VALIDATING ESG COMMITMENTS WITH RAG-ENHANCED LARGE LANGUAGE MODELS: TOWARD TRANSPARENT AND RELIABLE SUSTAINABILITY DISCLOSURE

Hsin-Ting Lu

Min-Yuh Day*

Graduate Institute of Information Management, National Taipei University, New Taipei City, Taiwan
myday@gm.ntpu.edu.tw*

Keywords: Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), ESG Reports, ESG Commitment Validation, Greenwashing Detection

OUTLINE

1. Introduction

- Research Background
- Research Motivation
- Research Objective

1.1 Research Background

Business Perspective

- Corporate sustainability reports have become an important channel for companies to communicate their commitments to Environmental, Social, and Governance (ESG) principles (Xu, Miao, Xiao, & Lin, 2025).

AI Perspective

- In response to the complexity of ESG disclosures and the surge in information volume, researchers have started leveraging natural language processing (NLP) techniques to improve the accuracy and efficiency of ESG data extraction and analysis (Xu, Miao, Xiao, & Lin, 2025).

1.2 Research Motivation

- Many ESG commitments in corporate sustainability reports are **vague**, **unverifiable**, or **selectively disclosed**, raising concerns about greenwashing.

- As ESG disclosure becomes central to corporate governance and investor trust, ensuring transparency and verifiability is increasingly critical.

1.3 Research Objective

Main Objective

- To develop an ESG commitment verification framework that integrates **Retrieval-Augmented Generation (RAG)** with **Large Language Models (LLMs)** of different scales to enhance classification and reasoning accuracy,evaluated on the **ML-Promise French subset (~400 samples) (Seki et al., 2024)**.

Research Question

- **RQ1:** Can RAG significantly improve LLM performance in ESG promise verification tasks compared with non-RAG baselines?
- **RQ2:** Do RAG-enhanced LLMs show different performance across the four subtasks (Promise Identification, Supporting Evidence Assessment, Evidence Quality, and Timing for Verification)?
- **RQ3:** How does model scale affect the effectiveness of RAG in ESG verification, and can smaller models benefit from retrieval to close the gap with larger models?

2. Literature Review

- ESG Reporting and the Challenge of Greenwashing
- Large Language Models: Capabilities and Scalability
- Retrieval-Augmented Generation for Knowledge-Intensive Tasks
- The ML-Promise Dataset for Multilingual ESG Commitment Verification

2.1 ESG Reporting and the Challenge of Greenwashing

Importance of ESG Reports

- Serve as a key reference for assessing corporate performance across **Environmental, Social, and Governance (ESG)** dimensions.
- Act as a crucial channel for firms to communicate commitments and demonstrate accountability.

The Emergence of Greenwashing

- Greenwashing occurs when companies selectively highlight positive ESG data to attract stakeholders, deliberately **hiding negative environmental impacts**.

Recent Detection Studies

- Introduced the **A3CG dataset** as a novel benchmark for robust ESG analysis under greenwashing contexts (Ong et al., 2025).
- Fine-tuned the **ClimateBERT model** to improve accuracy in identifying misleading disclosures (Vinella et al., 2023).

2.2 Large Language Models: Capabilities and Scalability

- Achieved remarkable performance in NLP tasks, especially text generation (Xie et al., 2024).
- Large models offer superior complex reasoning.

- Trade-offs in Model Scale:
  - Large Models: Offer superior performance, but face extremely high computational and financial costs.
  - Small/Medium Models: More efficient and easier to deploy, but their performance is often limited.

- Challenges:
  - **Hallucination**: Factual errors undermining reliability (Lin et al., 2025).
  - **High Cost**: Prohibitive deployment cost, limiting accessibility.

2.3 Retrieval-Augmented Generation for Knowledge-Intensive Tasks

- LLMs face limitations in **hallucinations** and knowledge access (lacking current/domain-specific data) (Wallat et al., 2025).

- RAG Solution and Benefits:
  - RAG employs a hybrid architecture coupling a **Retriever** with a **Generator** (Zhang et al., 2025).
  - Benefits: Improves factuality and interpretability.
  - Performance: RAG strategies significantly enhance model performance, leading to steady gains in **complex reasoning and knowledge-intensive tasks** (Li et al., 2025; Krishna et al., 2024).

- Applications:
  - Widely used in **open-domain QA**, **multi-hop reasoning**, and **specialized text analytics** (e.g., clinical trial data analysis (Zheng et al., 2025), legal document processing).

2.4 The ML-Promise Dataset for Multilingual ESG Commitment Verification

Dataset Overview (Seki et al., 2024):

- Scale: Approx. 3,010 samples collected from ESG reports across five countries.
- Languages Covered: Includes English, French, Chinese, Japanese, and Korean.
- Core Goal: To address challenges in evaluating corporate sustainability commitments, especially in response to Greenwashing.

Verification Tasks (Seki et al., 2024):

- Promise Identification
- Supporting Evidence Assessment
- Evidence Quality
- Timing for Verification

3. Research Methodology

- System Architecture
- Dataset
- Model Selection
- Retrieval Corpus and Indexing
- Evaluation Metric

3.1 System Architecture

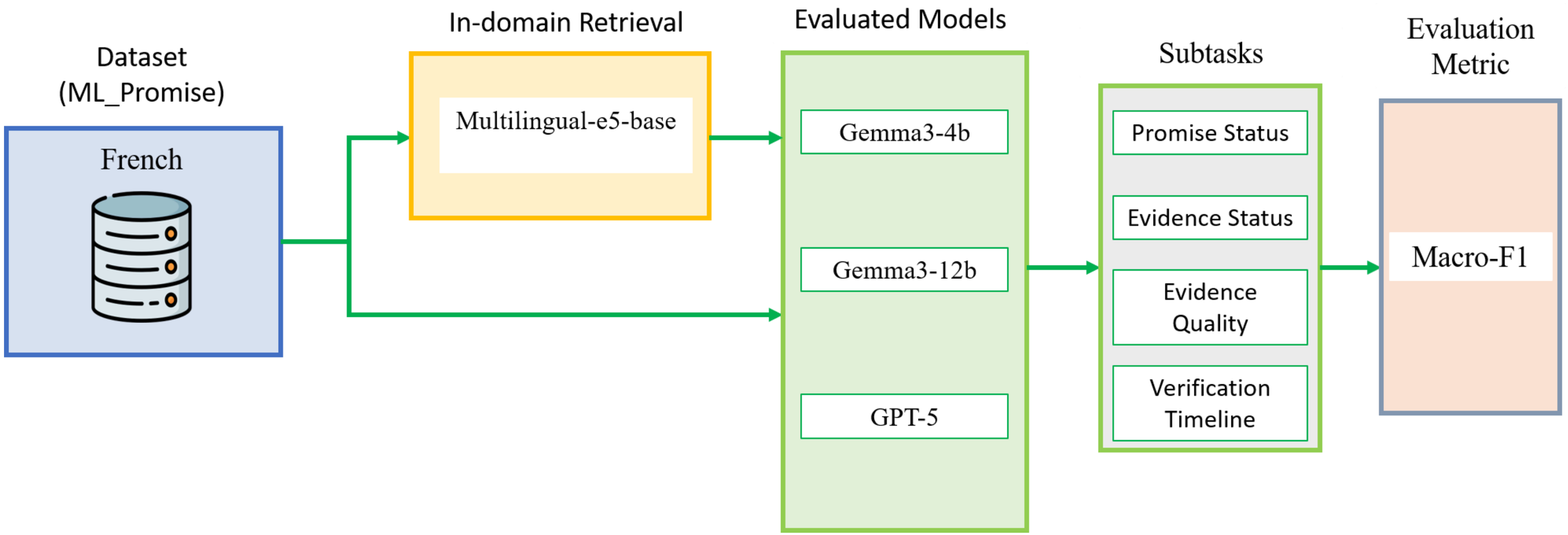

Figure 1. Proposed research workflow for ESG promise verification

Source: This study

3.2 Dataset

- Source: ML-Promise
- Language: French
- Sample Size: N = 400

Subtasks:

- Promise Status: whether a concrete or organization-level commitment is present (Yes / No).
- Evidence Status: whether verifiable supporting evidence is provided (Yes / No).
- Evidence Quality: clarity of the evidence (Clear, Not Clear, Misleading, N/A).
- Verification Timeline: expected timeframe for fulfilling the commitment (Already, Less than 2 years, 2 to 5 years, More than 5 years, N/A).

3.2 Dataset

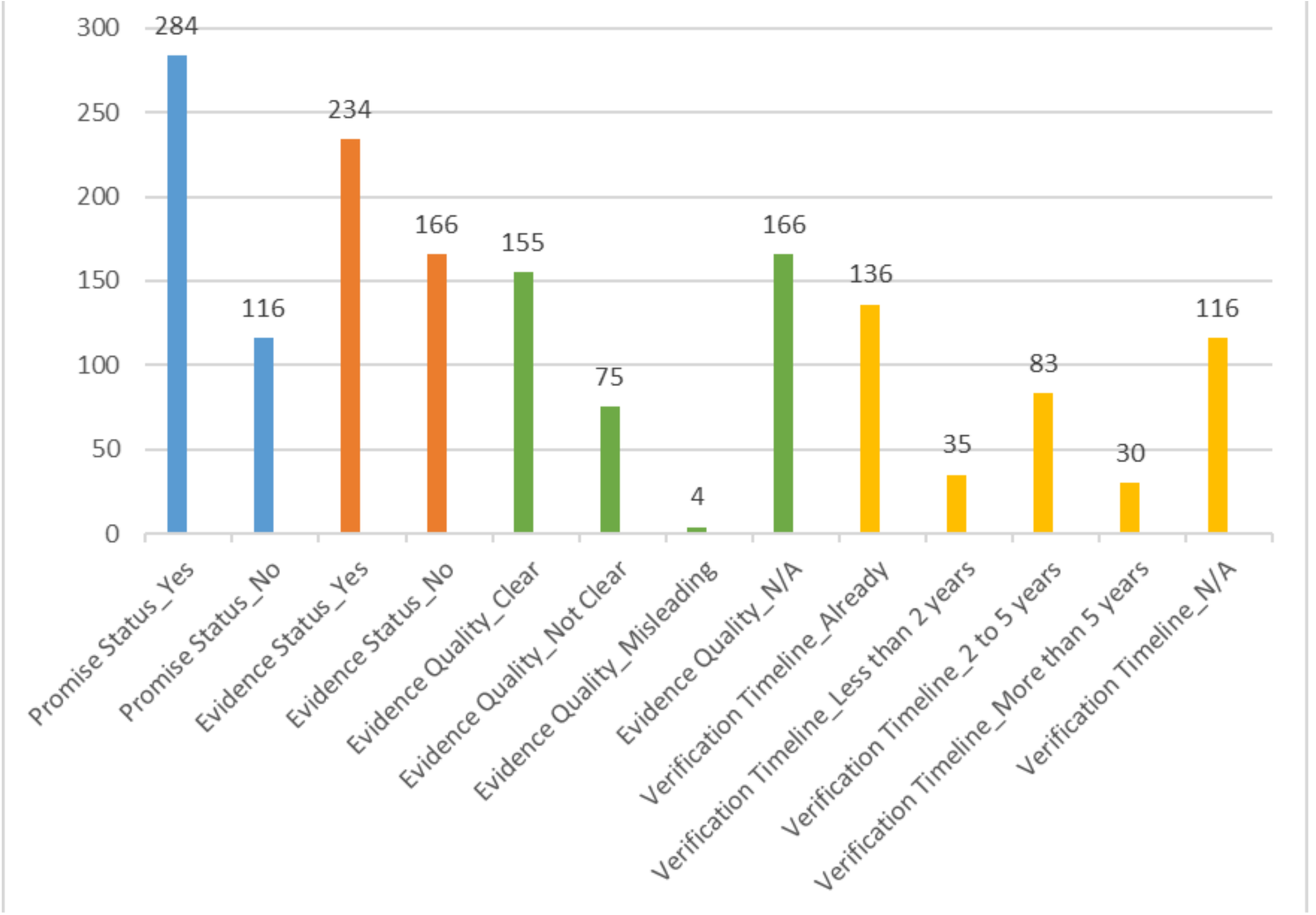

Figure 2. Label distribution of the French test set (n = 400), which is used for evaluation in this study

Source: This study

3.3 Model Selection

This study evaluated three language models spanning small, medium, and large scales:

- Small Scale: Gemma 3: 4B
- Medium Scale: Gemma 3: 12B
- Large Scale: GPT-5

Purpose of Selection:

- Systematically examine how **Retrieval-Augmented Generation** interacts with different model scales.
- Investigate whether retrieval techniques can help **small** and **medium** models narrow the performance gap with the large model under comparable Macro-F1 evaluation.

3.4 Retrieval Corpus and Indexing

Retrieval Corpus:

- Source: the **French training** split of the ML-Promise dataset
- Tools: Use **multilingual-e5-base model**, and used to construct a **FAISS index**.

Inference Process:

- The system retrieves the **top-6** most relevant passages from the index for each test instance.
- The retrieved top-6 content is appended to the **LLM prompt** to serve as contextual evidence.

3.5 Evaluation Metric

- **Metric Chosen:**
  - Adopted the **Macro-Averaged F1 Score (Macro-F1)**.

- **Advantages:**
  - Ensures equal importance for both **majority** and **minority** classes.
  - Compared with Accuracy, which tends to be biased toward majority classes, Macro-F1 provides a **fairer** and **more reliable** assessment of classification and reasoning performance.

4. Experiment Results and Analysis

- Overall Results with Baseline
- Subtask-Level Performance Analysis
- Effect of Model Scale

4.1 Overall Results with Baseline

RAG Setting	Task	Gemma3-4B	Gemma3-12B	GPT-5	ML_Promise French Dataset
W/O RAG	Promise Identification	0.509	0.734	0.687	0.816
	Supporting Evidence	0.573	0.528	0.787	0.746
	Evidence Quality	0.238	0.269	0.365	0.443
	Verification Timeline	0.211	0.422	0.418	0.523
W/ RAG	Promise Identification	0.625	0.754	0.756	0.798
	Supporting Evidence	0.523	0.666	0.749	0.732
	Evidence Quality	0.285	0.330	0.419	0.487
	Verification Timeline	0.301	0.411	0.420	0.601

Table 1. Overall Experimental Results on French ESG Promise Verification (Macro-F1), with Comparisons to ML-Promise Baseline

Source: This study

4.2 Subtask-Level Performance Analysis

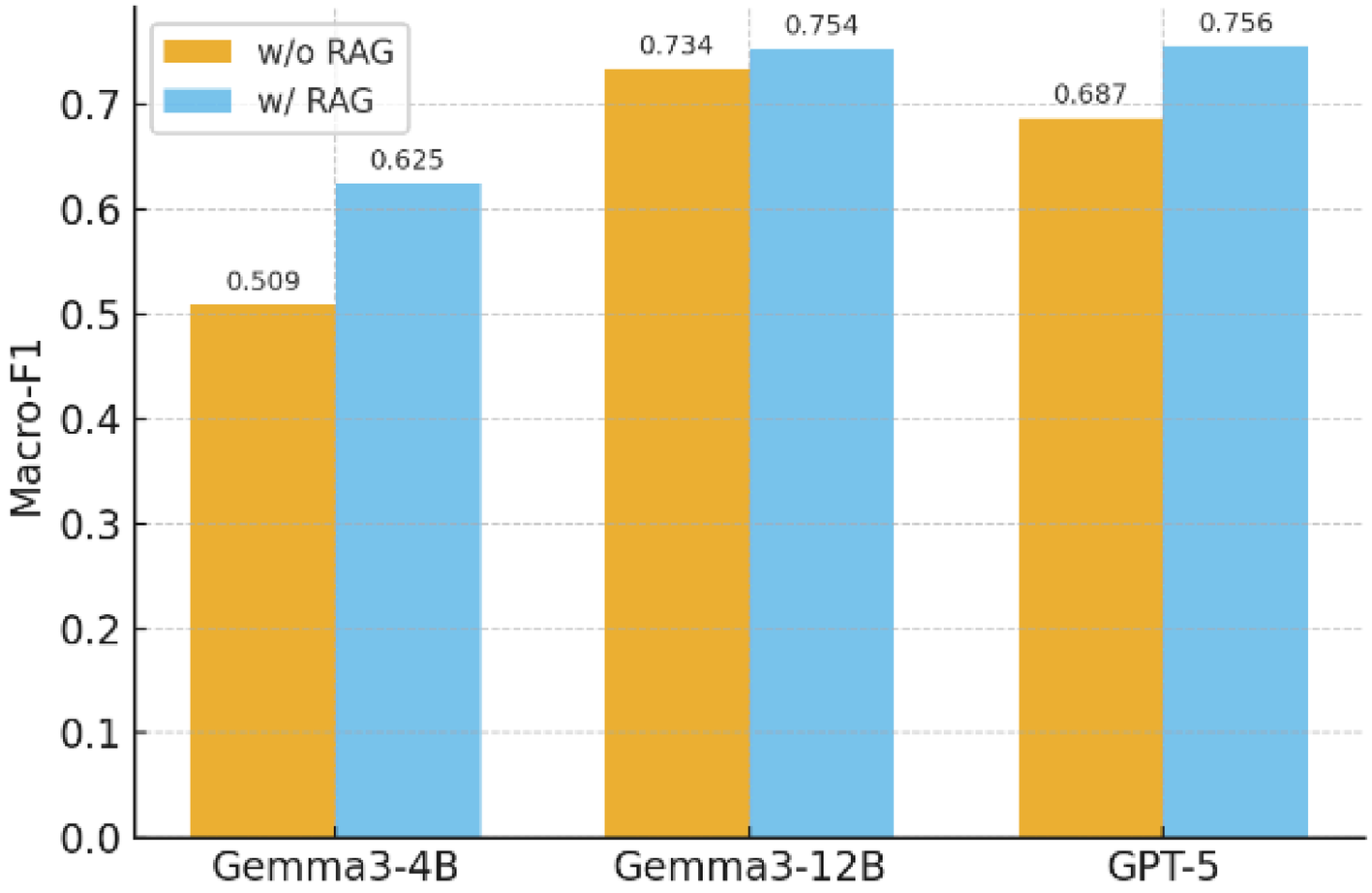

Figure 3. Subtask-level performance on Promise Identification (w/ vs. w/o RAG across models).

Source: This study

4.2 Subtask-Level Performance Analysis

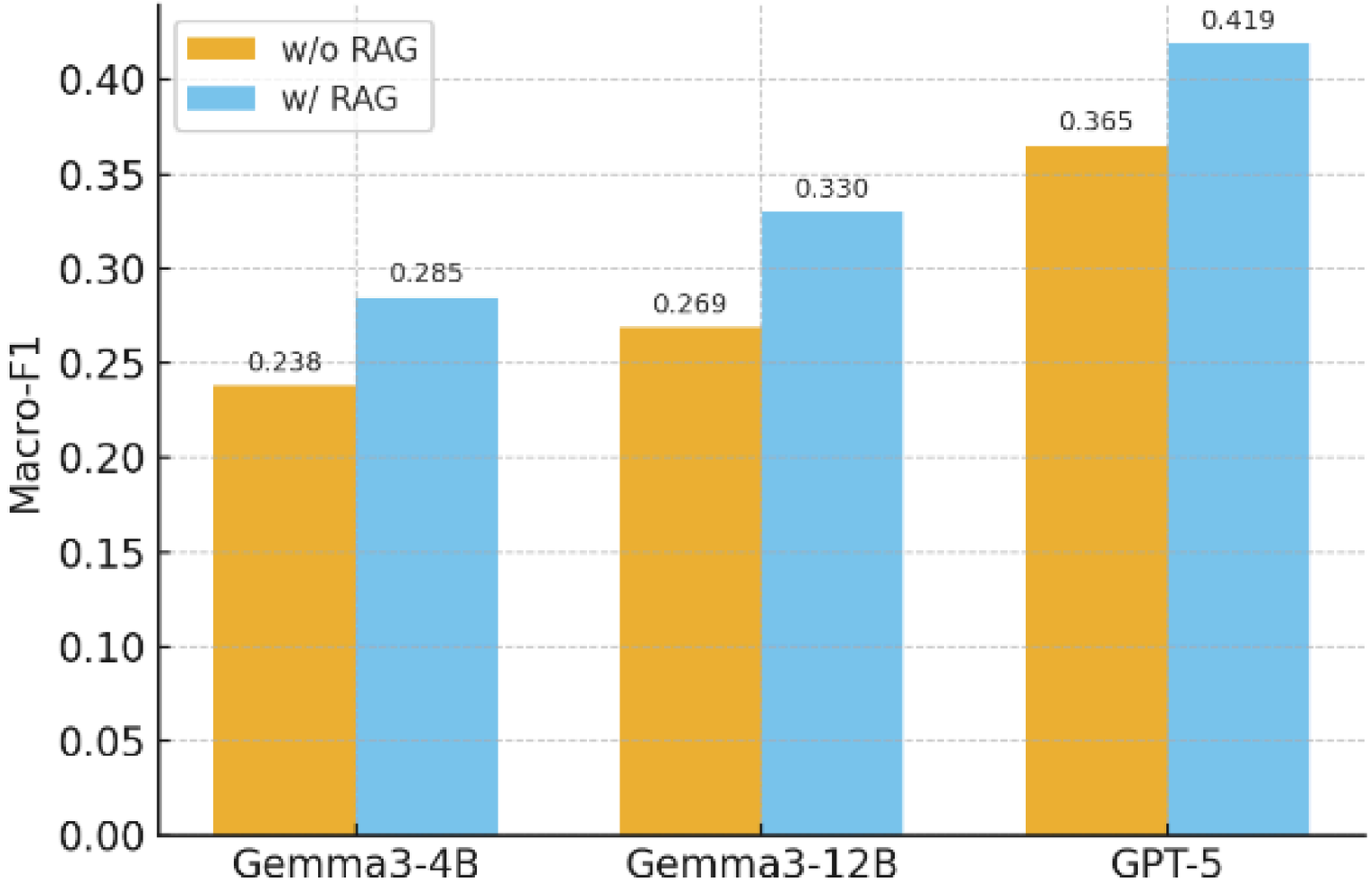

Figure 4. Subtask-level performance on Supporting Evidence Assessment (w/ vs. w/o RAG across models).

Source: This study

4.2 Subtask-Level Performance Analysis

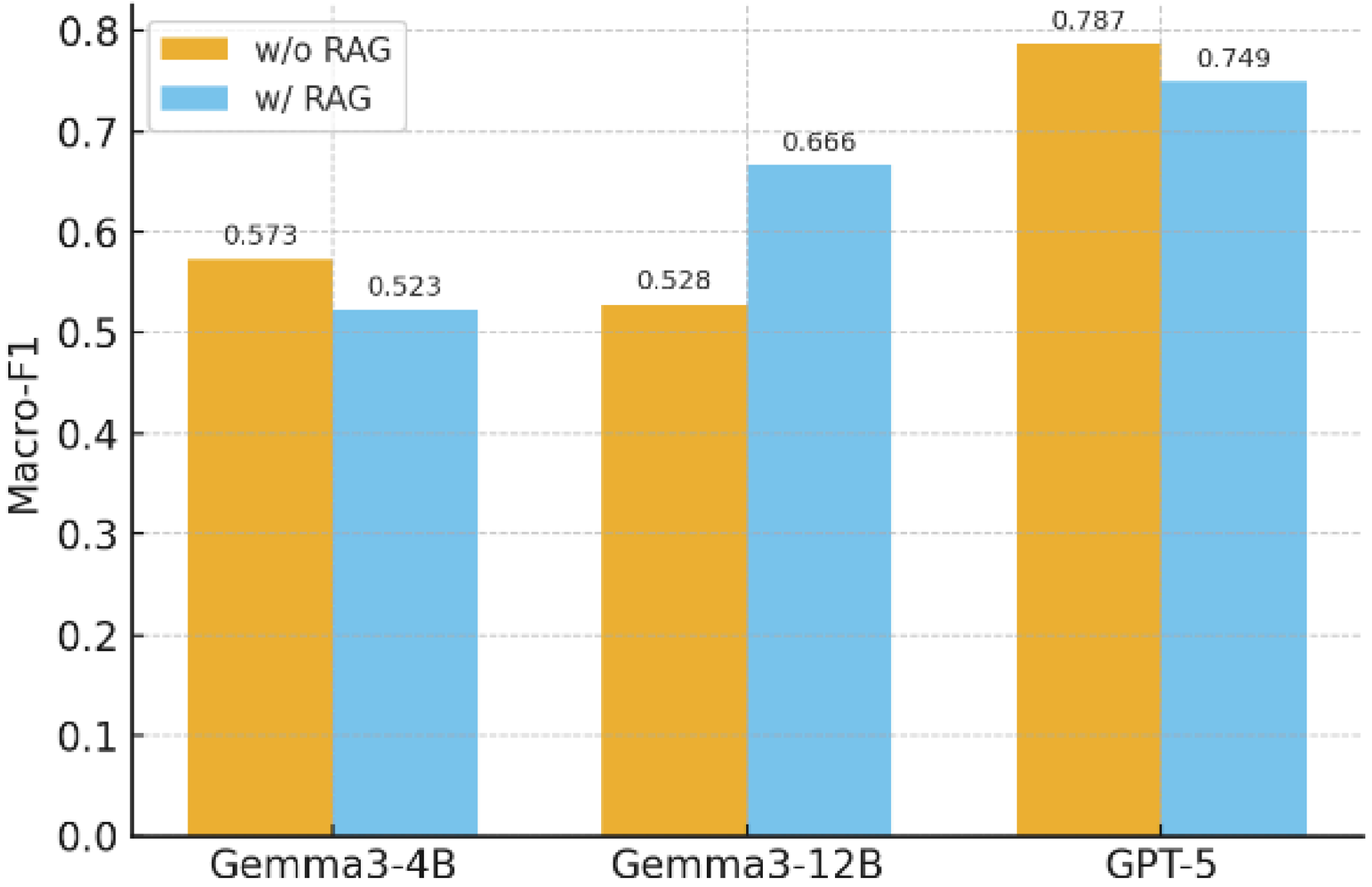

Figure 5. Subtask-level performance on evidence quality of the Promise–Evidence Pair (w/ vs. w/o RAG across models)..

Source: This study

4.2 Subtask-Level Performance Analysis

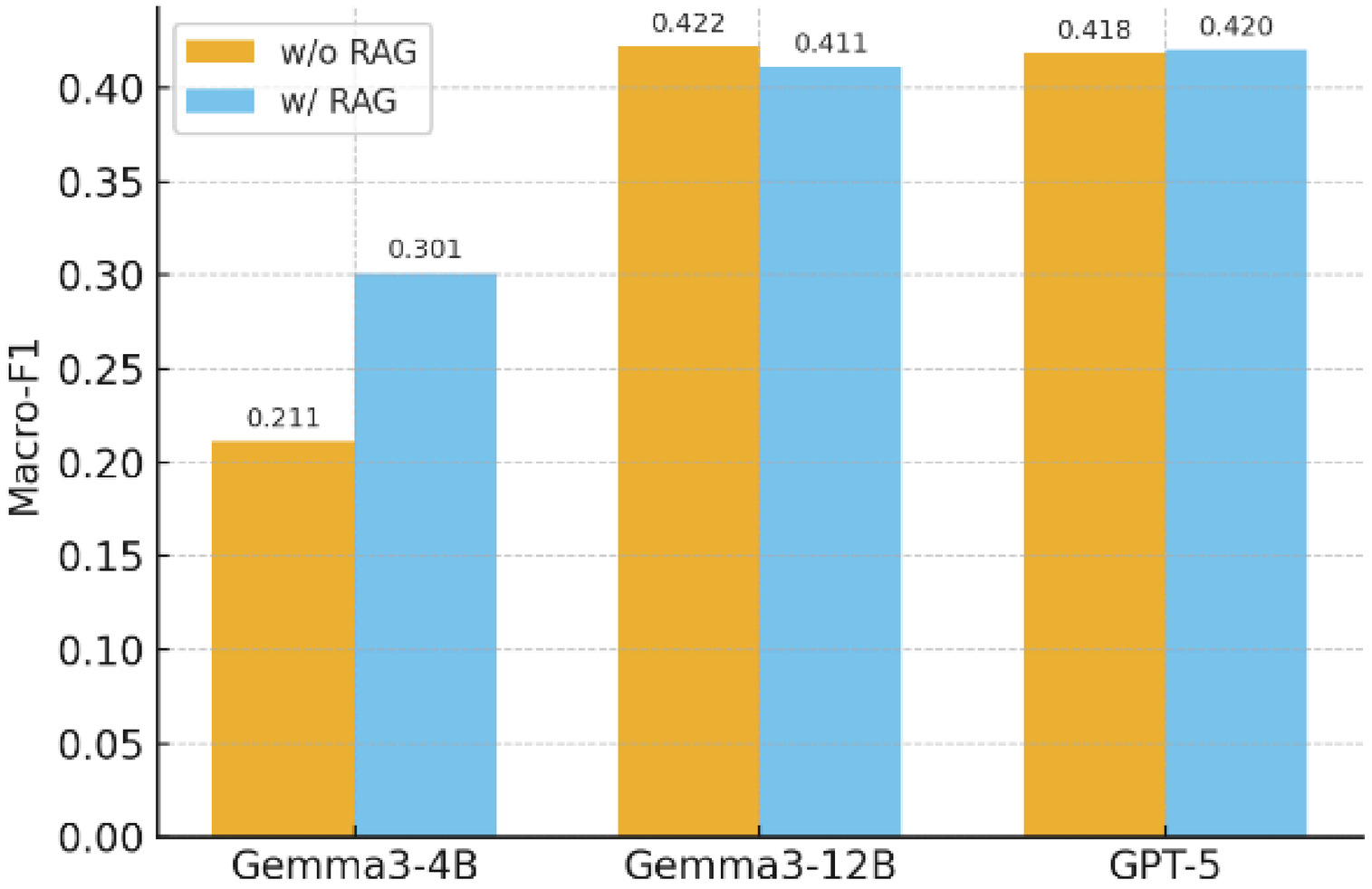

Figure 6. Subtask-level performance on Timing for Verification (w/ vs. w/o RAG across models).

Source: This study

4.3 Effect of Model Scale

Task	Gemma3-4B (ΔF1)	Gemma3-12B (ΔF1)	GPT5 (ΔF1)
Promise Identification	0.625 (+0.116)	0.754 (+0.020)	**0.756** (+0.069)
Supporting Evidence	0.523 (–0.050)	0.666 (+0.138)	**0.749** (–0.038)
Evidence Quality	0.285 (+0.047)	0.330 (+0.061)	**0.419** (+0.054)
Verification Timeline	0.301 (+0.090)	0.411 (–0.011)	**0.420** (+0.002)

Table 2. Subtask-level Macro-F1 with RAG across small (Gemma3-4B), medium (Gemma3-12B), and large (GPT-5) models, with ΔF1 relative to no-RAG baseline. Bold values indicate the best performance per subtask.

Source: This study

5. Conclusion

- Research Contributions
- Managerial Implications
- Future work

5.1 Research Contributions

- Offers empirical analysis results for the verification of ESG commitments specifically in a single language

- Demonstrates how **Retrieval-Augmented Generation (RAG)** impacts the performance of **Large Language Models (LLMs) of varying sizes** (large, medium, and small) in the ESG verification tasks.

5.2 Managerial Implications

- The approach can help regulators more efficiently identify unsupported or exaggerated corporate sustainability commitments in reports.
- It provides guidance to companies on how to improve their sustainability disclosures, thereby enhancing their verifiability and credibility.
- The ultimate goal is to strengthen the trust of investors and the public in corporate sustainability reports.

5.3 Future work

- Extend the current RAG-enhanced framework to be applied to **larger multilingual** corpora for ESG verification.
- Optimize retrieval strategies to reduce noise and further enhance the **robustness** and **accuracy** of the system.
- Investigate the framework's applicability to other **languages** and **domain-specific ESG contexts**.

Reference

- C. Xu, Y. Miao, Y. Xiao, and C. Lin, "DeepGreen: Effective LLM-Driven Green-washing Monitoring System Designed for Empirical Testing--Evidence from China," arXiv preprint arXiv:2504.07733, 2025.
- Y. Seki, H. Shu, A. Lhuissier, H. Lee, J. Kang, M.-Y. Day, and C.-C. Chen, "ML-Promise: A Multilingual Dataset for Corporate Promise Verification," arXiv preprint arXiv:2411.04473, 2024.
- K. Ong, R. Mao, D. Varshney, E. Cambria, and G. Mengaldo, "Towards Robust ESG Analysis Against Greenwashing Risks: Aspect-Action Analysis with Cross-Category Generalization," arXiv preprint arXiv:2502.15821, 2025.
- A. Vinella, M. Capetz, R. Pattichis, C. Chance, and R. Ghosh, "Leveraging language models to detect greenwashing," arXiv preprint arXiv:2311.01469, 2023.
- Y. Xie, C. Wang, J. Yan, J. Zhou, F. Deng, and J. Huang, "Making small language models better multi-task learners with mixture-of-task-adapters," in Proceedings of the 17th ACM International Conference on Web Search and Data Mining, 2024, pp. 1094-1097.
- K. Li, L. Zhang, Y. Jiang, P. Xie, F. Huang, S. Wang, and M. Cheng, "LaRA: Benchmarking Retrieval-Augmented Generation and Long-Context LLMs--No Silver Bullet for LC or RAG Routing," arXiv preprint arXiv:2502.09977, 2025.
- X. Lin, Y. Ning, J. Zhang, Y. Dong, Y. Liu, Y. Wu, X. Qi, N. Sun, Y. Shang, and P. Cao, "LLM-based Agents Suffer from Hallucinations: A Survey of Taxonomy, Methods, and Directions," arXiv preprint arXiv:2509.18970, 2025.
- J. Wallat, M. Heuss, M. D. Rijke, and A. Anand, "Correctness is not Faithfulness in Retrieval Augmented Generation Attributions," in ICTIR 2025 - Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval, 2025, pp. 22-32, doi: 10.1145/3731120.3744592. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105013792040&doi=10.1145%2f3731120.3744592&partnerID=40&md5=dbd1f068b642483b10c28d3e9921b088
- L. Zhang, Z. Jiang, H. Chi, H. Chen, M. Elkoumy, F. Wang, Q. Wu, Z. Zhou, S. Pan, and S. Wang, "Diagnosing and Addressing Pitfalls in KG-RAG Datasets: Toward More Reliable Benchmarking," arXiv preprint arXiv:2505.23495, 2025.
- S. Krishna, K. Krishna, A. Mohananey, S. Schwarcz, A. Stambler, S. Upadhyay, and M. Faruqui, "Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation," arXiv preprint arXiv:2409.12941, 2024.
- Y. Zheng, B. Li, Z. Lin, Y. Luo, X. Zhou, C. Lin, G. Li, and J. Su, "Revolutionizing Database Q&A with Large Language Models: Comprehensive Benchmark and Evaluation," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2025, vol. 2, pp. 5960-5971, doi: 10.1145/3711896.3737405. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-105014326087&doi=10.1145%2f3711896.3737405&partnerID=40&md5=57754945496427b3f9fda76ba90da80c

VALIDATING ESG COMMITMENTS WITH RAG-ENHANCED LARGE LANGUAGE MODELS: TOWARD TRANSPARENT AND RELIABLE SUSTAINABILITY DISCLOSURE

Q&A

Hsin-Ting Lu

Min-Yuh Day*

Graduate Institute of Information Management, National Taipei University, New Taipei City, Taiwan
myday@gm.ntpu.edu.tw*

Keywords: Retrieval-Augmented Generation (RAG), Large Language Models (LLMs), ESG Reports, ESG Commitment Validation, Greenwashing Detection

The 31st International Conference on IT Applications and Management (ITAM), Jinju, Korea, October 26-30, 2025.