the premium level is affected by holding structure and secondary market competition. [11] also points out that tax treatment changes the comparable basis of yields. [17] finds that the improvement of credit quality changes the direction of the premium. These mechanisms are not mutually exclusive. The dominant mechanisms may differ across markets and periods, and part of the empirical differences below can be understood from the perspective of mechanism differences.

2.3 Traditional Identification Methods and Their Limitations

Existing studies mainly identify greenium through matching methods. [5] matches green bonds with similar non-green bonds issued by the same issuer on the basis of daily spreads, and finds that in the full sample green bonds are only slightly tighter than comparable conventional bonds by about 1 basis point. [19] uses propensity score matching and coarsened exact matching to estimate 1,928 global green bonds, and finds that green bond yields are on average 15 to 20 basis points lower than those of conventional bonds. [20] introduces externally

assessed greenness ratings on the basis of matching, and finds that deep green bonds have a significant premium relative to light green bonds. [8] uses almost fully comparable bond pairs issued by the same issuer on the same day, and finds almost no pricing difference between green bonds and homogeneous non-green bonds. The limitations of matching methods are relatively obvious. The number of comparable bond pairs is relatively small, and the estimation results directly depend on the selection of matching variables. The premium close to zero obtained by [8] in an almost fully comparable sample suggests that some previous estimates may have mixed in bias caused by insufficient comparability. Another limitation lies in the measurement of green attributes. Dichotomous methods and discrete ratings treat green attributes as homogeneous or roughly layered. The substantive differences in the degree of greenness across different uses of proceeds are difficult to depict by these methods, and the heterogeneity within green bonds is therefore easily ignored. If the degree of greenness needs to be measured continuously, environmental information must be extracted directly from bond texts. This demand provides conditions for introducing text measurement methods.

2.4 Empirical Differences and the Need for Text Measurement

Empirical studies on whether greenium exists have not yet formed a consensus. The differences are not disorderly. There is a systematic correspondence between the evidence structure and the strictness of matching methods. The systematic review by [4] shows that 56% of primary market studies and 70% of secondary market studies confirm the existence of a premium. However, the confirmation rate is directly related to the matching criterion adopted by the study. In studies with looser matching, the premium is more common. [2] estimates an average of about 2 basis points. [21] obtains about 18 basis points in the European primary market. The global sample of [19] gives 15 to 20 basis points. When the comparability requirement becomes stricter, the premium shrinks significantly. [8] uses bond pairs issued by the same issuer on the same day and obtains a result close to zero. [22] finds that the premium is no longer significant after controlling for firm characteristics. This gradient indicates that part of the price difference attributed to green attributes may come from uncontrolled traditional pricing factors. Another source of differences lies in issuer structure. The results of [23] and [24] both point to systematic differences in issuer types. The premium direction can be opposite for institutional issuers and corporate issuers. [17] finds that the value of the green label in the U.S. municipal bond market changes over time from positive to negative and then to positive again. This shows that the premium reflects the result of a market learning process rather than a static parameter. Overall, greenium is more likely to exist conditionally. Its appearance depends on the joint effect of investor structure, market development stage, and measurement criterion. Empirical differences also show that consistent conclusions are difficult to obtain only by comparing labels and prices. The identification of the degree of greenness is the last link in various measurement criteria. Insufficient identification accuracy will be directly transformed into estimation bias. This provides the motivation for discussing text measurement below.

3. Technological Evolution of Language Models in Financial Text Analysis

3.1 Dictionary Methods and Their Limitations

The domain adaptation limitation of dictionary methods is the primary factor promoting the evolution of financial text analysis methods. Early financial text analysis mainly relied on dictionary methods to measure textual tone. [25] finds that the Harvard general negative dictionary has serious misclassification in 10-K texts. About 73.8% of words marked as negative do not have negative meanings in the financial context. Based on this, that study constructs word lists such as Fin-Neg for financial texts. These word lists are significantly more correlated with document daily returns, trading volume, and future earnings surprises than general dictionaries. This result shows that semantic measurement of financial texts must first solve the problem of domain adaptation. General language resources are difficult to transfer directly to financial scenarios. After dictionary methods, some studies use bag-of-words vectors with traditional machine learning models such as support vector machines or logistic regression for text classification. Such methods still rely on manually designed features. [26] points out that such classification methods have implementation pitfalls in accounting and financial texts. Dictionary methods themselves are static word matching and have difficulty dealing with polysemy and context dependence. Taking green bond texts as an example, the same word in descriptions of the use of proceeds may convey different green intensity in different contexts. Dictionary methods cannot

capture such differences. It is precisely these limitations that have led researchers to turn to context-based language models.

3.2 Domain Adaptation of Encoder Models

Using domain corpora for continued pre-training on a general architecture is the key path for financial language models to achieve performance improvement. The core of this path is domain adaptation, that is, continuing pre-training with domain-specific texts on the basis of a general language model, so that the model can master the terminology and expression habits of the domain.

BERT is the general starting point of this path. BERT stands for Bidirectional Encoder Representations from Transformers, and was proposed by Google in 2018 [27]. Its core design is bidirectional encoding. When the model processes a word, it uses both the left and right contextual information of the word. This feature is implemented through masked language modeling, in which the model randomly masks part of the input words and then predicts the masked words according to the context. In addition to masked language modeling, BERT also designs a next sentence prediction task to help the model understand the semantic relationship between sentences. These two pre-training tasks enable BERT to learn general language representations from unlabeled texts and then adapt to downstream tasks through fine-tuning. FinBERT for financial texts is formed by domain adaptation on the basis of BERT. To continue pre-training BERT, [28] uses financial corpora such as corporate reports, earnings conference calls, and analyst reports with 4.9 billion tokens. In financial sentiment classification tasks, FinBERT significantly outperforms general BERT. To enable the model to capture financial vocabulary and semantic knowledge more fully, [29] designs six pre-training tasks for FinBERT, and training is conducted on both general and financial corpora. [30] also independently proposes a FinBERT variant for financial sentiment analysis. These studies show that the semantic gap between general models and financial scenarios can be effectively narrowed through domain adaptation. FinBERT-type models can identify expressions related to green attributes in green bond texts, such as prospectuses, and this provides a semantic basis for subsequently constructing greenness indicators. Such models still belong to the encoder architecture and are mainly applicable to classification tasks. The core function of an encoder is to transform input text into vector representations. In operation, the aggregated representation of the whole sequence is often taken from the final hidden state of the special classification token at the beginning of the sequence for classification. This design makes BERT and FinBERT perform well in tasks such as text classification, sentiment analysis, and entity recognition. However, the encoder architecture does not have autoregressive generation capability and has difficulty directly completing text generation tasks. This architectural boundary therefore becomes the direct motivation for the evolution toward generative models.

3.3 Generative Models and Instruction Tuning

Recent research focuses on generative models and instruction tuning, and such methods attempt to complete financial tasks with less labeled data. Instruct-FinGPT proposed by [31] can transform financial sentiment classification data into instruction form and fine-tune LLaMA. The experimental results show that Instruct-FinGPT outperforms traditional supervised models and general large language models on multiple financial sentiment analysis datasets. Taking the FPB dataset as an example, the model has an accuracy of 0.76 and an F1 score of 0.74. ChatGPT 3.5 has an accuracy of 0.64 and an F1 score of 0.51. LLaMA-7B has an accuracy of 0.60 and an F1 score of 0.40. Overall, in terms of accuracy and F1 score, this method can achieve performance improvement of 15% to 48%. In terms of numerical sensitivity, when a news item mentions that losses decreased from 2.2 million to 0.3 million, FinBERT judges it as negative, but ChatGPT 4.0 and Instruct-FinGPT can correctly identify the positive signal corresponding to this obvious decrease. At the same time, BloombergGPT trained by [32] demonstrates the feasibility of large-scale pre-training on proprietary financial corpora. The potential of ChatGPT in predicting the direction of stock returns is also verified by [33]. The core implication of these advances is that generative models can adapt to financial tasks with less labeled data, but their performance depends to a large extent on domain instruction design and numerical reasoning capability. Financial text analysis has evolved from dictionary methods to pre-trained language models and then to generative models, experiencing a complete technological evolution process. This process provides a technical basis for the automatic analysis of green finance texts.

3.4 Model Reliability and Application Boundaries

In financial text analysis, the application of large language models is still restricted by reliability problems, among which hallucination is the most central. Hallucination is defined by [34] as content generated by a model that is superficially reasonable but lacks basis or contains factual errors. That study further points out that a fundamental defect of large language models is hallucination, which is especially prominent in fields such as finance, education, and law. The reason why hallucination can become a core topic is that financial decisions have extremely high requirements for information accuracy. The empirical study of [34] finds that off-the-shelf large language models show serious hallucination behavior in financial tasks. In the financial field, erroneous outputs may cause investment decision errors, regulatory violations, or economic losses.

To alleviate this problem, existing studies have tried external knowledge grounding. ChatClimate proposed by [35] grounds GPT-4 in authoritative sources such as the Sixth Assessment Report of the IPCC, and the accuracy and verifiability of climate question answering are significantly improved. [36] introduces source attribution in ChatReport to reduce hallucination risk. These studies show that the key mechanisms for improving model credibility are external knowledge grounding and answer traceability. Overall, there are currently two main paths to improve the credibility of financial text analysis, namely knowledge grounding and answer traceability. In high-stakes decision scenarios, manual review still cannot be replaced. This judgment also applies to the green finance text applications discussed below.

4. Applications of Language Models in Green Finance Texts

4.1 Semantic Parsing and Measurement of Substantive Greenness

As stated in Section 2.4, the basic conclusion of green bond pricing research is that evidence on greenium is not consistent. A simple price difference is insufficient to measure the value of green attributes. [2] uses a paired method to estimate green bond premiums and finds that green bond yields are on average about 2 basis points lower during the sample period. This effect is more obvious in low-rated bonds and financial bonds. [15] points out that the market reaction is positive after firms issue green bonds, but matched green bonds and conventional bonds show no significant difference in yields. Meanwhile, issuers' subsequent environmental performance improves. The different results of the two studies show that the pricing effect depends on sample selection, matching method, and bond type. This difference reminds researchers that they need to look for a more direct measure of substantive greenness beyond price indicators.

Text data provide supplementary information for measuring green attributes beyond price indicators. This is an important extension direction of green bond research. [37] develops firm-level climate exposure indicators based on earnings conference calls. That study uses a machine learning keyword discovery algorithm to capture three types of climate shocks: opportunity, physical, and regulatory shocks. These indicators can predict green employment and green patents. ClimateBERT pre-trained by [38] is designed specifically for climate-related texts, and performs better than general BERT in climate topic detection and sentiment classification tasks. These text indicators can reflect heterogeneous climate exposure among firms in the same industry. However, their data sources are mostly public conference calls or news, and they less often directly parse green attributes in bond prospectuses. This field still requires further research.

Applying natural language processing to green bond texts is expected to achieve more fine-grained measurement of substantive greenness. [39] uses ClimateBERT to analyze corporate climate risk disclosure and finds cheap talk and selective disclosure. [40] uses NLP methods to quantify the environmental, social, and governance dimensions in corporate ESG communication. Although these studies take corporate climate disclosure and ESG communication texts as their objects, they show that language models have the ability to assess specificity, commitment strength, and sentiment tendency in descriptions of the use of proceeds. Related methods are expected to be transferred to identify the deviation between nominal greenness and actual greenness of green bonds. Connecting text indicators with post-issuance environmental performance is a key research link that connects green bond pricing with real environmental impact.

4.2 Automatic Evaluation of Disclosure Quality

Corporate environmental disclosure is shifting from voluntary to mandatory, and the volume of disclosure is expanding accordingly. Automatic quality evaluation has thus become a practical need. For non-financial reports of Spanish companies from 2015 to 2022, [41] compares voluntary and mandatory situations. That study finds that mandatory disclosure systems improve the specificity and commitment level of climate-risk-related texts. However, there are differences across industries and institutional stages. After the implementation of the EU Non-Financial Reporting Directive and the Corporate Sustainability Reporting Directive, the scope of information that firms need to disclose will expand, and manual evaluation will have difficulty covering massive reports. The gap between disclosure scale and evaluation capacity directly promotes the development of text evaluation methods.

Large language models provide a scalable solution for disclosure quality evaluation. The ChatReport system developed by [36] is based on TCFD recommendations and can automatically evaluate the compliance of corporate sustainability reports. The system provides traceable question answering and scoring through retrieval-augmented generation. [40] uses ClimateBERT-NetZero to detect and evaluate corporate net-zero targets and emission reduction commitments. When reports are long and structurally complex, these systems can extract key information and conduct consistent evaluation according to disclosure frameworks. Whether such systems can be effective depends on prompt engineering, retrieval accuracy, and the integration of domain expert feedback. This is also an important condition for whether they can enter regulatory practice. The reliability problems of models in disclosure analysis have been discussed in Section 3.4.

4.3 Greenwashing Identification and Credibility Assessment of Claims

The core issue of greenwashing research is how to distinguish corporate environmental claims from actual performance. The basis of the concept of greenwashing lies precisely in the mismatch between the two. [42] defines greenwashing as a combination of poor environmental performance and positive environmental communication, and points out that loose regulation and weak enforcement are the main determinants. [43] further expands the conceptual boundary of greenwashing and argues that any misleading environmental communication that causes stakeholders to form overly positive perceptions can be included. Greenwashing types at the corporate and product levels and at the execution and claim levels are reviewed in the systematic review of [44]. These conceptual frameworks provide judgment criteria for empirical detection. The credibility of green claims depends on the fit among claim content, environmental performance, and communication intention.

Natural language processing provides scalable tools for greenwashing detection. Detection capability has thereby expanded from case-by-case review to batch text processing. [45] constructs a greenwashing risk indicator based on four dimensions: sentiment, commitment, specificity, and hedging language. That study fine-tunes ClimateBERT to classify corporate sustainability reports, and obtains an average accuracy of 86.34% on the test set. [39] proposes red-flag indicators for net-zero transition plans to identify inconsistency between targets and actions. These automated methods can capture text patterns that are difficult for traditional manual review to cover, such as vague commitments and overly positive tone. However, existing studies mostly rely on small-scale expert-labeled data. Model generalization ability is still constrained by training samples. The contradiction between large-scale detection demand and labeled data constraints is the main bottleneck currently faced by greenwashing identification research.

The credibility assessment of green claims needs to combine text signals with performance data. [46] emphasizes that corporate environmentalism contains both material and symbolic dimensions. Researchers need to avoid directly equating symbolic indicators such as certification labels with real environmental improvement. [45] introduces corporate carbon emissions data into model validation and finds inconsistency between the text risk scores of high-emission firms and actual emissions. This cross-validation between text and performance is the key path to distinguishing credible green claims from greenwashing. Future research needs to develop an integrated evaluation framework that can simultaneously understand linguistic details, verify factual bases, and track subsequent performance.

5. Paths, Challenges, and Research Agenda for Text Measurement and Economic Identification

5.1 Text Measurement Stage

Whether greenium exists depends on how green attributes are defined and measured. This is the logical starting point of this Section. As stated in Section 2.4, existing empirical studies mainly rely on price comparisons, but their conclusions are not consistent. [2] uses a paired method and finds that green bond yields are on average about 2 basis points lower. The premium is more obvious for low-rated bonds and financial bonds. [15] finds that corporate green bond issuance has a positive announcement effect, but matched green bonds and conventional bonds show no significant difference in yields. [8] further points out, based on a study of U.S. municipal bonds, that the green label itself is difficult to use to explain yield differences. These differences are not accidental, but reflect the inherent limitations of research methods. Matching methods have difficulty fully controlling issuer characteristics. During the sample period, the green bond market was still developing. Differences in secondary market liquidity also interfere with price comparisons. More importantly, the green label is a binary variable and cannot reflect the degree of greenness of the use of proceeds. This information needs to be obtained from texts. Therefore, incorporating the semantic features of bond texts into greenium research is a necessary direction to make up for the insufficiency of price comparison methods.

The specific process of the text measurement stage is as follows. First, semantic features are extracted from bond prospectuses, certification documents, and environmental reports. Second, indicators are constructed separately according to the three constructs defined in Section 4, namely substantive greenness, disclosure quality, and claim credibility. Third, the validity of the indicators is verified by using green taxonomies, expert annotation, and post-issuance environmental performance. The role of language models at this stage is variable construction. The systematic review of [4] points out that the main determinants of greenium include use of proceeds, certification status, issuer characteristics, and market conditions. Information such as use of proceeds and certification status mostly exists in textual form. Text is therefore the main carrier of green attributes. [9] has made an initial attempt along this path. That study uses BERT to encode use-of-proceeds texts and constructs a continuous greenness indicator, confirming that textual greenness is associated with bond yields. However, this attempt still has obvious limitations. The model it relies on is still a BERT pre-trained in a general domain and has not been adapted to the terminology system and contextual characteristics of green bond texts. The extraction accuracy of green semantic features is therefore constrained. In addition, that study constructs only a single greenness indicator and does not distinguish substantive greenness, disclosure quality, and claim credibility. The three constructs should be tested separately rather than merged into one score. Detailed disclosure does not mean that the project is greener. Sufficient certification also does not mean that actual benefits are higher.

The quality of text measurement depends on training data and model adaptation. The measurement of green attributes lacks a unified standard. This is a deep problem faced by training data for text models. [47] compares data from six mainstream ESG rating agencies and finds that measurement divergence contributes 56% of rating divergence, scope divergence accounts for 38%, and weight divergence accounts for 6%. Rating agencies show systematic inconsistency in scores for the same firm. [48] further finds that more ESG disclosure instead intensifies rating divergence. The reason is that rating agencies interpret the same information differently. These findings provide direct warnings for text measurement. If a model uses rating results or labels as training targets, the divergence in the labels themselves will be transmitted to the model output. An increase in disclosure does not necessarily bring an improvement in consistency. Simply expanding the scale of training data cannot solve the problem of inconsistent annotation standards. Researchers need to disclose the generation process of model training data, clarify the criteria for determining green labels, and conduct cross-validation among multiple annotation sources.

Domain adaptation is a precondition for the text measurement stage. Existing studies support this judgment. FinBERT proposed by [30] adapts BERT to financial texts through continued pre-training and significantly outperforms general BERT in sentiment classification tasks. ClimateBERT pre-trained by [38] continues pre-training for climate-related texts and performs better in climate topic identification and sentiment classification. The two studies jointly show that there are systematic differences between the professional terminology, expressions, and contexts of the financial and climate domains and those of general texts. The effect of domain adaptation depends not only on the model architecture, but also on the scale, source, and time coverage of the

pre-training corpus. Updates to climate policy and the evolution of disclosure frameworks will change the semantic distribution of texts. Static domain models may fail over time. Cross-market text distribution differences further restrict the generalization ability of models. The review by [49] points out that when the vocabulary, syntax, and topic distributions of the source domain and the target domain differ greatly, model performance declines significantly. Green bond disclosure in the European market follows the EU taxonomy. The green bond standards in the Chinese market differ from it. A model trained on texts from a single market may produce systematic bias in another market.

5.2 Economic Identification Stage

Text measurement only solves the problem of variable construction. Whether greenium exists and how large it is need to be answered by the economic identification stage. This stage places text indicators into a bond pricing model, controls for traditional pricing factors such as maturity, coupon, credit rating, liquidity, tax, and issuer characteristics, and then estimates the price effect of green attributes. It needs to be emphasized that prediction is not equal to causal identification. [50] positions machine learning as a prediction tool, emphasizes its advantages in high-dimensional variable selection and nonlinear modeling, but points out that prediction and causal identification are two different matters. [51] reviews machine learning methods that economists should know, with particular attention to the estimation of average treatment effects in causal inference. Textual greenness may be simultaneously related to issuer size, credit quality, disclosure ability, certification choice, and project quality. Simple regression can only show correlation and cannot prove that green semantics cause financing costs to decline. Therefore, text variables should not be directly interpreted as evidence of a premium. Causal identification is the basic requirement of the economic identification stage.

Strategies in the economic identification stage can draw on the existing literature on greenium. The case in which the same issuer simultaneously issues green and ordinary bonds provides the closest paired sample. [8] uses bond pairs issued by the same issuer on the same day, which is representative of this strategy. Issuer fixed effects can absorb issuer characteristics that do not change over time. Policy shocks provide exogenous variation. [17] finds that label value changes over time, showing that changes in policy and information environments can be used as timing points for identifying the premium. When data conditions allow, methods such as difference-in-differences or instrumental variables can also be used for identification. After text indicators enter the pricing equation, it is necessary to test their marginal contribution to yields and judge whether this contribution changes with certification status, issuer type, or market environment.

The economic identification stage faces five main types of threats. The first type is strategic disclosure. Issuers may selectively emphasize favorable information and weaken unfavorable information. [39] finds cheap talk and selective disclosure in corporate climate disclosure. If text indicators reflect disclosure strategy rather than real green attributes, the price effect will be wrongly attributed. The second type is selection bias. Whether to issue green bonds is itself the result of issuers' self-selection. The signaling view of [15] shows that issuance decisions convey information about environmental commitment, and this selection process may affect text features and pricing at the same time. The third type is reverse causality. Market pricing may in turn affect issuers' disclosure behavior, causing the correlation between text indicators and yields to contain bidirectional components. The fourth type is data leakage. The time split between model training and testing data must be strict; otherwise, evaluation results will overestimate model performance. The fifth type is cross-market transfer. As stated in Section 5.1, a model trained in a single market may produce systematic bias in other markets. These five types of threats need to be addressed one by one in the research design; otherwise, the estimated price effect of text indicators is not reliable.

5.3 Interpretability and Regulatory Applicability

The application of large language models in financial decisions raises concerns about interpretability. Interpretability should be embedded in model design rather than added afterward. [52] points out that post-hoc explanations of black-box models may conceal their internal defects. In high-stakes decisions, inherently interpretable models should be prioritized. In the view of [53], interpretability is not only a technical issue, but is also related to whether models can be safely deployed in the real world. Greenium identification directly affects asset pricing and investment allocation. If the model decision process is opaque, it is difficult to distinguish whether its judgment is based on real green attributes or statistical spurious correlation. Even if post-hoc explanation methods are adopted, the results may be misleading. [54] demonstrates that attention

weights do not necessarily correspond to the true basis of model decisions. Completely different attention distributions may also produce the same prediction. Attention heat maps or feature importance may be mistakenly regarded by regulators and investors as causal evidence, but these tools actually reflect statistical correlation of the model. Therefore, model design should strike a balance between accuracy and interpretability, rather than trying to explain model behavior after the model has been completed.

Whether green finance text analysis has ultimate value depends on whether it can enter regulatory and compliance processes. This places higher requirements on model outputs. In a review of climate finance, [55] points out that identifying and pricing climate risk requires the combination of text data, market data, and policy signals, and a single data source is difficult to use to support regulatory decisions. [56] discusses the strategic initiatives of the EU to address greenwashing and emphasizes the role of disclosure standards and third-party verification in preventing greenwashing. [57] analyzes the effectiveness of EU disclosure regulation in limiting greenwashing and finds that the regulation still faces consistency challenges during implementation. These studies jointly show that what regulators need is not only the result of model judgment, but also the basis of judgment, data sources, and reproducible processes. If language models cannot provide a traceable evidence chain, their conclusions are difficult to adopt in regulatory practice. Model design needs to incorporate the three standards of accuracy, interpretability, and regulatory compliance at the same time. The implementation framework of the three standards is discussed in Section 5.4.

5.4 Research Agenda

Corresponding to the above framework, the future research agenda can be divided into the following four directions. The first direction is to establish multidimensional green text constructs and expert-annotated benchmarks. Section 4 has shown that substantive greenness, disclosure quality, and claim credibility are three different constructs. Existing research still lacks annotated datasets for green bond texts. Future research should invite experts in green finance to annotate prospectuses, certification documents, and environmental reports, and establish publicly available benchmarks for the three constructs separately. Such benchmarks can be used for model training as well as for indicator validity testing.

Green semantic indicators should also be connected with priced risk factors. [58] finds that firms with higher carbon emissions in the U.S. stock market obtain higher returns. This shows that investors have required compensation for carbon emission risk. [59] extends this evidence to 77 countries and points out that the carbon transition risk premium is higher in countries with stricter domestic climate policies. [60] proposes a pollution premium and argues that environmental policy uncertainty is a new source of systematic risk. If green semantic indicators cannot be linked with carbon emissions, transition risk, or policy sensitivity, it is difficult for them to explain asset price differences. Therefore, indicator validity testing requires not only expert annotation, but also connection with these priced risk factors.

The second direction is to carry out domain adaptation for financial, climate, and bond texts. As stated in Section 5.1, general models and static domain models are both insufficient for processing green bond texts. [61] systematically reviews the technical routes of large language models in the financial field and points out that domain adaptation, instruction tuning, and hallucination control are key challenges. [62] reviews the pre-training, post-training, and utilization strategies of large language models from the technical level, and emphasizes the importance of evaluation and alignment. Future research needs to strike a balance between public financial corpora and proprietary green bond texts, and avoid overfitting to specific labels or market environments. At the same time, a dynamically updated and auditable green semantic benchmark should be established, so that model outputs not only reflect language patterns but can also track subsequent environmental performance.

The third direction is to combine text measurement with strict econometric identification. This direction directly responds to the threats proposed in Section 5.2. [26] points out that text analysis in accounting and finance is much rougher than traditional quantitative methods and has many pitfalls in implementation. Future research should report evidence on the validity of text indicators, the applicable conditions of identification strategies, and robustness test results in the same framework. The study of [3] shows that the secondary market performance of green bonds is similar to that of conventional bonds, and the premium is mainly generated at the primary issuance stage. Future research can accordingly estimate the price effects of the primary market and the secondary market separately, and examine the performance of text indicators under different policy

cycles and market sentiment. The survey of 861 financial experts, practitioners, and regulators by [63] shows that respondents generally believe that asset prices underestimate climate risk, and regulatory risk receives the most attention in the short term. Investor beliefs themselves change with policy and market environments. If text signals measure such beliefs, their explanatory power may also change over time.

The fourth direction is to establish an interpretable, traceable, and auditable regulatory framework. This direction responds to the discussion in Section 5.3. [64] distinguishes transparency from post-hoc explanation and points out that many studies define interpretability vaguely, and that linear models are not necessarily naturally interpretable. The SHAP method proposed by [65] provides a unified local explanation framework for complex models. In the view of [66], explanations in interactive learning are important for user trust. The LIME method proposed by [67] makes local explanations of any classifier possible. [52] argues that inherently interpretable models should be prioritized in high-risk scenarios. Transferring these principles to the field of green finance, models should not only output indicator values, but also provide a verifiable evidence chain and remain consistent with certification standards and emissions data. Future research should establish evaluation standards for explanation methods and distinguish explanations that serve understanding from explanations that serve regulatory accountability.

In summary, this study reviews the concept, mechanism, and empirical differences of greenium, summarizes the technological evolution of language models and their applications in green finance texts, and proposes a two-stage framework of text measurement and economic identification. Language models are expected to improve the measurement accuracy of green attributes. Whether they can improve the credibility of greenium identification depends on construct validity, external performance validation, and econometric identification strategies. Only when this premise is met can language models truly support greenium identification, greenwashing detection, and regulatory decisions.

6. Conclusions

This study systematically reviews the literature in the two fields of greenium research and language model applications from 2016 to 2026, and the main findings are as follows. First, the empirical evidence on greenium is not consistent. Whether the premium appears and how large it is depend on investor structure, market development stage, and measurement criterion. Green attributes are more likely to be conditionally priced. Second, traditional identification methods rely on certification labels for binary classification and cannot depict continuous differences in the degree of greenness. Once the matching criterion differs, the estimation results also change. Third, financial text analysis has evolved from dictionary methods to encoder models and then to generative models. Language models can already extract green semantic features from bond texts, but they have not yet been combined with greenium identification research.

Accordingly, this study proposes a two-stage framework of text measurement and economic identification. Language models are responsible for constructing green attribute indicators from prospectuses, certification documents, and environmental reports. Econometric models are responsible for estimating price effects after controlling for traditional pricing factors such as maturity, credit risk, and liquidity. The division of labor between the two is clear, and they cannot replace each other. This framework expands greenium measurement from binary classification to continuous measurement, provides methodological reference for empirical research, and also provides a basis for regulators to improve certification standards and for investors to price green assets.

The limitations of this study mainly include three aspects. First, as a review study, the proposed two-stage framework has not yet been tested with empirical data, and its validity remains to be verified by subsequent research. Second, annotated benchmarks for green bond texts have not yet been established, and the definitions of the three constructs of substantive greenness, disclosure quality, and claim credibility still need to be refined. Third, model hallucination, cross-market transfer, and interpretability problems have not been fully solved.

Future research can proceed from four directions. First, expert-annotated green text benchmarks should be established, and semantic indicators should be connected with priced risk factors such as carbon emissions. Second, domain adaptation of green bond texts should be carried out to improve feature extraction accuracy. Third, the price effect of text indicators should be tested under strict econometric identification, and threats such as strategic disclosure, selection bias, and reverse causality should be addressed. Fourth, an interpretable,

traceable, and auditable evaluation process should be established. At the policy level, regulators can promote standardized disclosure of green bond offering texts, introduce text technology to assist certification review and greenwashing monitoring, and gradually incorporate automatic evaluation into the green finance regulatory framework.

References

- [1] Agliardi, E., & Agliardi, R. (2019). Financing environmentally sustainable projects with green bonds. *Environment and Development Economics*, 24(6), 608-623.
- [2] Zerbib, O. D. (2019). The effect of pro-environmental preferences on bond prices: Evidence from green bonds. *Journal of Banking & Finance*, 98, 39-60.
- [3] Ehlers, T., & Packer, F. (2017). Green bond finance and certification. *BIS Quarterly Review*, September 2017, 89-104.
- [4] MacAskill, S., Roca, E., Liu, B., Stewart, R. A., & Sahin, O. (2021). Is there a green premium in the green bond market? Systematic literature review revealing premium determinants. *Journal of Cleaner Production*, 280, 124491.
- [5] Hachenberg, B., & Schiereck, D. (2018). Are green bonds priced differently from conventional bonds? *Journal of Asset Management*, 19(6), 371-383.
- [6] Zhang, L., Liu, J., & Wang, H. (2021). Does greenium exist? Evidence from China's green bond market. *China Journal of Economics*, 8(2), 45-72. <https://doi.org/10.16513/j.cnki.cje.20210602.002>.
- [7] Shen, H., & Wang, K. (2026). Rating inflation in the green bond market: Evidence from large language models. *Chinese Review of Financial Studies*, 18(2), 72-103.
- [8] Larcker, D. F., & Watts, E. M. (2020). Where's the greenium? *Journal of Accounting and Economics*, 69(2-3), 101312.
- [9] Benincasa, E., Fu, J., Mishra, M., & Paranjape, A. (2022). Different shades of green: Estimating the green bond premium using natural language processing. *Swiss Finance Institute Research Paper Series 22-64*, Swiss Finance Institute. SSRN Electronic Journal. https://www.ssrn.com/abstract_id,4198065.
- [10] Fama, E. F., & French, K. R. (2007). Disagreement, tastes, and asset prices. *Journal of Financial Economics*, 83(3), 667-689.
- [11] Baker, M., Bergstresser, D., Serafeim, G., & Wurgler, J. (2022). The pricing and ownership of U.S. green bonds. *Annual Review of Financial Economics*, 14, 415-437.
- [12] Krueger, P., Sautner, Z., & Starks, L. T. (2020). The importance of climate risks for institutional investors. *The Review of Financial Studies*, 33(3), 1067-1111.
- [13] Bolton, P., & Kacperczyk, M. (2021). Global pricing of carbon-transition risk. *Journal of Financial Economics*, 142(2), 468-496.
- [14] Hyun, S., Park, D., & Tian, S. (2020). The price of going green: The role of greenness in green bond markets. *Accounting & Finance*, 60(1), 73-95.
- [15] Flammer, C. (2021). Corporate green bonds. *Journal of Financial Economics*, 142(2), 499-516.
- [16] Kapraun, J., Latino, C., Scheins, C., & Schlag, C. (2021). (In)-credibly green: Which bonds trade at a green bond premium? In *Proceedings of Paris December 2019 Finance Meeting EUROFIDAI-ESSEC*.
- [17] Karpf, A., & Mandel, A. (2018). The changing value of the green label on the U.S. municipal bond market. *Nature Climate Change*, 8(2), 161-165.
- [18] Partridge, C., & Medda, F. R. (2020). Green bond pricing: The search for greenium. *The Journal of Alternative Investments*, 23(1), 49-56.
- [19] Löffler, K. U., Petreski, A., & Stephan, A. (2021). Drivers of green bond issuance and new evidence on the greenium. *Eurasian Economic Review*, 11(1), 1-24.

- [20] Dorfleitner, G., Utz, S., & Zhang, R. (2022). The pricing of green bonds: External reviews and the shades of green. *Review of Managerial Science*, 16(3), 797-834.
- [21] Gianfrate, G., & Peri, M. (2019). The green advantage: Exploring the convenience of issuing green bonds. *Journal of Cleaner Production*, 219, 127-135.
- [22] Tang, D. Y., & Zhang, Y. (2020). Do shareholders benefit from green bonds? *Journal of Corporate Finance*, 61, 101427.
- [23] Bachelet, M. J., Becchetti, L., & Manfredonia, S. (2019). The green bonds premium puzzle: The role of issuer characteristics and third-party verification. *Sustainability*, 11(4), 1098.
- [24] Fatica, S., Panzica, R., & Rancan, M. (2021). The pricing of green bonds: Are financial institutions special? *Journal of Financial Stability*, 54, 100873.
- [25] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35-65.
- [26] Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), 1187-1230.
- [27] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171-4186).
- [28] Yang, Y., Uy, M. C. S., & Huang, A. (2020). FinBERT: A pretrained language model for financial communications. arXiv. <https://doi.org/10.48550/arXiv.2006.08097>
- [29] Liu, Z., Huang, D., Huang, K., Li, Z., & Zhao, J. (2020). FinBERT: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI 2020)* (pp. 4513-4519).
- [30] Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv. <https://doi.org/10.48550/arXiv.1908.10063>
- [31] Zhang, B., Yang, H., & Liu, X.-Y. (2023). Instruct-FinGPT: Financial sentiment analysis by instruction tuning of general-purpose large language models. arXiv. <https://doi.org/10.48550/arXiv.2306.12659>
- [32] Wu, S., Irsoy, O., Lu, S., Dabrowski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., & Mann, G. (2023). BloombergGPT: A large language model for finance. arXiv. <https://doi.org/10.48550/arXiv.2303.17564>
- [33] Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. arXiv. <https://doi.org/10.48550/arXiv.2304.07619>
- [34] Kang, H., & Liu, X.-Y. (2023). Deficiency of large language models in finance: An empirical examination of hallucination. arXiv. <https://doi.org/10.48550/arXiv.2311.15548>
- [35] Vaghefi, S. A., Stambach, D., Muccione, V., Bingler, J., Ni, J., Kraus, M., ... & Leippold, M. (2023). ChatClimate: Grounding conversational AI in climate science. *Communications Earth & Environment*, 4(1), 480.
- [36] Ni, J., Bingler, J., Colesanti Senni, C., Kraus, M., Gostlow, G., Schimanski, T., Stambach, D., Vaghefi, S. A., Wang, Q., Webersinke, N., Wekhof, T., Yu, T., & Leippold, M. (2023). ChatReport: Democratizing sustainability disclosure analysis through LLM-based tools. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 21-51).
- [37] Sautner, Z., van Lent, L., Vilkov, G., & Zhang, R. (2023). Firm-level climate change exposure. *The Journal of Finance*, 78(3), 1449-1498.
- [38] Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2022). ClimateBERT: A pretrained language model for climate-related text. arXiv. <https://doi.org/10.48550/arXiv.2110.12010>
- [39] Bingler, J. A., Kraus, M., Leippold, M., & Webersinke, N. (2022). Cheap talk and cherry-picking: What ClimateBERT has to say on corporate climate risk disclosures. *Finance Research Letters*, 47, 102776.

- [40] Schimanski, T., Reding, A., Reding, N., Bingler, J., Kraus, M., & Leippold, M. (2024). Bridging the gap in ESG measurement: Using NLP to quantify environmental, social, and governance communication. *Finance Research Letters*, 61, 104979.
- [41] Villacampa-Porta, J., Coronado-Vaca, M., & Garrido-Merchan, E. C. (2025). Impact of EU non-financial reporting regulation on Spanish companies' environmental disclosure: A cutting-edge natural language processing approach. *Environmental Sciences Europe*, 37, 29.
- [42] Delmas, M. A., & Burbano, V. C. (2011). The drivers of greenwashing. *California Management Review*, 54(1), 64-87.
- [43] Lyon, T. P., & Montgomery, A. W. (2015). The means and end of greenwash. *Organization & Environment*, 28(2), 223-249.
- [44] de Freitas Netto, S. V., Sobral, M. F. F., Ribeiro, A. R. B., & Soares, G. R. D. L. (2020). Concepts and forms of greenwashing: A systematic review. *Environmental Sciences Europe*, 32, 19.
- [45] Vinella, A., Capetz, M., Pattichis, R., Chance, C., & Ghosh, R. (2023). Leveraging language models to detect greenwashing. arXiv. <https://doi.org/10.48550/arXiv.2311.01469>
- [46] Bowen, F., & Aragon-Correa, J. A. (2014). Greenwashing in corporate environmentalism research and practice: The importance of what we say and do. *Organization & Environment*, 27(2), 107-112.
- [47] Berg, F., Kolbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, 26(6), 1315-1344.
- [48] Christensen, D. M., Serafeim, G., & Sikochi, A. (2022). Why is corporate virtue in the eye of the beholder? The case of ESG ratings. *The Accounting Review*, 97(1), 147-175.
- [49] Ramponi, A., & Plank, B. (2020). Neural unsupervised domain adaptation in NLP: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 6838-6855).
- [50] Mullainathan, S., & Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2), 87-106.
- [51] Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, 11, 685-725.
- [52] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215.
- [53] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv. <https://doi.org/10.48550/arXiv.1702.08608>
- [54] Jain, S., & Wallace, B. C. (2019). Attention is not explanation. In *Proceedings of NAACL-HLT 2019* (pp. 3543-3556).
- [55] Giglio, S., Kelly, B., & Stroebe, J. (2021). Climate finance. *Annual Review of Financial Economics*, 13, 15-36.
- [56] Buttigieg, C. P., & Pulis, S. (2024). Strategic initiatives to address greenwashing. *ERA Forum*, 25, 327-337.
- [57] Nieto, M. J., & Papathanassiou, C. (2024). Different shades of green: An analysis of the EU disclosure regulation and its effectiveness on limiting greenwashing. *EBI Working Paper Series No. 177*.
- [58] Bolton, P., & Kacperczyk, M. (2021). Do investors care about carbon risk? *Journal of Financial Economics*, 142(2), 517-549.
- [59] Bolton, P., & Kacperczyk, M. (2023). Global pricing of carbon-transition risk. *The Journal of Finance*, 78(6), 3677-3754.
- [60] Hsu, P.-H., Li, K., & Tsou, C.-Y. (2023). The pollution premium. *The Journal of Finance*, 78(3), 1343-1392.

- [61] Lee, J., Stevens, N., Han, S. C., & Song, M. (2024). A survey of large language models in finance (FinLLMs). arXiv. <https://doi.org/10.48550/arXiv.2402.02315>
- [62] Zhao, W. X., Zhou, K., Li, J., Tang, T., Dong, Z., Hou, Y., Zhang, B., Min, Y., Zhang, J., Liu, P., Wang, X., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., ... Wen, J.-R. (2026). A survey of large language models. *Frontiers of Computer Science*, 20, Article 2012627. <https://doi.org/10.1007/s11704-026-60308-3>
- [63] Stroebel, J., & Wurgler, J. (2021). What do you think about climate finance? *Journal of Financial Economics*, 142(2), 487-498.
- [64] Lipton, Z. C. (2018). The mythos of model interpretability. *Communications of the ACM*, 61(10), 36-43.
- [65] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 30 (pp. 4765-4774).
- [66] Teso, S., & Kersting, K. (2018). Why should I trust interactive learners? Explaining interactive queries of classifiers to users. arXiv. <https://doi.org/10.48550/arXiv.1805.08578>
- [67] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). <https://doi.org/10.48550/arXiv.1602.04938>

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Open Access

This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

