

Evaluating Retrieval Augmentation for Multilingual ESG Promise Verification Across Model Capability Tiers and Modeling Paradigms



Min-Yuh Day, Ph.D.

Graduate Institute of
Information Management,
National Taipei University
Presenter: Hsin-Ting Lu
Instructor: Min-Yuh Day, Ph.D.

Date: 2026.10.20



Hsin-Ting Lu

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

OUTLINE

- 1 **Introduction**
- 2 **Literature Review**
- 3 **Research Methodology**
- 4 **Experimental Results and Analysis**
- 5 **Conclusions**

1. Introduction

1.1 Research Background

1.2 Research Motivation

1.3 Research Objective

1.1 Research Background

The Rise of ESG Reporting

- ESG reporting has become a key foundation for modern corporate responsibility and sustainable investment, with stakeholders increasingly relying on these disclosures to judge whether a company can be trusted (Agrawal, 2025).

The Threat of Greenwashing

- Some companies spread **misleading or unsupported claims** about their environmental performance, which seriously threatens market stability (Lagasio, 2024).
- The **complicated language** and **huge volume** of corporate reports make it very **time-consuming and biased** to verify them manually (Mousavian Anaraki et al., 2025).

The Need for Automated Solutions

- There is an urgent need **for automated tools** that combine NLP and RAG to **accurately detect false or misleading claims** in corporate sustainability reports (Kaoukis et al., 2025; Chen et al., 2025).

1.2 Research Motivation

Limitations of Current Practice

- Even though more companies are required to publish ESG reports, some use **complicated language** to hide their poor environmental performance (Seki et al., 2024).
- Traditional audit methods are too **slow and limited** to catch these subtle and technical forms of deception at scale (Ong et al., 2025; Chuang et al., 2025).

Gap in Research

- There is no standard checking mechanism, making it **hard for investors and policymakers** to tell whether corporate statements are genuine commitments or just marketing (Seki et al., 2024).
- Existing research still lacks a complete framework that **combines language analysis and RAG techniques** to verify whether corporate claims are actually true (He et al., 2025).

1.3 Research Objective

Main Objectives

- Evaluate whether **RAG and reasoning-enhanced RAG** improve Macro-F1 in multilingual ESG promise verification, and measure the added value of reasoning.
- Examine how RAG effectiveness varies across **model sizes, datasets, languages, and tasks**, and whether it reduces performance gaps between models.
- Compare API-based generative LLMs with **supervised encoder classifiers** on identical held-out test sets.



Research Question

- **RQ1:** Do **RAG and reasoning** improve Macro-F1 over non-retrieval baselines across ESG verification tasks, datasets, and languages? Does reasoning add gains beyond RAG?
- **RQ2:** Do RAG effects vary by **model tier, dataset, language, and task**? Can RAG narrow the performance gap between models?
- **RQ3:** How do API-based LLMs compare with **language-appropriate fine-tuned encoder** classifiers on identical test sets?

2. Literature Review

2.1 ESG Reporting and the Challenge of Greenwashing

2.2 Corporate Promise Verification

2.3 Large Language Models

2.4 Retrieval-Augmented Generation

2.5 Summary

2.1 ESG Reporting and the Challenge of Greenwashing

- ESG reporting has become a core mechanism for demonstrating corporate accountability and sustainability commitments (Forliano et al., 2025).
- The growing volume of disclosures has been accompanied by strategic behaviors that systematically undermine transparency (Ruggeri et al., 2025).

2.2 Corporate Promise Verification

From Rhetoric to Evidence

- Advanced information systems reduce the gap between reported performance and actual environmental outcomes (Luo et al., 2026).
- Recent frameworks explicitly link sustainability claims with verifiable corporate actions rather than vague promises (Ong et al., 2025).

Computational Approaches

- Fine-tuned language models can quantify greenwashing risk through mathematical formulations (Vinella et al., 2023).
- RAG-based frameworks extract and verify structured data points from complex ESG documents (Zou et al., 2025).

2.3 Large Language Models

LLMs broadly encompass both **decoder-based generative models** and **encoder-based pre-trained models**, each serving distinct roles in NLP tasks.

Two Core Model Architectures

- Decoder-based models (e.g., GPT series, Gemini-3, Claude Opus 5 & Sonnet 5) leverage large-scale pre-training and are optimized via prompt engineering and in-context learning without modifying underlying weights (Mienye et al., 2025; Zhang et al., 2026).
- Encoder-based models employ **supervised fine-tuning**, making them suitable for high-precision domain-specific tasks.
- Distinguishing these two architectures is essential for building robust ESG promise verification frameworks (Liu et al., 2025; Mienye et al., 2025).

2.4 Retrieval-Augmented Generation

Table 1 Comparison of Five RAG Paradigms

Paradigm	Key Features	Strengths	References
Naïve RAG	Single-pass retrieve-then-generate	Simple and low-cost	Lewis et al. (2020); Gao et al. (2023)
Advanced RAG	Dense retrieval with query rewriting and reranking	More accurate, less noisy context	Karpukhin et al. (2020); Gao et al. (2023)
Modular RAG	Interchangeable modules with flexible flows	Flexible and reconfigurable	Gao et al. (2024)
Graph RAG	Entity–relation graph index for retrieval	Relational and multi-hop reasoning	Edge et al. (2024); Peng et al. (2024)
Agentic RAG	Agents dynamically plan and refine retrieval	Adapts to complex, multi-step tasks	Singh et al. (2025)

2.5 Summary

Table 2 Comparison of PRAGMA and Prior Approaches in ESG Verification

Method	Verification Scope	Grounded Evidence	Multilingual	Source
Rule-based	Keyword detection	Low (No external evidence)	Limited to a single language	Ong et al., (2025)
LLM	Scoring of 4 semantic attributes	None (Model fine-tuning only)	English	Vinella et al. (2023)
Existing RAG	4-task promise verification	Yes (RAG-based)	Multilingual (English, French, Chinese, Japanese, Korean)	Seki et al. (2024)
PRAGMA (Promise RAG Agentic Grounded Multilingual Analyzer)	4-dimensional logical verification	Strong (Dual-track dynamic RAG)	English, Traditional Chinese	This Study

Source: This study

3. Research Methodology

3.1 System Architecture

3.2 Dataset

3.3 Retrieval Corpus and Indexing

3.4 Model Selection

3.5 Evaluation Metric

3.1 PRAGMA System Architecture

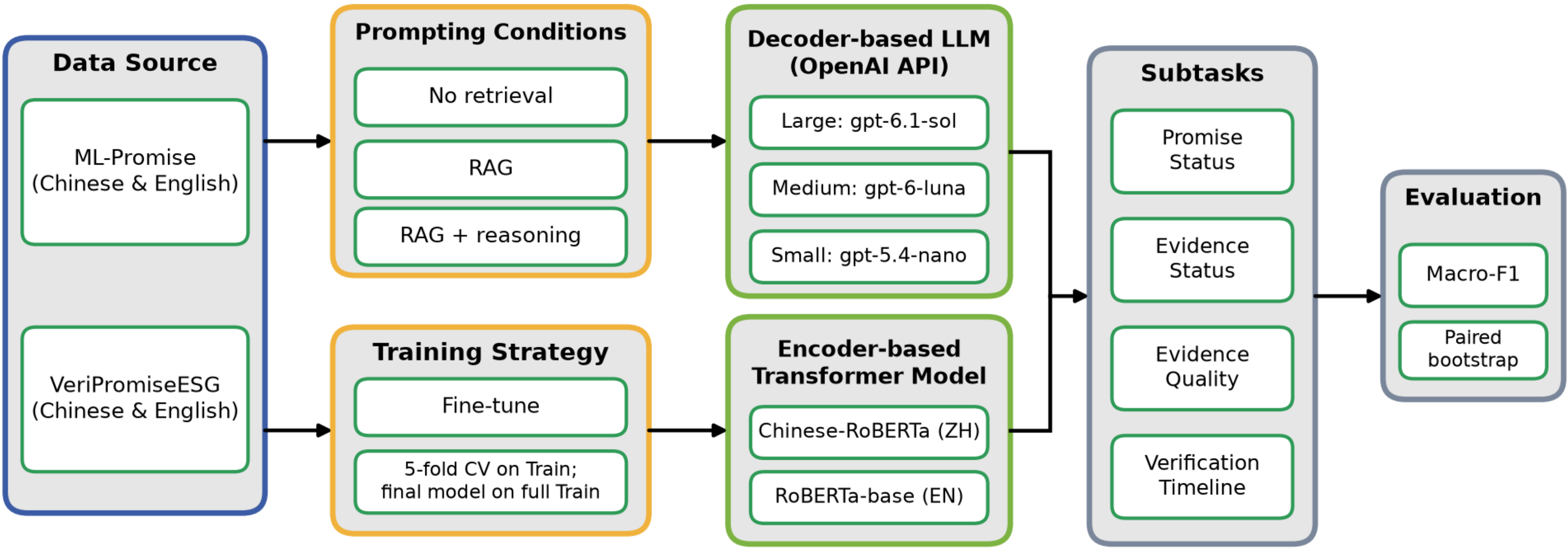


Figure 1 Proposed Dual-Path Research Architecture Integrating Retrieval-Augmented Decoder Models and Fine-Tuned Encoder Models for Multilingual ESG Promise Verification

3.2 Dataset (ML_Promise)

- Source: ML_Promise
- Language: Chinese & English
- Sample Size: Chinese: 410 train / 489 test; English: 400 train / 400 test

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 (VeripromiseESG)

- Source: **VeripromiseESG**
- Language: **Traditional** Chinese & English
- Sample Size: **Train 1000 / Val 1000 / Test 2000**
- **Coverage: Taiwan 50 Index (TWSE: 0050), 15 industry sectors**

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).

3.3 Retrieval Corpus and Indexing

Retrieval Corpus:

- Source: training splits of **ML-Promise (Chinese & English)** and **VeriPromiseESG (Chinese & English)** datasets
- Tools: Use **multilingual-e5-base model (dense similarity search)**

Inference Process:

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

3.4 Model Selection

Decoder-based Models (RAG Path) :

- Small Scale: GPT-5.4-nano
- Medium Scale: GPT-6-luna
- Large Scale: GPT-6.1 Sol

Encoder-based Model (Fine-tuning Path) :

- Chinese-RoBERTa (Chinese)
- RoBERTa-base (English)

3.5 Evaluation Metric

- **Metric Chosen:**
 - Adopted the **Macro-Averaged F1 Score (Macro-F1)** across all four subtasks: Promise Status, Evidence Status, Evidence Quality, and Verification Timeline.
 - Paired bootstrap, 95% CI on Macro-F1 differences
- **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

4.1 Overall Performance

4.2 Effect of RAG

4.3 Incremental Effect of Reasoning Prompting beyond RAG

4.4 Effect of Model Capability Tier

4.5 Variation across Datasets, Language Versions, and Tasks

4.6 Generative LLM Systems versus Supervised Encoder Classifiers

4.1 Overall Performance

Table 3 Mean Macro-F1 of the API tiers and methods and the RoBERTa encoder.

Model	NORAG	RAG	RAG+Reasoning	Encoder Model
<i>Chinese (mean of 8 dataset × task cells)</i>				
GPT-6.1-Sol	0.489	0.621	0.621	
GPT-6-Luna	0.493	0.617	0.620	
GPT-5.4-Nano	0.382	0.531	0.510	
Chinese-RoBERTa				0.623
<i>English (mean of 8 dataset × task cells)</i>				
GPT-6.1-Sol	0.532	0.598	0.598	
GPT-6-Luna	0.518	0.581	0.576	
GPT-5.4-Nano	0.455	0.499	0.505	
RoBERTa-base				0.549

Source: This study

1. RAG improves all three model tiers by a similar margin (+0.097 / +0.093 / +0.099 for Nano / Luna / Sol).
2. The gain is larger in Chinese (+0.123 to +0.149) than in English (+0.044 to +0.066).
3. RAG improves performance across tiers, while adding reasoning (RAG+Reasoning) brings no further gain (at most −0.008 vs. RAG).

4.2 Effect of RAG

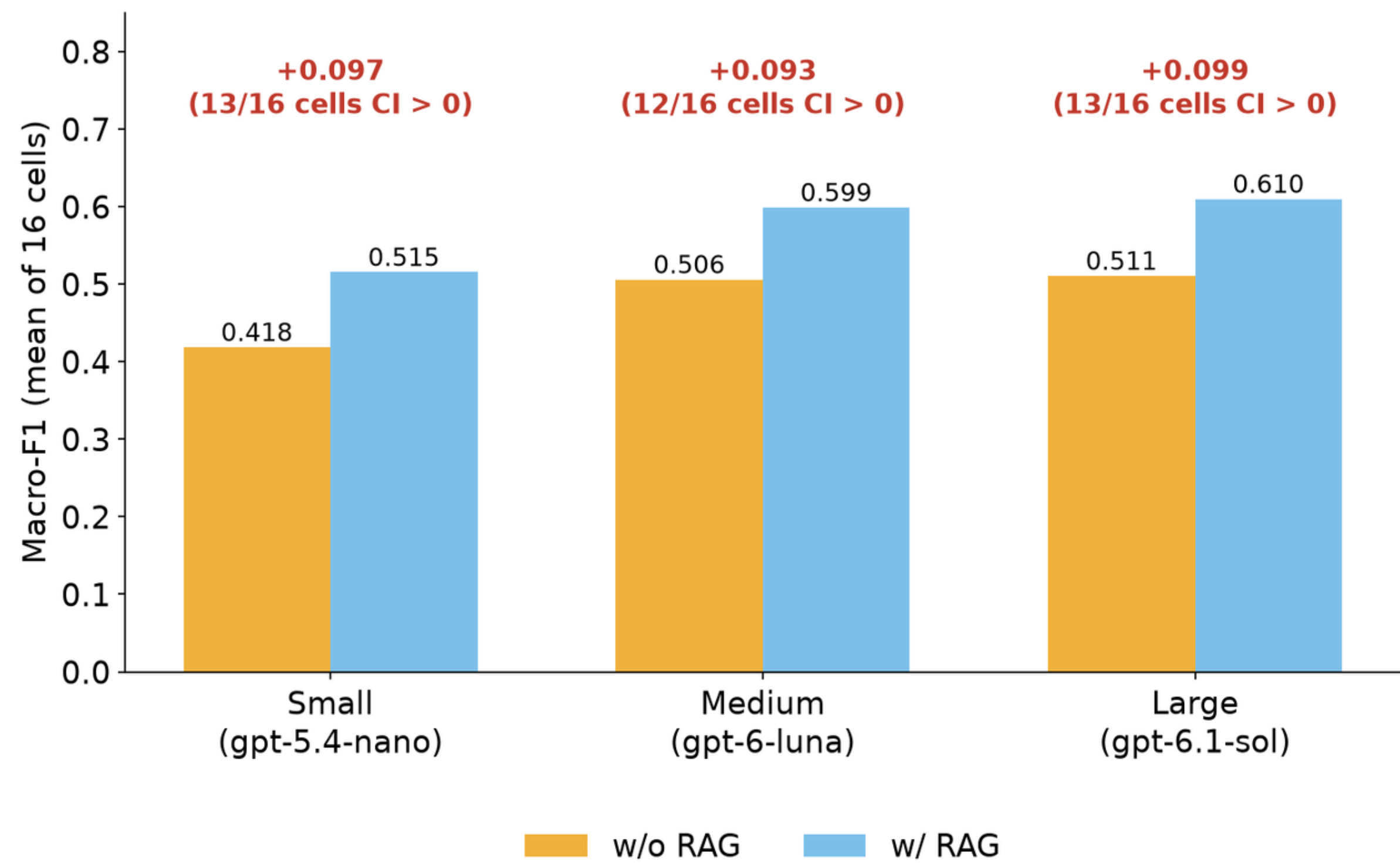


Figure 2. Mean Macro-F1 over 16 cells, with vs. without RAG (red: mean gain; cells with 95% CI above 0).

Source: This study

4.3 Incremental Effect of Reasoning Prompting beyond RAG

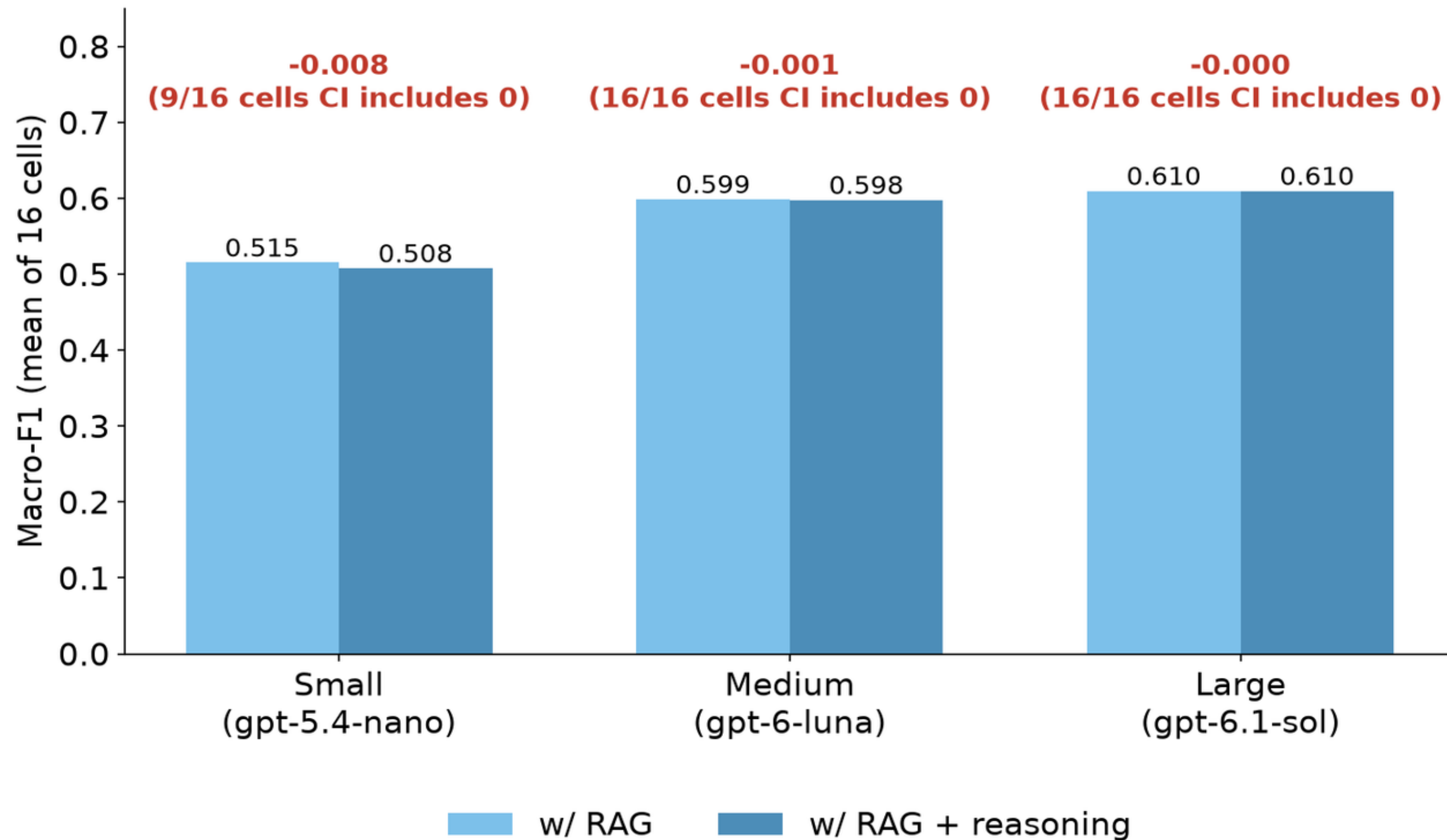
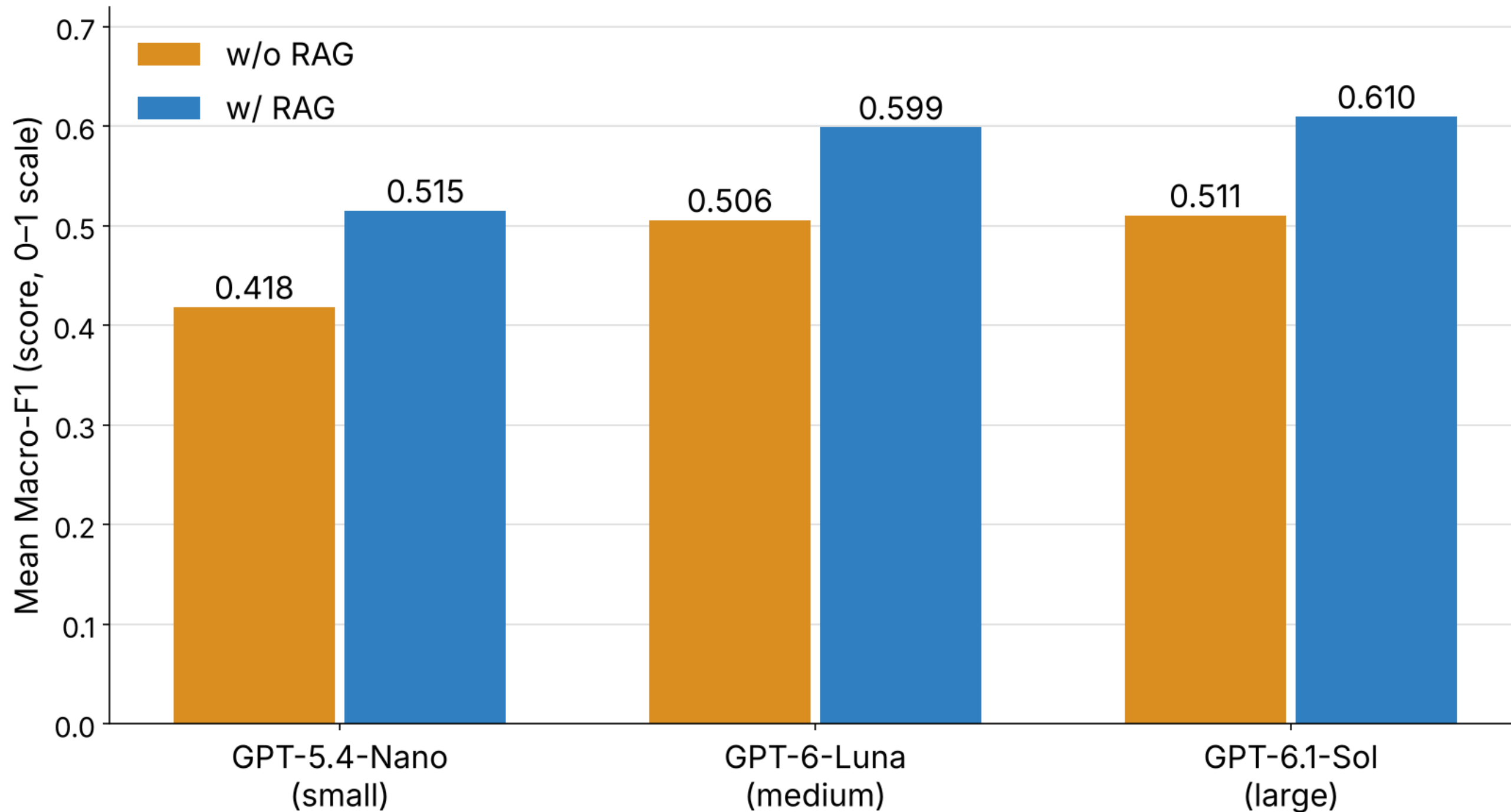


Figure 3. RAG vs. RAG + reasoning (red: mean gain; cells whose CI includes 0).

Source: This study

4.4 Effect of Model Capability Tier



- Small → Medium: large gain.** Macro-F1 rises by about +0.09 (+0.087 w/o RAG, +0.083 w/ RAG).
- Medium → Large: small gain.** The improvement is only +0.005 w/o RAG and +0.011 w/ RAG.
- RAG helps at every tier.** Macro-F1 rises by about +0.09 to +0.10 at each tier.

Figure 4. Mean Macro-F1 by Model Tier, with and without RAG

Source: This study

4.5 Variation across Datasets, Language Versions, and Tasks

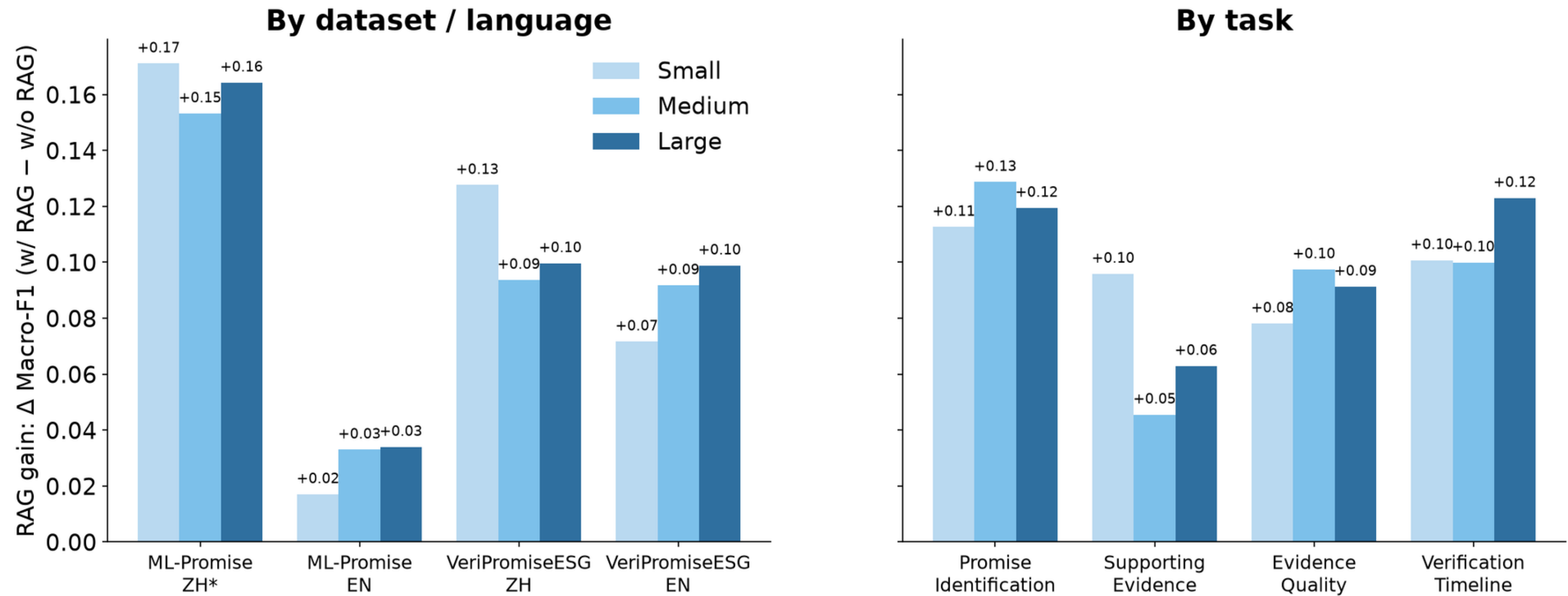


Figure 5. RAG gain by dataset / language and by task

Source: This study

4.6 Generative LLM Systems versus Supervised Encoder Classifiers

Cells where API is better / worse / indistinguishable from Encoder (of 16)

	w/o RAG	w/ RAG	w/ RAG + reasoning
Small (gpt-5.4-nano)	0 / 13 / 3	1 / 11 / 4	0 / 10 / 6
Medium (gpt-6-luna)	2 / 11 / 3	5 / 6 / 5	5 / 6 / 5
Large (gpt-6.1-sol)	2 / 10 / 4	8 / 6 / 2	8 / 6 / 2

Encoder baseline: mean Macro-F1 = 0.586

Figure 6. Number of cells (of 16) where each API setting is better than, worse than, or indistinguishable from the RoBERTa encoder (paired bootstrap, 95% CI).

5. Conclusion

5.1 Major Finding

5.2 Academic Research Contributions

5.3 Managerial Implications

5.4 Future work

5.1 Major Finding

Retrieval drives the gains

1. RAG raised mean Macro-F1 by about **+0.09 to +0.10 in all three tiers** (e.g., Large: 0.511 → 0.610).
2. RAG + reasoning showed no statistical evidence of an additional gain over RAG alone.

RAG-enhanced LLMs are competitive with the encoder

1. Medium and Large with RAG matched the Chinese encoder **(0.62 vs. 0.62)** and exceeded the English encoder **(0.58–0.60 vs. 0.55)**.
2. Moving from Small to Medium helped most **(about +0.08 to +0.09)**; Medium to Large added only about **+0.01**.

5.2 Academic Research Contributions

1. Controlled evaluation

Separates the effects of RAG from reasoning prompting.

2. Cross-condition analysis

Examines RAG effectiveness across model tiers, datasets, languages, and ESG tasks.

3. Fair model comparison

Compares generative LLMs and encoder classifiers on identical held-out test data.

4. Multilingual robustness

Evaluates performance across naturally multilingual and translated data.

5.3 Managerial Implications

1. Identifies when RAG provides meaningful performance gains.
2. Examines whether retrieval-enhanced smaller models can approach larger models.
3. Supports model selection under different cost and resource constraints.
4. Highlights multilingual and translation-related performance distortions.

5.4 Future work

1. Optimize retrieval performance by incorporating **reranker and hybrid search** mechanisms to improve the precision of retrieved evidence for ESG promise verification.
2. Enhance model **explainability** through evidence attribution, enabling users to trace which retrieved passages support each verification decision.
3. Explore real-world **deployment in regulatory systems** to automate large-scale ESG disclosure monitoring and corporate accountability auditing.

References

- Abo El-Enen, M., Saad, S., & Nazmy, T. (2025). A survey on retrieval-augmentation generation (RAG) models for healthcare applications. *Neural Computing and Applications*, 37(33), 28191-28267. <https://doi.org/10.1007/s00521-025-11666-9>
- Agrawal, A. (2025, 16-17 Feb. 2025). Assessing ESG Compliance and Impact: A Zero-Shot Learning Approach to Analyzing Fortune 500 Companies' Sustainability Reports. 2025 8th International Conference on Data Science and Machine Learning Applications (CDMA),
- Calamai, T., Balalau, O., Guenedal, T. e. L., & Suchanek, F. M. (2025). Detecting Greenwashing: A Natural Language Processing Literature Survey.
- Chae, Y., & Davidson, T. (2025). Large Language Models for Text Classification: From Zero-Shot Learning to Instruction-Tuning [Article]. *Sociological Methods and Research*. <https://doi.org/10.1177/00491241251325243>
- Chen, C.-C., Seki, Y., Shu, H., Lhuissier, A., Kang, J., Lee, H., Day, M.-Y., & Takamura, H. (2025, July). SemEval-2025 Task 6: Multinational, Multilingual, Multi-Industry Promise Verification. In S. Rosenthal, A. Rosá, D. Ghosh, & M. Zampieri, *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)* Vienna, Austria.
- Chuang, M., Chuang, G., Chuang, C., & Chuang, J. (2025, July). Judging It, Washing It: Scoring and Greenwashing Corporate Climate Disclosures using Large Language Models. In K. Dutia, P. Henderson, M. Leippold, C. Manning, G. Morio, V. Muccione, J. Ni, T. Schimanski, D. Stambach, A. Singh, A. Su, & S. A. Vaghefi, *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)* Vienna, Austria.
- Forliano, C., Battisti, E., de Bernardi, P., & Klietik, T. (2025). Mapping the greenwashing research landscape: a theoretical and field analysis [Review]. *Review of Managerial Science*, 19(11), 3407-3456. <https://doi.org/10.1007/s11846-025-00856-3>
- Gardazi, N. M., Daud, A., Malik, M. K., Bukhari, A., Alsahfi, T., & Alshemaimri, B. (2025). BERT applications in natural language processing: a review. *Artificial Intelligence Review*, 58(6), 166. <https://doi.org/10.1007/s10462-025-11162-5>
- Gorovaia, N., & Makrominas, M. (2024). Identifying greenwashing in corporate-social responsibility reports using natural-language processing. *European Financial Management*, 31. <https://doi.org/10.1111/eufm.12509>
- Hao, X., Tian, T., Dong, L., Wong, C. W. Y., & Lai, K.-h. (2026). Unmasking greenwashing in ESG disclosure: insights from evolutionary game analysis. *Annals of Operations Research*, 356(1), 667-695. <https://doi.org/10.1007/s10479-025-06538-3>
- He, C., Zhou, X., Wu, Y., Yu, X., Zhang, Y., Zhang, L., Wang, D., Lyu, S., Xu, H., Xiaoqiao, W., Liu, W., & Miao, C. (2025, November). ESGenius: Benchmarking LLMs on Environmental, Social, and Governance (ESG) and Sustainability Knowledge. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* Suzhou, China.
- Hindi, M., Mohammed, L., Maaz, O., & Alwarafy, A. (2025). Enhancing the Precision and Interpretability of Retrieval-Augmented Generation (RAG) in Legal Technology: A Survey. *IEEE Access*, 13, 46171-46189. <https://doi.org/10.1109/ACCESS.2025.3550145>
- Jiang, N., Wu, G., Zou, W., & Chen, X. (2026). Responsible use of artificial intelligence and the mitigation of corporate greenwashing: The role of executive oversight. *Socio-Economic Planning Sciences*, 105, 102436. <https://doi.org/https://doi.org/10.1016/j.seps.2026.102436>
- Kaoukis, G., Koufopoulos, I. A., Psaroudaki, E., Karidi, D. P., Pitoura, E., Papastefanatos, G., & Tsaparas, P. (2025). EmeraldMind: A Knowledge Graph-Augmented Framework for Greenwashing Detection. *arXiv*, abs/2512.11506.
- Klesel, M., & Wittmann, H. F. (2025). Retrieval-Augmented Generation (RAG). *Business & Information Systems Engineering*, 67(4), 551-561. <https://doi.org/10.1007/s12599-025-00945-3>

References

- Lagasio, V. (2024). ESG-washing detection in corporate sustainability reports. *International Review of Financial Analysis*, 96, 103742. <https://doi.org/https://doi.org/10.1016/j.irfa.2024.103742>
- Lee, H., Lee, S. H., Park, H., Kim, J. H., & Jung, H. S. (2024). ESG2PreEM: Automated ESG grade assessment framework using pre-trained ensemble models. *Heliyon*, 10(4), e26404. <https://doi.org/https://doi.org/10.1016/j.heliyon.2024.e26404>
- Liu, L., Ma, Z., Zhou, Y., Fan, M., & Han, M. (2024). Trust in ESG reporting: The intelligent Veri-Green solution for incentivized verification. *Blockchain: Research and Applications*, 5(2), 100189. <https://doi.org/https://doi.org/10.1016/j.bcra.2024.100189>
- Liu, Y., Cao, J., Liu, C., Ding, K., & Jin, L. (2025). Datasets for large language models: a comprehensive survey. *Artificial Intelligence Review*, 58(12), 403. <https://doi.org/10.1007/s10462-025-11403-7>
- Lublóy, Á., Keresztúri, J. L., & Berlinger, E. (2025). Quantifying firm-level greenwashing: A systematic literature review. *Journal of Environmental Management*, 373, 123399. <https://doi.org/https://doi.org/10.1016/j.jenvman.2024.123399>
- Luo, K., Xue, J., Rassiah, P., & Lim, E. K. (2026). Smart accountability: leveraging AI to align ESG disclosure with practice. *International Journal of Accounting Information Systems*, 57, 100773. <https://doi.org/https://doi.org/10.1016/j.accinf.2026.100773>
- Ma, Y., & Han, W.-Z. (2025). How false-disclosure greenwashing affects corporate financial sustainability. *Discover Sustainability*, 7(1), 136. <https://doi.org/10.1007/s43621-025-02465-2>
- Ma, Y., Nie, H., Chen, C., Zhang, J., Jiang, J., Wang, B., & Xia, Y. (2025). A Survey of Retrieval-Augmented Generation (RAG) for Large Language Models 2025 International Conference on Trustworthy Big Data and Artificial Intelligence (ICTBAI), <https://doi.ieeecomputersociety.org/10.1109/ICTBAI68361.2025.00008>
- Mehra, S., Louka, R., & Zhang, Y. (2022). Esgbert: Language model to help with classification tasks related to companies environmental, social, and governance practices. *arXiv preprint arXiv:2203.16788*.
- Mienye, I. D., Jere, N., Obaido, G., Ogunraku, O. O., Esenogho, E., & Modisane, C. (2025). Large language models: an overview of foundational architectures, recent trends, and a new taxonomy [Review]. *Discover Applied Sciences*, 7(9), Article 1027. <https://doi.org/10.1007/s42452-025-07668-w>
- Mohamed, A., Rashid, M. E., & Shaalan, K. (2025). In-Context Learning in Large Language Models (LLMs): Mechanisms, Capabilities, and Implications for Advanced Knowledge Representation and Reasoning. *IEEE Access*, 13, 95574-95593. <https://doi.org/10.1109/ACCESS.2025.3575303>
- Mousavian Anaraki, S. A., Croce, D., & Basili, R. (2025). Large language models for sustainability reporting: A systematic review and research agenda. *Sustainable Futures*, 10, 101494. <https://doi.org/https://doi.org/10.1016/j.sftr.2025.101494>
- Ong, K., Mao, R., Varshney, D., Cambria, E., & Mengaldo, G. (2025). Towards Robust ESG Analysis Against Greenwashing Risks: Aspect-Action Analysis with Cross-Category Generalization. Proceedings of the Annual Meeting of the Association for Computational Linguistics,
- Ramesh, G., Sahil, M., Palan, S. A., Bhandary, D., Ashok, T. A., Shreyas, J., & Sowjanya, N. (2025). A review on NLP zero-shot and few-shot learning: methods and applications. *Discover Applied Sciences*, 7(9), 966. <https://doi.org/10.1007/s42452-025-07225-5>
- Rostam, Z. R. K., & Kertész, G. (2024, 23-25 Oct. 2024). Fine-Tuning Large Language Models for Scientific Text Classification: A Comparative Study. 2024 IEEE 6th International Symposium on Logistics and Industrial Informatics (LINDI),
- Ruggeri, M., Vinci, G., Ruggieri, R., & Savastano, M. (2025). Facing the Risk of Greenwashing in the ESG Report of Global Companies: The Importance of Life Cycle Thinking [Article]. *Corporate Social Responsibility and Environmental Management*, 32(3), 4216-4234. <https://doi.org/10.1002/csr.3178>
- 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. <https://doi.org/10.1016/j.frl.2024.104979>
- Seki, Y., Shu, H., Lhuissier, A., Lee, H., Kang, J., Day, M.-Y., & Chen, C.-C. (2024). ML-Promise: A Multilingual Dataset for Corporate Promise Verification. *arXiv preprint arXiv:2411.04473*.
- Sneideriene, A., & Legenzova, R. (2025). Greenwashing prevention in environmental, social, and governance (ESG) disclosures: A bibliometric analysis. *Research in International Business and Finance*, 74, 102720. <https://doi.org/https://doi.org/10.1016/j.ribaf.2024.102720>

References

- Vinella, A., Capetz, M., Pattichis, R., Chance, C., & Ghosh, R. (2023). Leveraging language models to detect greenwashing. *arXiv preprint arXiv:2311.01469*.
- Wang, S., Yang, H., & Liu, W. (2025). Research on the construction and application of retrieval enhanced generation (RAG) model based on knowledge graph. *Scientific Reports*, 15(1), 40425. <https://doi.org/10.1038/s41598-025-21222-z>
- Wang, X., Gao, X., & Sun, M. (2025). Construction and analysis of corporate greenwashing index: a deep learning approach. *EPJ Data Science*, 14(1), 44. <https://doi.org/10.1140/epjds/s13688-025-00562-w>
- Webersinke, N., Kraus, M., Bingler, J. A., & Leippold, M. (2021). Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*.
- Yadav, S., Samadhiya, A., Kumar, A., Luthra, S., & Pandey, K. K. (2025). Environmental, Social, and Governance (ESG) Reporting and Missing (M) Scores in the Industry 5.0 Era: Broadening Firms' and Investors' Decisions to Achieve Sustainable Development Goals. *Sustainable Development*, 33(3), 3455-3477. <https://doi.org/https://doi.org/10.1002/sd.3306>
- Zhang, M., Shen, Q., Zhao, Z., Wang, S., & Huang, G. Q. (2025). Optimizing ESG reporting: Innovating with E-BERT models in nature language processing. *Expert Systems with Applications*, 265, 125931. <https://doi.org/https://doi.org/10.1016/j.eswa.2024.125931>
- Zhang, Y., Zhao, M., Zhang, Y., & Cheung, Y. m. (2026). Trending Applications of Large Language Models: A User Perspective Survey. *IEEE Transactions on Artificial Intelligence*, 7(4), 1835-1852. <https://doi.org/10.1109/TAI.2025.3620272>
- Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J., & Cui, B. (2026). Retrieval-Augmented Generation for AI-Generated Content: A Survey. *Data Science and Engineering*, 11(1), 1-29. <https://doi.org/10.1007/s41019-025-00335-5>
- Zou, Y., Shi, M., Chen, Z., Deng, Z., Lei, Z., Zeng, Z., Yang, S., Tong, H., Xiao, L., & Zhou, W. (2025). ESGReveal: An LLM-based approach for extracting structured data from ESG reports [Article]. *Journal of Cleaner Production*, 489, Article 144572. <https://doi.org/10.1016/j.jclepro.2024.144572>

Evaluating Retrieval Augmentation for Multilingual ESG Promise Verification Across Model Capability Tiers and Modeling Paradigms

Q&A



Min-Yuh Day, Ph.D.

Graduate Institute of
Information Management,
National Taipei University

Presenter: Hsin-Ting Lu

Instructor: Min-Yuh Day, Ph.D.

Date: 2026.10.20



Hsin-Ting Lu

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