CYUT at SemEval-2025 Task 6: Prompting with Precision – ESG Analysis via Structured Prompts

Shih-Hung Wu, Zhi-Hong Lin, Ping-Hsuan Lee

Department of Computer Science and Information Engineering, Chaoyang University of Technology, Wufeng, Taichung, Taiwan

shwu@cyut.edu.tw, s11327609@gm.cyut.edu.tw,
s11327603@gm.cyut.edu.tw

Abstract

In response to the increasing demand for efficient ESG verification, we introduce a novel natural language processing (NLP) framework designed to automate the assessment of corporate sustainability claims. This approach combines Retrieval-Augmented Generation (RAG), Chain-of-Thought (CoT) reasoning, and structured prompt engineering to accurately process and categorize a wide range of multilingual ESG disclosures. In the SemEval-2025 PromiseEval competition, our system achieved a score of 0.5611—ranking 4th on the private English leaderboard—and a score of 0.5747—securing 1st place on the private French leaderboard. These results represent substantial improvements over traditional machine learning methods and underscore the framework's potential as a scalable, transparent, and robust solution for ESG evaluation in corporate settings.

1 Introduction

The concept of Environmental, Social, and Governance (ESG) sustainability has emerged as a critical framework for assessing corporate responsibility and long-term viability. As concerns over climate change, social inequality, and governance practices continue to escalate, corporations are increasingly required to demonstrate measurable commitments. However, evaluating these commitments presents significant challenges. Traditional assessment methods heavily rely on manual reviews of corporate reports, third-party evaluations, and media sources—approaches that are labor-intensive, costly, difficult to scale, and often inconsistent across regions and languages.

To tackle these challenges, our team participated in the SemEval-2025 Task 6: PromiseEval—Multinational, Multilingual, Multi-Industry Promise Verification competition(Chen et al., 2025). This competition introduces a novel multilingual dataset encompassing English, French, Chinese, Japanese, and Korean, designed to assess corporate commitments and their fulfillment in the ESG domain. The primary objective is to develop NLP methodologies that automate corporate promise verification by identifying commitments, evaluating supporting evidence, assessing clarity, and inferring appropriate verification timelines.

Advancements in Natural Language Processing (NLP) have demonstrated immense potential in automating the evaluation of large-scale textual data. Early NLP techniques, including sentiment analysis, topic modeling, and named entity recognition, have been widely applied to extract structured insights from ESG disclosures. Nevertheless, these methods remain constrained by rule-based systems, which struggle to adapt to dynamic and diverse ESG datasets. Transformer-based models(Vaswani et al., 2023), such as BERT(Devlin et al., 2019) and GPT(OpenAI et al., 2024), have revolutionized the field through context-aware text analysis, enhancing the scalability and robustness of NLP applications.(Chung and Latifi, 2024) evaluated ESG-specific pre-trained Large Language Models (LLMs), such as FinBERT-ESG and fine-tuned LLaMA models, demonstrating their superior performance over traditional machine learning techniques like SVM and XGBoost in ESG text classification tasks. These models excel at capturing semantic and contextual nuances within ESG-related texts, making them particularly well-suited for analyzing abstract concepts and complex interrelations.

Despite these advancements, challenges remain in applying NLP techniques to ESG evaluation. ESG data originate from diverse formats, sources, and languages, necessitating sophisticated approaches capable of integrating both structured and unstructured information.(Peng et al., 2024) propose an advanced methodology for processing unstructured ESG data, addressing challenges in text extraction, multilingual content, and diverse

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 494–501
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

document formats to improve the accuracy of ESG assessments. Additionally, many ESG indicators—such as descriptions of social responsibility initiatives or governance strategies—are inherently qualitative, requiring models to not only extract data but also comprehend and reason about complex relationships.(Sokolov et al., 2021) highlight the difficulties in automating ESG scoring using NLP, particularly in handling qualitative ESG factors that require contextual reasoning.

To address these limitations, this study integrates state-of-the-art NLP techniques, including Retrieval-Augmented Generation (RAG)(Gao et al., 2024), Chain-of-Thought (CoT)(Yu et al., 2023),(Wei et al., 2023) reasoning, and Prompt Engineering(Sahoo et al., 2024),(Vatsal and Dubey, 2024), to enhance the automation of ESG commitment verification. The Structured Prompt framework systematically guides the model through a multi-stage reasoning process using explicit definitions, clarification rules, concrete examples, constrained label outputs, and stepwise instructions. This design enables the model to accurately comprehend classification standards and make consistent decisions across diverse contexts. By leveraging these methods, our framework provides a multilingual, efficient, and scalable solution that significantly narrows the gap between corporate commitments and measurable outcomes, while enhancing interpretability and reliability in ESG evaluations.

The ClimateBERT model fine-tuned by (Vinella et al., 2024) demonstrated an accuracy of 86.34% in assessing greenwashing risks within corporate sustainability reports, underscoring the promising capabilities of language models in the domain of greenwashing detection.

By applying cutting-edge NLP methodologies to ESG evaluation and testing them within the ML-Promise challenge framework, we aim to advance more transparent and reliable corporate ESG oversight mechanisms, ultimately fostering sustainable development practices.

2 Dataset

The dataset used in this study is derived from SemEval-2025 Task 6 (Chen et al., 2025), which focuses on verifying corporate commitments disclosed in Environmental, Social, and Governance (ESG) reports. It comprises textual data from multiple companies, annotated with structured labels designed to support the identification and evaluation of corporate pledges and their supporting evidence, thereby facilitating effective ESG statement verification through Natural Language Processing (NLP) models.

Annotations are organized across four principal dimensions: (1) Promise Status, indicating whether a clear commitment has been made ("Yes" or "No"); (2) Verification Timeline, classifying the expected timeframe for fulfillment as "Already," "Less than 2 years," "2 to 5 years," "More than 5 years," or "N/A"; (3) Evidence Status, reflecting the existence of verifiable documentation ("Yes" or "No"); and (4) Evidence Quality, evaluating the clarity and credibility of supporting evidence as "Clear," "Not Clear," "Misleading," or "N/A."

Statistical analysis of the dataset reveals that most corporate statements contain explicit commitments, yet many lack short-term fulfillment targets, potentially undermining their credibility. While documentation often supports pledges, inconsistencies remain due to unverifiable claims and variable evidence quality. Practical challenges such as spelling errors, linguistic variation, and unstructured text also impact multilingual model performance. Nevertheless, the dataset offers a strong structural foundation for ESG commitment verification and highlights directions for future improvements in bias mitigation and model robustness.

3 Methodology

3.1 Structured Prompt Design for ESG Classification

To improve the accuracy, consistency, and interpretability of large language models (LLMs) in ESG-related classification tasks, this study proposes a Structured Prompting approach tailored specifically for corporate sustainability analysis. Traditional methods—such as keyword matching or rule-based classification—often suffer from limitations in handling context, ambiguity, and domain-specific interpretation. To address these issues, we designed a modular structured prompt architecture that guides the model through a multi-step reasoning process, mimicking human annotation logic.

The Structured Prompt comprises five synergistic components:

- **Definition**: Establishes explicit criteria for what constitutes a valid ESG commitment, filtering out vague or aspirational language.

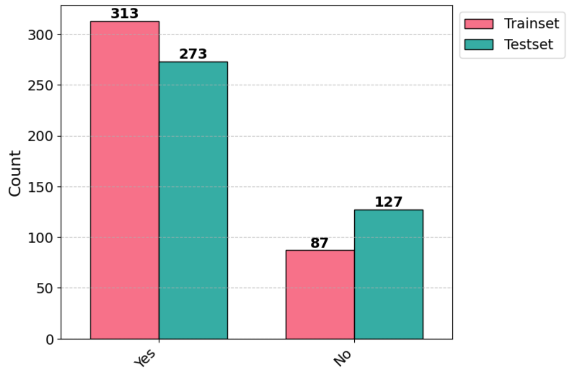

Figure 1: Promise Status Distribution

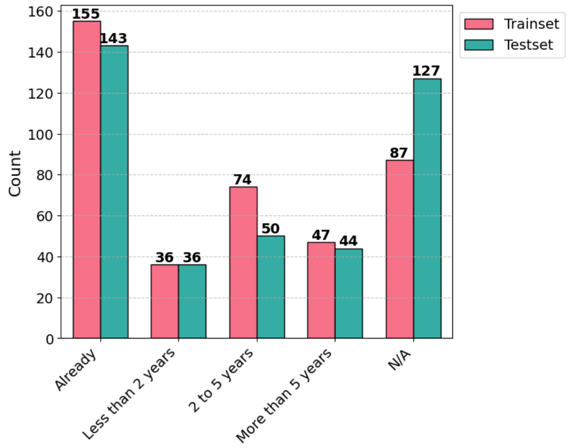

Figure 2: Verification Timeline Distribution

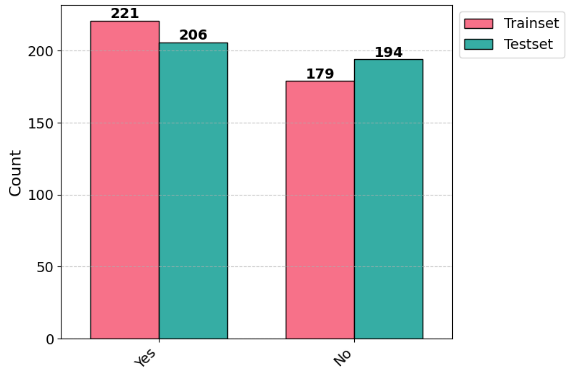

Figure 3: Evidence Status Distribution

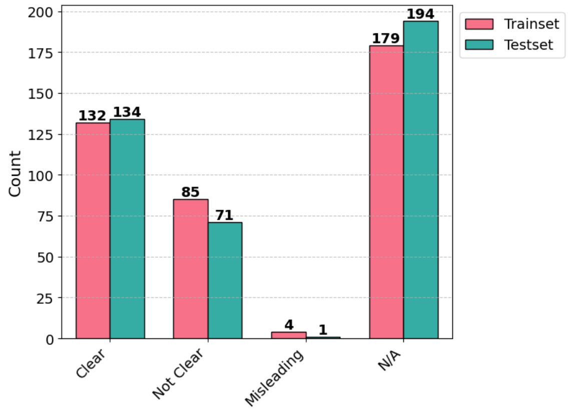

Figure 4: Evidence Quality Distribution

- **Clarification**: Provides further elaboration on borderline cases, helping reduce overclassification by emphasizing semantic precision.
- **Example**: Supplies positive and negative illustrations to operationalize the abstract classification principles and anchor model interpretation.
- **Labels**: Standardizes output to binary classifications (e.g., {"promise_status": "Yes"}), enabling structured evaluation and automation.
- **Instructions**: Enforces conservative reasoning under uncertainty and reinforces task-specific constraints (e.g., ignoring irrelevant corporate statements).

This design was not arbitrary but emerged from iterative testing on noisy, multilingual ESG disclosures. Early prompt variants often led to inconsistent predictions, especially when faced with vague language or complex governance terminology. By integrating structured prompting with logical Chain-of-Thought (CoT), Self-Consistency, and Tree-of-Thought (ToT) mechanisms, the model is prompted to evaluate ESG statements in a context-aware, sequential manner.

Additionally, more advanced prompting strategies such as System 2 Attention and Graph-of-Thoughts (GoT) are optionally applied to encourage deliberate, multi-domain reasoning, particularly in cases involving cross-sectional ESG categories.

This modular yet principled design enables LLMs to simulate human annotation logic at scale, ensuring interpretability, robustness, and alignment with ESG classification standards. The performance gains observed through ablation studies (Section 5.4) further validate the contribution of each prompt component to overall model effectiveness.

3.2 Advantages of the Structured Prompt Approach

To enhance both the accuracy and consistency of ESG-related text classification, we developed a structured prompting framework comprising six interlocking steps. These steps guide large language models (LLMs) to perform multi-dimensional classification based on explicit standards. Among

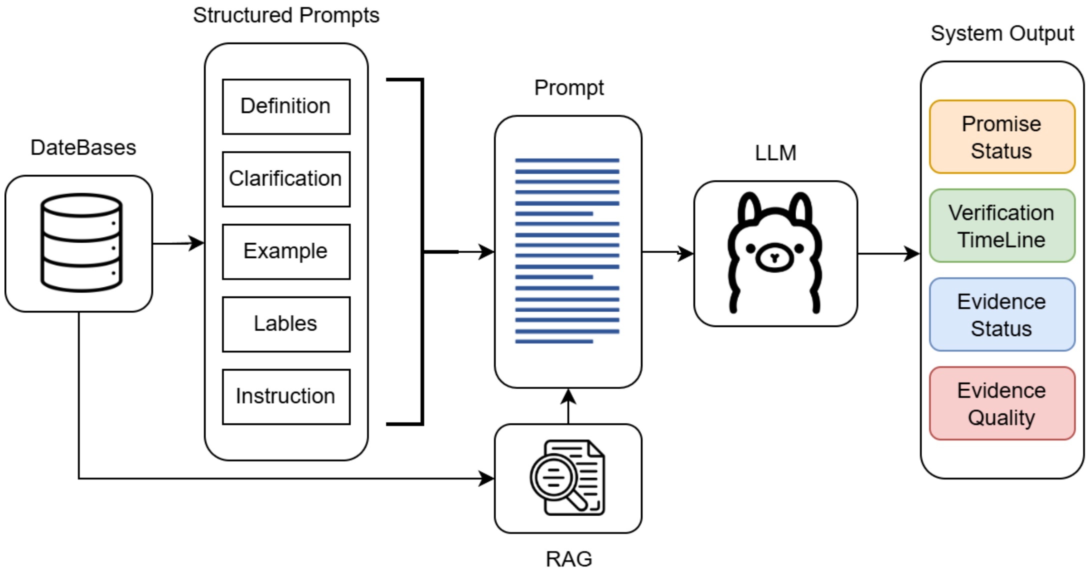

Figure 5: Structured Prompts with Retrieval-Augmented Generation Workflow

them, three core techniques—Definition/Schema-Priming Prompting, Chain-of-Thought (CoT) Prompting, and Few-shot Prompting—serve as the foundation for semantic precision and reasoning reliability(e.g. in Table 1).

To ensure output consistency and machine readability, structured-output prompting was also integrated. The model was instructed to return results in a standardized JSON schema, facilitating downstream processing. Additionally, conditional prompting was used to enforce logical constraints between fields (e.g., if Commitment = No, then Timeline and Evidence Quality must be marked as "N/A").

Collectively, this structured prompting framework provides a reproducible and interpretable mechanism for guiding LLMs in ESG classification tasks. It not only articulates what the model should judge, but also dictates how it should reason and in what format the results should be conveyed—thereby offering a concrete blueprint for future high-consistency, scalable semantic classification applications.

4 Experimental setup

To rigorously evaluate the effectiveness of our proposed ESG verification framework, we conducted experiments focusing on computational efficiency, model accuracy, and multilingual generalization. This section outlines the technical framework, hardware configuration, retrieval methodology, preprocessing techniques, and evaluation strategy.

4.1 Core Framework and Hardware Configuration

Our system is built on Ollama, a lightweight yet powerful framework optimized for large-scale NLP applications. It runs on a high-performance setup featuring an Intel i7-12900K processor, an RTX 3090 Ti GPU, and 32GB of RAM, enabling efficient ESG text processing and fast document retrieval with reasonable computational costs.

4.2 RAG

We employ FAISS (Douze et al., 2025) for scalable nearest-neighbor search, enabling rapid retrieval of relevant ESG documents. For embedding generation, we use Multilingual-E5 (Wang et al., 2024), a transformer-based model designed for cross-lingual tasks. The integration of FAISS and RAG ensures efficient retrieval of semantically relevant ESG statements, enhancing verification accuracy.

4.3 Base Model and Preprocessing

The Base Model serves as the foundation for ESG commitment verification. During preprocessing, defaultdict is utilized to optimize data structure handling, improving the speed and accuracy of classification.

4.4 Experimental Evaluation

Experiments were conducted to assess different RAG configurations and prompt structures, aiming to identify the most effective setup for ESG-related tasks. The focus was on analyzing how retrieval strategies and prompt engineering impact key performance metrics.

4.5 Evaluation Metrics

We adopt standard classification metrics—Accuracy, Recall, Precision, and F1-score—to evaluate model performance, with Macro F1-score reported unless otherwise specified. For different ESG subtasks (Promise Status, Verification Timeline, Evidence Status, Evidence Quality), we select metrics tailored to task-specific requirements to ensure comprehensive performance analysis.

5 Results

5.1 Model and RAG Quantity Analysis

In this study, we evaluated different model configurations for the ESG commitment classification task by varying the number of Retrieval-Augmented Generation (RAG) instances. The RAG quantity refers to the number of top-relevant documents retrieved during inference—for example, RAG-3 retrieves the three most similar results to support the model's reasoning. An appropriate number of retrieved documents can significantly enhance prediction accuracy.

We systematically compared the F1-scores across various model and RAG configurations, as shown in Table 2. The results indicate that model performance varies considerably depending on the RAG setting, underscoring the importance of tuning retrieval parameters(e.g. in Table 2).

Additionally, a comparative analysis was performed between models with and without RAG to assess the necessity and effectiveness of RAG in improving performance (e.g. in Table 3).

5.2 Detailed Evaluation of the Optimal Model

Once the best-performing model and the optimal RAG quantity were identified, further analysis was conducted to evaluate specific task components, including promise status, verification timeline, evidence status, and evidence quality. The performance of each of these aspects was recorded and analyzed in detail (e.g. in Table 4).

5.3 Comparative Experiments Using CoT vs. Not Using CoT

To investigate the impact of Chain-of-Thought (CoT) on reasoning tasks, we conducted comparative experiments for "with CoT" and "without CoT" across four subtasks: Promise Status, Verification Time, Evidence Status, and Evidence Quality. In all experiments, we used Retrieval-Augmented Generation (RAG) with a retrieval count of 6, and the unified base model was Llama 3.1 (70B) (Grattafiori et al., 2024) ,(Touvron et al., 2023). (e.g. Table 5) summarizes the Accuracy performance for both configurations across the four subtasks:

The results show that, for Promise Status and Evidence Status, the model using CoT achieves noticeably higher Accuracy than the one without CoT. Meanwhile, for Verification Time and Evidence Quality, which require deeper reasoning, the CoT-based model also significantly outperforms the non-CoT setting, demonstrating CoT's advantages in multi-step reasoning scenarios. Since all experiments in this study fixed RAG at 6 and utilized Llama 3.1 (70B) as the base model, future adjustments to the retrieval count or strategy may further affect performance on different subtasks and could serve as a reference for subsequent prompt engineering and model optimization.

5.4 Ablation Study on Prompt Engineering

To assess the impact of our prompt engineering strategies, we conducted an ablation study by systematically removing different structural components of the prompt. The performance variations observed across different configurations provided insights into the contribution of each prompt component to the overall system effectiveness in (e.g. in Table 6).

Overall, the results demonstrate the effectiveness of our approach in optimizing the ESG promise task, highlighting the importance of both RAG and structured prompt engineering in achieving high performance.

6 Conclusion

This study presents an advanced NLP-driven framework for ESG commitment verification, addressing the limitations of traditional assessment methods. By leveraging Retrieval-Augmented Generation, Chain-of-Thought reasoning, and structured prompt engineering, our approach enhances the automation, accuracy, and interpretability of ESG evaluations. The experimental results from our participation in the SemEval-2025 PromiseEval task validate the effectiveness of our model, demonstrating its superior performance in classifying corporate commitments, verifying evidence, and assessing the credibility of ESG-related claims. Future research could explore further improvements in multilingual adaptability and the integration of external knowledge sources to enhance contextual understanding. Ultimately, our methodology contributes to the development of scalable, reliable, and transparent ESG verification systems, supporting global efforts in corporate sustainability assessment.

Acknowledgments

This study was supported by the National Science and Technology Council under the grant number NSTC 113-2221-E-324-009.

References

Chung-Chi Chen, Yohei Seki, Hakusen Shu, Anaïs Lhuissier, Juyeon Kang, Hanwool Lee, Min-Yuh Day, and Hiroya Takamura. 2025. SemEval-2025 task 6: Multinational, multilingual, multi-industry promise verification. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Tin Yuet Chung and Majid Latifi. 2024. Evaluating the performance of state-of-the-art esg domain-specific pre-trained large language models in text classification against existing models and traditional machine learning techniques.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and et al. 2024. The llama 3 herd of models.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, and et al. 2024. Gpt-4 technical report.

Jiahui Peng, Jing Gao, Xin Tong, Jing Guo, Hang Yang, Jianchuan Qi, Ruiqiao Li, Nan Li, and Ming Xu. 2024. Advanced unstructured data processing for esg reports: A methodology for structured transformation and enhanced analysis.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications.

Alik Sokolov, Jonathan Mostovoy, Jack Ding, and Luis A. Seco. 2021. Building machine learning systems for automated esg scoring. In *The Journal of Impact and ESG Investing*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Shubham Vatsal and Harsh Dubey. 2024. A survey of prompt engineering methods in large language models for different nlp tasks.

Avalon Vinella, Margaret Capetz, Rebecca Pattichis, Christina Chance, Reshmi Ghosh, and Kai-Wei Chang. 2024. Leveraging language models to detect greenwashing.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. 2023. Towards better chain-of-thought prompting strategies: A survey.

Appendices A–F

A Structured Prompting Pipeline for ESG Text Classification

B Evaluation of Model Performance Across Multiple RAG Settings

C Performance Comparison of ESG Verification Models: Baseline vs. RAG Variants

D Performance Metrics of the Optimal Model in ESG Verification

E Effectiveness of CoT Reasoning in ESG Verification

F of Structured Prompt Components on ESG Classification Labels

Step	Prompt Engineering Method	Description	Example / Use Case
1	Instruction-based Prompting	Use a clear sentence or paragraph to explicitly tell the model what to do, establishing a clear objective.	Start with a prompt like: *“Objective: Systematically assess… ”*
2	Definition / Schema-Priming Prompting	Define key terms or schema to ensure consistent and accurate understanding.	*“A promise must… ”*; define all key labels.
3	Chain-of-Thought (CoT) Prompting	Guide the model through logical steps such as “Step 1… Step n” to encourage multi-step reasoning.	*“1. Promise Status → 2. Verification Timeline → 3… ”* style step guidance.
4	Few-shot / Demonstration Prompting	Provide 1 to k real examples to help the model learn the output pattern and reduce bias.	Each item includes Example: *“Yes: … ”, “No: … ”*.
5	Structured-Output Prompting / JSON-schema Prompting	Specify the output format, such as JSON or table, to enhance structure and consistency.	Final section defines the output format, e.g., JSON schema.
6 (Optional)	Conditional / Guardrail Prompting	Set conditional rules (e.g., if–then statements) to handle exceptions and enforce constraints.	Rules like: *“If Promise = No, then Timeline = N/A”*.

Table 1: Structured Prompting Pipeline for ESG Text Classification

Model	RAG-1	RAG-2	RAG-3	RAG-4	RAG-5	RAG-6
`llama3.1:8b`	0.5623	**0.5770**	0.5630	0.5633	0.5494	0.5477
`llama3.1:70b`	0.5456	0.5676	0.5571	0.5707	**0.5893**	0.5769
`llama3.2:3b`	0.4748	**0.4905**	0.4578	0.4708	0.4889	0.4741
`phi4:14b`	**0.5555**	0.5443	0.5347	0.5295	0.5456	0.5384
`qwq:32b-prev`	0.5444	0.5401	0.5499	0.5279	0.5188	**0.5450**

Table 2: Evaluation of Model Performance Across Multiple RAG Settings

	With RAG	W/o RAG
Promise Status	**0.7956**	0.6629
Verification Timeline	**0.5083**	0.4442
Evidence Status	0.6975	**0.7224**
Evidence Quality	0.3800	**0.3918**

Table 3: Performance Comparison of ESG Verification Models: Baseline vs. RAG Variants

	Accuracy	Recall	Precision	F1 Score
Promise Status	0.8100	0.8100	0.7956	0.8154
Verification Timeline	0.6075	0.6083	0.5083	0.5915
Evidence Status	0.7000	0.7009	0.6975	0.6985
Evidence Quality	0.5800	0.5800	0.3561	0.5551

Table 4: Performance Metrics of the Optimal Model in ESG Verification

	With CoT	W/o CoT
Promise Status	**0.7956**	0.7150
Verification Timeline	**0.5083**	0.3300
Evidence Status	**0.6975**	0.5900
Evidence Quality	**0.3800**	0.3350

Table 5: Effectiveness of CoT Reasoning in ESG Verification

	Definition	Clarification	Example	Labels	Instructions
Promise Status	0.7706	0.7597	**0.7708**	0.7429	0.7523
Verification Timeline	**0.5011**	0.4615	0.4902	0.4898	0.4513
Evidence Status	0.6984	0.7090	0.7042	0.7010	**0.7325**
Evidence Quality	0.3246	**0.3777**	0.3506	0.3391	0.3378

Table 6: Impact of Structured Prompt Components on ESG Classification Labels