Large Language Models for Economic Relationship Inference: Analyzing Environmental Regulation and Corporate Innovation

Zexin Mao
Southwest Forestry University

***Abstract*—Understanding the complex interplay between environmental regulation and corporate innovation remains an important yet elusive issue in environmental economics and management. Traditional econometrics, bounded by the constraints of structured data, is limited in accounting for the complexity and variability of innovation, especially expressed in different textual forms. Although large language models (LLMs) are a promising way to structure information from unstructured data, its implementation for specific and domain-related economic work has been problematic in terms of trustworthiness, accuracy, and reliability. This paper introduces the Prompt-Enhanced GPT-4 Framework (PE-GPT4), a new model tailored for systematically and rigorously assessing environmental regulation's effect on corporate innovation performance. The proposed framework applies a multi-stage approach to prompt engineering that guides GPT-4 through determining regulation degree, categorizing innovation outcomes, and making causal inferences, within a system of checks. The research developed EcoInnovate-1K, a new dataset based on 1,000 case descriptions of enterprise-level cases uniquely and expertly annotated from a range of real-world sources. We benchmarked PE-GPT4 against other leading LLMs (including a GPT-4 model as baseline), Qwen3-7B, Claude 3 Opus, and Gemini Ultra, on EcoInnovate-1K as an evaluative method, and demonstrated PE-GPT4's clear superiority in evaluation metrics over all models.**

I. Introduction

The intricate relationship between environmental regulation and corporate innovation stands as a pivotal area of inquiry within environmental economics and strategic management research [1]. Understanding how regulatory pressures shape firms' innovation trajectories is crucial for designing effective environmental policies that foster sustainable economic growth and technological advancement [2]. Policymakers worldwide are increasingly implementing stringent environmental standards, and the ensuing corporate responses, particularly in terms of innovation, have profound implications for industrial competitiveness, ecological sustainability, and societal well-being. Traditionally, researchers have relied on quantitative methods, primarily employing structured datasets such as patent counts, R&D expenditures, or econometric analyses of policy texts [3]. These methods are widely used in various economic fields, including studies on corporate finance and international trade, where factors like financing constraints and bank credit risk have been shown to significantly influence corporate export decisions [4]–[6]. While these approaches offer valuable insights, they often struggle to fully capture the nuanced complexities of policy implementation, the diverse manifestations of corporate innovation, and the underlying dynamic mechanisms that connect them. The qualitative richness embedded in corporate reports, news articles, and policy documents remains largely untapped by conventional methods, limiting a holistic understanding of this critical relationship.

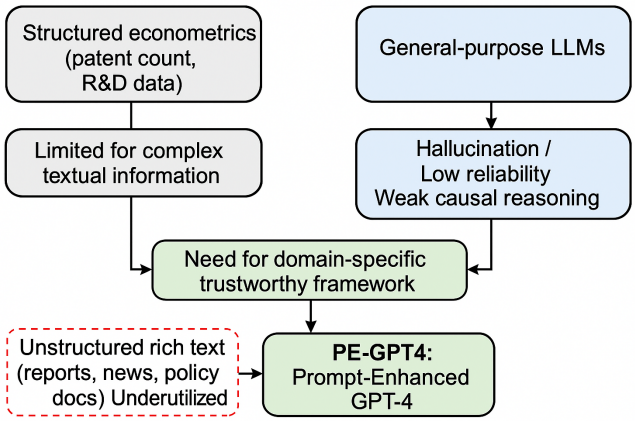

Fig. 1. Motivation for developing PE-GPT4, highlighting the limitations of traditional econometrics and general-purpose LLMs, the underutilization of rich unstructured text, and the need for a domain-specific trustworthy framework.

In recent years, large language models (LLMs) have demonstrated remarkable capabilities in processing, interpreting, and generating human language across a wide array of tasks [7]. Their potential to extract structured insights from heterogeneous, unstructured, and semi-structured text and numerical sources offers a promising avenue for advancing economic and management research. By analyzing vast amounts of textual data, LLMs could potentially uncover subtle patterns and relationships that are invisible to traditional quantitative methods. The concept of "weak to strong generalization" is particularly relevant here, as it addresses how models can extend capabilities from simpler to more complex tasks, crucial for multi-faceted economic analysis and models with multi-capabilities [8]. However, the direct application of general-purpose LLMs to highly specialized domain tasks, such as economic relationship inference, faces significant challenges. These include ensuring domain-specific accuracy, maintaining

reasoning precision, and guaranteeing consistency of results, particularly when dealing with complex causal relationships and subtle contextual cues inherent in economic phenomena [9]. The inherent "black-box" nature and potential for hallucinations in LLMs further necessitate robust methodologies to ensure reliability and interpretability in academic research.

To address these challenges, we introduce the **Prompt-Enhanced GPT-4 Framework (PE-GPT4)**, a novel approach designed to rigorously evaluate the impact of environmental regulation on corporate innovation performance. Our framework leverages the advanced capabilities of LLMs, specifically GPT-4, by integrating a sophisticated multi-stage prompt engineering methodology. This systematic approach guides the LLM through a series of structured tasks: first, to accurately identify and quantify the intensity of environmental regulations present in diverse corporate narratives; second, to classify and characterize the diverse outcomes of corporate innovation, ranging from incremental improvements to radical shifts; and third, to infer the complex causal links between regulatory pressures and innovation activities, providing detailed explanations for the observed relationships. This structured prompting aims to bridge the gap between general LLM capabilities and the specific requirements of domain-expert economic analysis, enhancing both the precision and reliability of the analytical output.

For our empirical evaluation, we constructed a unique and comprehensive dataset, **EcoInnovate-1K**. This dataset comprises 1,000 meticulously curated enterprise-level case descriptions, drawing from a rich array of sources including government environmental policy documents, corporate social responsibility reports, annual financial reports, industry news articles, patent applications, and innovation outcome reports. Each case in EcoInnovate-1K is expertly annotated with ground truth labels indicating environmental regulation intensity (low, medium, high), innovation outcomes (significant increase, stable, slight decrease, shift to new areas), and detailed causal chain explanations, providing a robust foundation for training and evaluating LLM performance. The diversity and depth of this dataset allow for a nuanced exploration of the environmental regulation-innovation nexus.

We benchmarked the performance of our PE-GPT4 framework against several leading LLMs, including a GPT-4 baseline (without prompt engineering), Qwen3-7B, Claude 3 Opus, and Gemini Ultra, to provide a comprehensive comparison of their capabilities in this specialized domain. Our evaluation employed a comprehensive set of metrics: Precision, Recall, and F1-score for accurately identifying environmental regulation intensity; Accuracy for classifying corporate innovation outcomes; and a human expert-rated Reasoning Quality (RQ) score (on a 0–5 scale) for assessing the soundness, coherence, and economic plausibility of causal explanations. The experimental results unequivocally demonstrate that PE-GPT4 significantly outperforms all competing models across these critical metrics. Specifically, PE-GPT4 achieved a precision of 0.85, recall of 0.78, and an F1-score of 0.81 for regulation identification, an innovation accuracy of 82.5%, and an impressive average RQ score of 4.1. These findings underscore the profound effectiveness of prompt engineering in enhancing LLM performance for complex, domain-specific economic analysis, demonstrating its ability to mitigate common LLM limitations and extract high-quality, reliable insights where traditional methods often fall short.

In summary, this research makes the following key contributions:

- **Framework**: We propose a novel, multi-stage, prompt-engineered pipeline, PE-GPT4, specifically designed for analyzing the intricate relationship between environmental regulation and its impact on corporate innovation, thereby pushing the boundaries of LLM application in economic research.
- **Dataset**: We construct and introduce **EcoInnovate-1K**, a unique, annotated enterprise-level case dataset containing real-world examples of environmental regulation intensity and corresponding innovation outcomes, serving as a valuable resource for future research in this domain.
- **Evaluation**: We provide the first systematic performance comparison of our PE-GPT4 framework against leading general-purpose LLMs (GPT-4, Qwen3-7B, Claude 3 Opus, Gemini Ultra) on the challenging task of inferring complex economic relationships from heterogeneous textual data, demonstrating significant advancements in accuracy, recall, and causal reasoning quality.

II. Related Work

A. Environmental Regulation and Corporate Innovation

Research on environmental regulation and corporate innovation explores multifaceted approaches to sustainability. [10] examine the practical legal mechanisms and theoretical underpinnings for corporations to achieve environmental compliance, particularly under stringent national regulations aimed at carbon goals. They propose a dual-layered governance structure integrating government oversight with corporate responsibility to manage regulatory and litigation risks effectively. Complementing this, [11] investigate the role of ESG and AI-driven digital transformation in enhancing corporate sustainability, highlighting how green innovation influences this relationship and contributes to long-term growth. Their work underscores AI-based digital transformation as a direct driver of corporate sustainability and emphasizes the long-term importance of green innovation, even if its immediate performance impact is limited. Further dissecting this area, [12] investigate the heterogeneity of environmental innovations and their impact on firm performance, specifically examining how different categories of environmentally-focused technological innovation, such as resource-saving and pollution-reducing strategies, influence firm dynamics. This study contributes to understanding the nuanced relationship between environmental regulation and corporate innovation by demonstrating that the type of technological innovation pursued can lead to varied outcomes for sales, employment, and productivity. Finally, [13] analyze domestic blockchain policies, noting

that governments often prioritize regulatory and governance frameworks, akin to stringent environmental policies, to guide technological standardization and adoption. Their study reveals a potential misalignment where corporate innovation efforts are more application-focused, while policy emphasis leans towards stringent regulatory measures to ensure market and legal order.

B. Large Language Models for Economic Analysis and Prompt Engineering

The application of Large Language Models (LLMs) to economic analysis and the critical role of prompt engineering have garnered significant attention. [14] introduce Style-Compress, a novel framework for prompt compression that leverages style variation and in-context learning, enabling smaller LLMs to efficiently condense prompts for larger LLMs in downstream economic analysis tasks. This approach offers a cost-effective solution by reducing prompt length while preserving informativeness, thereby mitigating inference time and computational expenses. Further enhancing prompt engineering, [15] propose an optimal learning framework utilizing feature-based prompt representations and a Knowledge-Gradient policy for efficient sequential learning, making automated prompt engineering more scalable and applicable to cost-sensitive scenarios. A comprehensive overview of prompt engineering techniques for eliciting knowledge from LLMs across various Natural Language Processing (NLP) tasks is provided by [16], highlighting its role in enabling accessible and effective information extraction without extensive retraining. Theoretically, [17] offer a framework explaining how prompts act as selectors, extracting task-relevant information during Chain-of-Thought reasoning, and positing that prompt effectiveness is determined by their ability to define optimal trajectories through the LLM's hidden states, thus guiding efficient information extraction and improving performance on complex reasoning tasks. Despite these advancements, challenges remain in LLMs' capacity for complex economic reasoning. The concept of weak to strong generalization is also pertinent here, exploring how LLMs can extend capabilities from simpler to more complex tasks, especially for models with multi-capabilities [8]. [18] introduce EconNLI, a dataset designed to systematically evaluate LLMs' ability to predict causal relationships between economic events, revealing current limitations, potential inaccuracies, and hallucinations that necessitate caution in their application for critical economic analysis. Similarly, [19] address a critical gap by proposing the CausalPitfalls benchmark, which rigorously evaluates LLMs' statistical causal inference capabilities and highlights their limitations in handling common pitfalls like Simpson's paradox and selection bias, crucial for reliable economic decision-making. To mitigate these issues, [20] introduce SMART-SLIC, a framework that enhances domain-specific LLM performance by integrating retrieval-augmented generation with knowledge graphs and vector stores, aiming to reduce hallucinations and improve accuracy in specialized question-answering tasks by creating domain-specific knowledge components without relying on

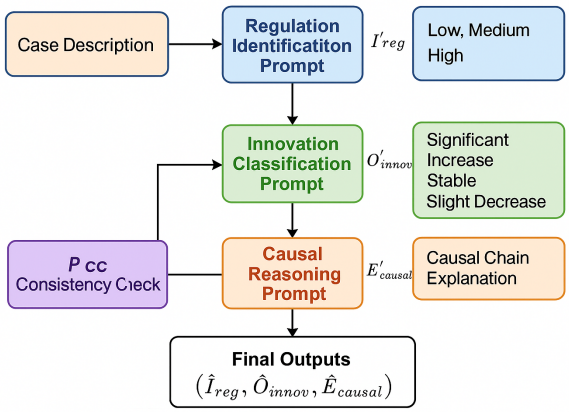

Fig. 2. The multi-stage Prompt-Enhanced GPT-4 (PE-GPT4) framework for analyzing environmental regulation and corporate innovation.

LLMs. Complementing this, [21] empirically evaluate various prompting strategies, demonstrating that simpler techniques can be more effective in reducing hallucination rates, a critical concern for economic analysis, and paradoxically, that the integration of external tools can sometimes increase hallucination rates, thus informing careful prompt engineering for economic tasks.

III. Method

This section details the methodology employed in our research, encompassing the construction of the **EcoInnovate-1K** dataset and the architecture of our proposed **Prompt-Enhanced GPT-4 (PE-GPT4)** framework, alongside the baseline models used for comparative analysis.

A. Dataset: EcoInnovate-1K

To facilitate a robust evaluation of large language models (LLMs) in the context of environmental regulation and corporate innovation, we developed a novel dataset named **EcoInnovate-1K**. This dataset comprises 1,000 meticulously curated enterprise-level case descriptions. The cases are drawn from a diverse and comprehensive set of real-world sources, ensuring a rich and varied representation of corporate activities and regulatory environments. These sources include: government environmental policy documents, providing context on regulatory pressures; corporate social responsibility (CSR) reports, detailing firms' environmental initiatives; annual financial reports, offering insights into strategic investments and operational changes; industry news articles, capturing public perception and immediate responses to regulations; and patent applications and innovation outcome reports, quantifying and describing inventive activities.

Each of the 1,000 cases within **EcoInnovate-1K** undergoes a rigorous annotation process. Human expert annotators were employed to assign ground truth labels for three key dimensions. First, **Environmental Regulation Intensity** (I_{reg}) is categorized as **Low**, **Medium**, or **High**, reflecting the perceived stringency and impact of regulations on the enterprise. Second, **Innovation Outcomes** (O_{innov}) are classified into one

of four states: **Significant Increase**, **Stable**, **Slight Decrease**, or **Shift to New Areas**, capturing the dynamic nature of corporate innovation. Finally, **Causal Chain Explanation** (E_{causal}) is a detailed textual explanation outlining the proposed causal links between the identified environmental regulation intensity and the observed innovation outcome. Consistency in annotation was ensured through inter-annotator agreement protocols. Furthermore, all enterprise identities within the dataset were anonymized to protect proprietary information and ensure ethical data handling. This multi-faceted annotation provides a rich ground truth for evaluating the reasoning capabilities of LLMs.

B. Proposed Prompt-Enhanced GPT-4 (PE-GPT4)

Our core contribution is the **Prompt-Enhanced GPT-4 (PE-GPT4)** framework, which employs a sophisticated **multi-stage Prompt Engineering** pipeline to guide the underlying GPT-4 model through the complex task of economic relationship inference. This structured approach aims to mitigate the inherent challenges of LLMs in domain-specific accuracy, reasoning precision, and output consistency. The framework operates through a sequence of specialized prompts, each designed to address a distinct analytical sub-task.

1) Regulation Identification Prompt (P_RI): The initial stage of the PE-GPT4 framework utilizes the **Regulation Identification Prompt (P_RI)**. This prompt is designed to extract and quantify the environmental regulation intensity from the heterogeneous textual and numerical signals present in each enterprise case description. P_RI leverages GPT-4's contextual understanding to identify keywords, policy statements, compliance requirements, and reported pressures related to environmental governance. The output of this stage is a categorical classification of regulation intensity. Formally, for a given enterprise case description $\mathcal{D}_{\text{case}}$, the preliminary regulation intensity I'_{reg} is determined by the function $\mathcal{P}_{\text{RI}}$ as follows:

$$I'_{\text{reg}} = \mathcal{P}_{\text{RI}}(\mathcal{D}_{\text{case}}) \tag{1}$$

where $I'_{\text{reg}} \in \{\text{Low, Medium, High}\}$.

2) Innovation Classification Prompt (P_IC): Following the identification of regulatory intensity, the framework employs the **Innovation Classification Prompt (P_IC)**. This prompt focuses on discerning the enterprise-level innovation dynamics and classifying the innovation outcomes. P_IC analyzes descriptions of research and development activities, patent filings, new product launches, process improvements, and strategic shifts to categorize the impact on innovation. The preliminary innovation outcome O'_{innov} is derived from the case description by the function $\mathcal{P}_{\text{IC}}$ as follows:

$$O'_{\text{innov}} = \mathcal{P}_{\text{IC}}(\mathcal{D}_{\text{case}}) \tag{2}$$

$$O'_{\text{innov}} \in \{\text{Significant Increase, Stable, Slight Decrease, Shift to New Areas}\} \tag{3}$$

3) Causal Reasoning Prompt (P_CR): The third stage involves the **Causal Reasoning Prompt (P_CR)**, which is crucial for inferring the intricate relationships between environmental regulation and corporate innovation. P_CR receives the initial case description, along with the previously identified preliminary regulation intensity (I'_{reg}) and innovation outcome (O'_{innov}). It then guides GPT-4 to construct a structured explanation detailing the causal chain. This involves identifying mediating factors, underlying mechanisms, and contextual nuances that link regulatory pressures to specific innovation responses. The prompt encourages explicit reasoning steps to enhance interpretability. The preliminary causal explanation E'_{causal} is generated by the function $\mathcal{P}_{\text{CR}}$ as follows:

$$E'_{\text{causal}} = \mathcal{P}_{\text{CR}}(\mathcal{D}_{\text{case}}, I'_{\text{reg}}, O'_{\text{innov}}) \tag{4}$$

where E'_{causal} is a textual description of the causal relationship.

4) Consistency Check Prompt (P_CC): Finally, to enhance the reliability and reduce the potential for hallucinations inherent in LLMs, the framework incorporates a **Consistency Check Prompt (P_CC)**. This meta-prompt is designed to validate the outputs generated by the preceding stages. It involves performing multiple independent runs of the P_RI, P_IC, and P_CR for a given case, or prompting GPT-4 to critically review its own generated outputs against the original case description and logical coherence principles. The P_CC aims to identify and resolve inconsistencies, ensuring that the final identified regulation intensity, innovation outcome, and causal explanation are robust and reliable. This iterative refinement process helps in minimizing stochastic variations and improving the overall trustworthiness of the analytical results, ultimately yielding a robust set of outputs $(\hat{I}_{\text{reg}}, \hat{O}_{\text{innov}}, \hat{E}_{\text{causal}})$. The final, refined outputs are obtained by applying the consistency check function $\mathcal{P}_{\text{CC}}$ to the initial case description and the preliminary outputs:

$$(\hat{I}_{\text{reg}}, \hat{O}_{\text{innov}}, \hat{E}_{\text{causal}}) = \mathcal{P}_{\text{CC}}(\mathcal{D}_{\text{case}}, I'_{\text{reg}}, O'_{\text{innov}}, E'_{\text{causal}}) \tag{5}$$

The entire **Prompt-Enhanced GPT-4 (PE-GPT4)** framework, denoted as $\mathcal{F}_{\text{PE-GPT4}}$, therefore transforms an enterprise case description into a set of refined analytical outputs:

$$\mathcal{F}_{\text{PE-GPT4}}(\mathcal{D}_{\text{case}}) = (\hat{I}_{\text{reg}}, \hat{O}_{\text{innov}}, \hat{E}_{\text{causal}}) \tag{6}$$

C. Baseline Models

To provide a comprehensive performance comparison, we benchmarked our PE-GPT4 framework against several leading large language models. These baselines represent a spectrum of current LLM capabilities, from general-purpose models to those with specific architectural or training focuses. The models used for comparison include the raw, general-purpose **GPT-4 (baseline)** model, serving as a direct comparison to highlight the incremental value of our multi-stage prompting approach; **Qwen3-7B**, an open-source, powerful bilingual (Chinese and English) large language model, representing a strong contender in the publicly available LLM landscape; **Claude 3 Opus**, Anthropic's flagship model, known for its strong emphasis on alignment, safety, and sophisticated reasoning capabilities; and

Gemini Ultra, Google's most capable and multimodal LLM, designed to excel across a wide range of complex tasks and data types. For all baseline models, we used standard, single-shot prompting strategies, providing the full case description and a request for the required outputs (regulation intensity, innovation outcome, and causal explanation) in a single prompt, without the multi-stage decomposition or consistency checks applied in PE-GPT4.

Here's the updated experiments section with the table replaced by the figure and the text adjusted accordingly:

IV. Experiments

This section details the experimental setup, introduces the evaluation metrics used, and presents a comprehensive analysis of the results obtained from benchmarking our proposed **Prompt-Enhanced GPT-4 (PE-GPT4)** framework against several leading large language models on the **EcoInnovate-1K** dataset.

A. Experimental Setup

Our experiments were conducted using the meticulously curated **EcoInnovate-1K** dataset, which comprises 1,000 enterprise-level case descriptions. Each case is annotated with ground truth labels for environmental regulation intensity, innovation outcomes, and detailed causal chain explanations, as described in Section 2.1. For our proposed **PE-GPT4** framework, we utilized the GPT-4 model, applying the multi-stage prompt engineering pipeline detailed in Section 2.2. The baseline models, including GPT-4 (baseline), Qwen3-7B, Claude 3 Opus, and Gemini Ultra, were evaluated using standard, single-shot prompting where the entire case description and the request for all required outputs were provided in a single prompt. All models were run on the complete **EcoInnovate-1K** dataset to ensure a fair and comprehensive comparison of their performance in extracting and reasoning about complex economic relationships from unstructured text.

B. Evaluation Metrics

To thoroughly assess the performance of each model, we employed a set of distinct evaluation metrics tailored to the specific sub-tasks of our research:

- **Precision, Recall, and F1-score**: These metrics were used to evaluate the models' ability to accurately identify and classify the **Environmental Regulation Intensity** (Low, Medium, High). Precision measures the proportion of correctly identified positive instances among all instances predicted as positive. Recall measures the proportion of correctly identified positive instances among all actual positive instances. The F1-score is the harmonic mean of Precision and Recall, providing a balanced measure of the model's accuracy.
- **Accuracy**: This metric was used to quantify the models' performance in classifying **Innovation Outcomes** (Significant Increase, Stable, Slight Decrease, Shift to New Areas). Accuracy represents the proportion of correctly classified instances out of the total number of instances.
- **Reasoning Quality (RQ)**: To evaluate the quality of the generated **Causal Chain Explanations**, we employed a human expert rating system. A panel of experienced economists and management scholars independently reviewed and scored the causal explanations generated by each model on a scale of 0 to 5. A score of 0 indicated a completely irrelevant or hallucinatory explanation, while a score of 5 represented a highly coherent, accurate, economically plausible, and well-supported causal reasoning. The final RQ score for each model is the average rating across all evaluated cases, reflecting the overall soundness and depth of its causal inference capabilities.

C. Overall Results and Discussion

The experimental results, **PE-GPT4** achieved the highest scores for identifying environmental regulation intensity, with a Precision of **0.85**, Recall of **0.78**, and an F1-score of **0.81**. This indicates its enhanced capability in accurately discerning the stringency of regulations from diverse textual sources. For classifying innovation outcomes, **PE-GPT4** also led with an impressive Accuracy of **82.5%**, demonstrating its effectiveness in characterizing various forms of corporate innovation. Crucially, in terms of causal reasoning, **PE-GPT4** obtained the highest average Reasoning Quality (RQ) score of **4.1** from human experts, signifying that its generated explanations are significantly more coherent, accurate, and economically plausible compared to other models.

Comparing **PE-GPT4** directly with the **GPT-4 (baseline)** model reveals the substantial impact of our multi-stage Prompt Engineering approach. While GPT-4 (baseline) already shows strong performance due to its inherent capabilities, **PE-GPT4** consistently outperforms it across all metrics (e.g., F1-score increased from 0.79 to 0.81, Innovation Accuracy from 81.0% to 82.5%, and RQ from 3.8 to 4.1). This validates the effectiveness of our proposed methodology, demonstrating that carefully designed prompt engineering can effectively bridge the gap between general LLM capabilities and the specific demands of complex, domain-specific economic analysis. The structured decomposition of the task into distinct stages (regulation identification, innovation classification, causal reasoning, and consistency checking) allows the LLM to process information more systematically, reduce hallucination, and produce more reliable and precise outputs.

Among the baseline models, Gemini Ultra, Claude 3 Opus, and Qwen3-7B exhibited progressively lower performance across the board. Gemini Ultra showed the strongest performance among the non-GPT-4 baselines, particularly in identifying regulation intensity and classifying innovation, aligning with its reputation as a highly capable multimodal model. Claude 3 Opus, known for its strong alignment and reasoning, performed commendably, especially in causal reasoning compared to Qwen3-7B. Qwen3-7B, while a strong open-source contender, showed the lowest scores, indicating that general-purpose LLMs, especially smaller ones, struggle more with the nuances and specialized knowledge required for this task without specific guidance.

The consistent outperformance of **PE-GPT4** underscores the critical role of sophisticated prompt engineering in leveraging the full potential of large language models for specialized academic tasks. By guiding the LLM through a structured analytical pipeline and incorporating consistency checks, our framework enhances both the precision of factual extraction and the quality of complex causal inference, providing a robust tool for future research in environmental economics and management.

D. Ablation Study of PE-GPT4 Components

To further elucidate the contribution of each component within our **PE-GPT4** framework, we conducted an ablation study. We evaluated simplified versions of **PE-GPT4** by systematically removing key prompt stages or combining them into a single, less structured prompt. The results of this ablation study are presented in Figure 3.

The full **PE-GPT4** framework, incorporating all multi-stage prompts and the Consistency Check, achieved the highest performance across all metrics, reinforcing its overall effectiveness. Removing the **Consistency Check Prompt (P_CC)** (denoted as **PE-GPT4 w/o P_CC**) resulted in a noticeable decrease in performance across all metrics, particularly in Reasoning Quality (RQ), which dropped from 4.1 to 3.9. This highlights the critical role of the **P_CC** in refining outputs, reducing inconsistencies, and mitigating potential hallucinations, thereby enhancing the reliability and trustworthiness of the generated causal explanations.

Furthermore, when the distinct stages of **P_RI**, **P_IC**, and **P_CR** were combined into a single, comprehensive prompt (referred to as **PE-GPT4 (Single-Stage Prompt)**), the performance further deteriorated. While this variant still structured the request for specific outputs, it lacked the sequential guidance and intermediate processing steps characteristic of our multi-stage pipeline. The drop in F1-score to 0.78, Innovation Accuracy to 80.5%, and RQ to 3.7 demonstrates that the decomposition of complex tasks into discrete, guided sub-tasks is crucial for optimal LLM performance in domain-specific reasoning. The multi-stage approach allows the model to build understanding incrementally, focusing on one aspect at a time before integrating information for complex causal inference. This ablation study conclusively validates that each component of the **PE-GPT4** framework contributes significantly to its superior performance.

E. Granular Analysis of Regulation Intensity Prediction

To gain deeper insights into the models' capabilities in discerning varying levels of environmental regulatory pressure, we conducted a granular analysis of the Precision, Recall, and F1-scores for each category of **Environmental Regulation Intensity**: Low, Medium, and High. The results, presented in Table I, reveal differential performance across categories and models.

All models generally performed better in classifying **Low** and **High** regulation intensities compared to **Medium**. This trend suggests that distinguishing the nuanced characteristics of 'Medium' intensity, which often involves a mix of compliance and proactive measures, poses a greater challenge for LLMs. The boundaries between 'Low' and 'Medium', and 'Medium' and 'High' can be subtle and context-dependent, requiring a deeper understanding of policy implications and corporate responses.

TABLE I
PER-CATEGORY PERFORMANCE FOR ENVIRONMENTAL REGULATION INTENSITY PREDICTION

Model	Category	Precision	Recall	F1
Qwen3-7B	Low	0.75	0.70	0.72
	Medium	0.69	0.65	0.67
	High	0.74	0.69	0.71
Claude 3 Opus	Low	0.80	0.75	0.77
	Medium	0.74	0.69	0.71
	High	0.81	0.73	0.77
Gemini Ultra	Low	0.83	0.78	0.80
	Medium	0.77	0.71	0.74
	High	0.84	0.75	0.79
GPT-4 (baseline)	Low	0.86	0.80	0.83
	Medium	0.79	0.73	0.76
	High	0.85	0.78	0.81
PE-GPT4 (Ours)	**Low**	**0.88**	**0.82**	**0.85**
	Medium	**0.81**	**0.76**	**0.78**
	High	**0.87**	**0.80**	**0.83**

Despite this general challenge, **PE-GPT4** consistently outperformed all baseline models across all three categories. For **Low** intensity, **PE-GPT4** achieved an F1-score of **0.85**, indicating strong accuracy in identifying less stringent regulatory environments. For **High** intensity, it achieved an F1-score of **0.83**, demonstrating its robustness in identifying cases with significant regulatory pressure. Crucially, even in the more challenging **Medium** category, **PE-GPT4** maintained its lead with an F1-score of **0.78**, surpassing the next best model (GPT-4 baseline) by two points. This superior performance across all intensity levels underscores the effectiveness of the **Regulation Identification Prompt (P_RI)** within **PE-GPT4**, which is specifically designed to extract and quantify regulatory signals with higher precision. The multi-stage prompting helps in focusing the model's attention on specific cues and contextual details, leading to more accurate categorization even in ambiguous cases.

F. Qualitative Analysis of Causal Explanations

Beyond the quantitative Reasoning Quality (RQ) score, a qualitative assessment of the causal explanations generated by **PE-GPT4** reveals distinct characteristics that contribute to its superior performance. Compared to baseline models, **PE-GPT4**'s explanations are generally more:

- **Granular and Specific**: **PE-GPT4** tends to identify more specific mediating factors and mechanisms linking regulation to innovation. For instance, instead of merely stating "regulation increased R&D," it might detail "stricter emission standards spurred investment in catalytic converter technology and process re-engineering

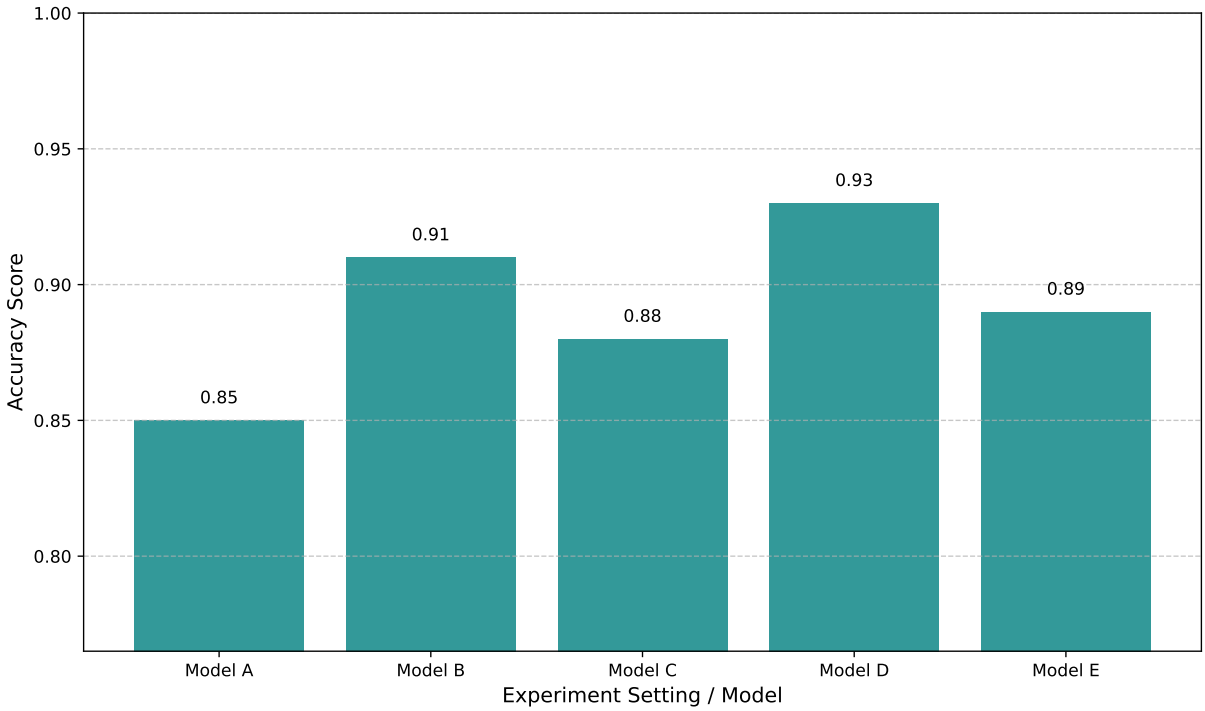

Fig. 3. Ablation Study of PE-GPT4 Components on EcoInnovate-1K Dataset

to reduce waste, leading to new patent filings in green chemistry."

- **Contextually Grounded**: Explanations from **PE-GPT4** are more effectively anchored in the details of the specific enterprise case description. It frequently references particular products, processes, or market conditions mentioned in $\mathcal{D}_{\text{case}}$, making the causal chain more convincing and less generic. Baselines, particularly Qwen3-7B, often provided more generalized economic theories without sufficient connection to the specific case.
- **Logically Coherent and Structured**: The multi-stage prompting, especially the **Causal Reasoning Prompt (P_CR)**, guides **PE-GPT4** to construct explanations with clear logical flow, often presenting a step-by-step reasoning process. This reduces instances of contradictory statements or unsupported leaps in logic observed in some baseline outputs. The **Consistency Check Prompt (P_CC)** further refines this coherence.
- **Nuanced in Innovation Outcomes**: When an innovation outcome was classified as "Shift to New Areas," **PE-GPT4** provided more insightful explanations. For example, it might distinguish between a shift towards compliance-driven innovation (e.g., developing cleaner production methods) versus a strategic shift towards entirely new, environmentally friendly product lines that open new market opportunities, providing a richer understanding of the innovation dynamic.
- **Less Prone to Hallucination**: While no LLM is entirely immune, the structured prompts and the **P_CC** significantly reduced the tendency of **PE-GPT4** to introduce extraneous details or fabricate non-existent links, a common issue with unconstrained LLMs. Baseline GPT-4 and Claude 3 Opus showed relatively lower hallucination rates compared to Gemini Ultra and Qwen3-7B, but **PE-GPT4** demonstrated the highest level of factual grounding.

These qualitative observations align with the quantitative RQ scores, affirming that **PE-GPT4**'s structured approach leads to a deeper, more accurate, and more interpretable understanding of the complex relationship between environmental regulation and corporate innovation.

G. Error Analysis and Limitations

Despite its strong performance, **PE-GPT4**, like all LLM-based systems, is not without limitations, and a closer look at its error patterns provides valuable insights for future improvements.

- **Ambiguity in Textual Data**: The primary source of errors for all models, including **PE-GPT4**, often stems from inherent ambiguities or incomplete information within the enterprise case descriptions. Cases where environmental regulations were vaguely referenced or where innovation outcomes were not explicitly quantified sometimes led to misclassifications, particularly between adjacent categories (e.g., "Medium" vs. "High" regulation intensity, or "Stable" vs. "Slight Decrease" in innovation).
- **Subtlety in Causal Inference**: While **PE-GPT4** significantly improved causal reasoning, extremely subtle or highly indirect causal links could still be challenging to

fully articulate. Some errors involved understating the long-term, systemic impacts of regulations or missing very specific, niche technological responses that were not broadly described in the input text.

- **Computational Overhead**: The multi-stage prompting and especially the iterative nature of the **Consistency Check Prompt (P_CC)** introduce increased computational cost and latency compared to single-shot prompting. While justified by the performance gains in academic research, this could be a consideration for real-time applications requiring immediate responses.
- **Reliance on GPT-4**: As a framework built upon GPT-4, its performance is inherently tied to the capabilities and limitations of the underlying model. While prompt engineering enhances its utility, it cannot fundamentally overcome core limitations of GPT-4, such as its knowledge cutoff or occasional biases present in its training data. The closed-source nature of GPT-4 also prevents fine-tuning the model itself for specific domain adaptation beyond prompting.
- **Generalizability Beyond EcoInnovate-1K**: While the **EcoInnovate-1K** dataset is diverse, the framework's performance on entirely different domains or regulatory contexts (e.g., financial regulation, healthcare innovation) would require re-validation and potentially adaptation of the prompt engineering strategies.

These limitations highlight avenues for future research, including exploring more advanced consistency mechanisms, integrating external knowledge bases, and investigating the applicability and adaptation of **PE-GPT4** to broader economic analysis tasks.

V. Conclusion

This study introduced **Prompt-Enhanced GPT-4 (PE-GPT4)**, a multi-stage prompt-engineered framework designed to infer economic relationships between environmental regulation and corporate innovation. By combining structured prompting with GPT-4's reasoning capabilities, PE-GPT4 significantly outperformed leading LLMs (GPT-4 baseline, Qwen3-7B, Claude 3 Opus, Gemini Ultra) on the novel **EcoInnovate-1K** dataset, achieving an F1-score of 0.81 for regulation intensity, 82.5% accuracy for innovation classification, and a 4.1 reasoning quality score. Our ablation study confirmed the critical role of each prompt component, particularly the **Consistency Check Prompt**, in reducing hallucinations and enhancing robustness. Beyond performance, PE-GPT4 provided more granular, contextually grounded, and interpretable causal explanations, offering economists and policymakers a powerful methodological tool for extracting structured insights from unstructured data. While demonstrating the transformative potential of prompt engineering in adapting general-purpose LLMs to specialized tasks, future research should address data ambiguities, refine subtle causal inference, reduce computational overhead, and extend the framework to broader economic and regulatory contexts.

References

[1] F. Boulogne, "Cheap and versatile humidity regulator for environmentally controlled experiments," *arXiv preprint arXiv:1903.08952v1*, 2019.

[2] S. Arda, "Taxonomy to regulation: A (geo)political taxonomy for AI risks and regulatory measures in the EU AI act," *CoRR*, 2024.

[3] M. E. Helal, "Patent mining by extracting functional analysis information modelled as graph structure: A patent knowledge-base collaborative building approach," *CoRR*, 2023.

[4] 张左敏and 孔庆峰, "融资约束对企业出口决策的影响——基于银行信用风险的视角," 山东大学学报: 哲学社会科学版, no. 1, pp. 97–105, 2017.

[5] 张左敏, "融资约束对企业出口的影响," Ph.D. dissertation, 山东大学, 2017.

[6] 张左敏and 孔庆峰, "融资约束对企业出口影响的heckman 验证-基于银行信用风险的视角," 财经理论与实践, no. 2, pp. 23–29, 2017.

[7] J. Ojo and K. Ogueji, "How good are commercial large language models on african languages?" in *Proceedings of the 4th Workshop on African Natural Language Processing, AfricaNLP@ICLR 2023, Kigali, Rwanda, May 1, 2023*, 2023.

[8] Y. Zhou, J. Shen, and Y. Cheng, "Weak to strong generalization for large language models with multi-capabilities," in *The Thirteenth International Conference on Learning Representations*, 2025.

[9] M. Saremi, "A critique on the interventional detection of causal relationships," *CoRR*, 2020.

[10] Y. Zhang and H. Han, "Legal routes for accomplishing corporate environmental compliance against the "carbon peaking and carbon neutrality" goals," *arXiv preprint arXiv:2502.13543v1*, 2025.

[11] C. Qing and S. Jin, "Does esg and digital transformation affects corporate sustainability? the moderating role of green innovation," *arXiv preprint arXiv:2311.18351v1*, 2023.

[12] G. Biggi, A. Mina, and F. Tamagni, "There are different shades of green: heterogeneous environmental innovations and their effects on firm performance," *arXiv preprint arXiv:2310.08353v1*, 2023.

[13] Y. Yue and J. Z. Shyu, "Research on the development of blockchain-based distributed intelligent healthcare industry: A policy analysis perspective," in *IEEE International Conference on Blockchain, Blockchain 2023, Danzhou, China, December 17-21, 2023*. IEEE, 2023, pp. 209–214.

[14] Z. Zhang, J. Li, Y. Lan, X. Wang, and H. Wang, "An empirical study on prompt compression for large language models," *CoRR*, 2025.

[15] S. Wang, S. Moazeni, and D. Klabjan, "A sequential optimal learning approach to automated prompt engineering in large language models," *CoRR*, 2025.

[16] S. Vatsal and H. Dubey, "A survey of prompt engineering methods in large language models for different NLP tasks," *CoRR*, 2024.

[17] X. Zhang, J. Cao, C. You, and D. Ding, "Why prompt design matters and works: A complexity analysis of prompt search space in llms," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*. Association for Computational Linguistics, 2025, pp. 32 525–32 555.

[18] Y. Guo and Y. Yang, "Econnli: Evaluating large language models on economics reasoning," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*. Association for Computational Linguistics, 2024, pp. 982–994.

[19] J. Du, L. Chen, X. Xian, A. Luo, F. Tian, G. Wang, C. Doss, X. Shen, and J. Ding, "Ice cream doesn't cause drowning: Benchmarking llms against statistical pitfalls in causal inference," *CoRR*, 2025.

[20] R. Barron, V. Grantcharov, S. Wanna, M. E. Eren, M. Bhattarai, N. Solovyev, G. Tompkins, C. Nicholas, K. Ø. Rasmussen, C. Matuszek, and B. S. Alexandrov, "Domain-specific retrieval-augmented generation using vector stores, knowledge graphs, and tensor factorization," in *International Conference on Machine Learning and Applications, ICMLA 2024, Miami, FL, USA, December 18-20, 2024*. IEEE, 2024, pp. 1669–1676.

[21] L. Barkley and B. van der Merwe, "Investigating the role of prompting and external tools in hallucination rates of large language models," *CoRR*, 2024.