ESG-KG: A Multi-modal Knowledge Graph System for Automated Compliance Assessment

Li-Yang Chang[1] **Chih-Ming Chen**[1] **Hen-Hsen Huang**[2]
An-Zi Yen[3] Ming-Feng Tsai[4,5] Chuan-Ju Wang[1]

[1]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[2]Institute of Information Science, Academia Sinica, Taiwan
[3]Department of Computer Science, National Yang Ming
Chiao Tung University, Taiwan
[4]Department of Computer Science, National Chengchi University, Taiwan
[5]Delta Electronics, Taiwan

Abstract

We present ESG-KG, a system that automates ESG compliance assessment through multi-modal information extraction and knowledge graph construction.ESG-KG processes corporate sustainability reports containing diverse data formats—text, tables, figures, and infographics—and extracts ESG-related entities, relationships, and metrics into a structured knowledge graph. This KG-based architecture enables precise cross-modal information retrieval and provides verifiable evidence grounding for downstream analysis.Built upon this foundation, ESG-KG integrates retrieval-augmented generation (RAG) with LLM-based reasoning to automatically evaluate compliance against ESG frameworks and standards. Our demonstration showcases the system's end-to-end pipeline, from multi-modal document processing to automated compliance scoring, highlighting its capability to handle real-world sustainability reports and generate interpretable assessment results with traceable evidence chains. To facilitate further research, we release our open-source Python toolkit for Automated Compliance Assessment at `https://github.com/cnclabs/website.kg.esg.demo.git`, and a live demonstration video is available at `https://youtu.be/Lj4Zp74J1nY`.

1 Introduction

The growing emphasis on Environmental, Social, and Governance (ESG) issues has created a complex landscape of sustainability reporting standards. Among these, the Global Reporting Initiative (GRI) Standards have emerged as the most widely adopted framework for ESG disclosure, providing structured reporting principles and standardized indicators for systematic assessment of organizational sustainability performance. However, manually assessing compliance remains challenging due to the volume and heterogeneity of modern sustainability reports, which integrate textual, tabular, and visual elements.

Existing automated approaches for ESG compliance assessment typically rely on fact-based retrieval methodologies that decompose statements into individual claims and perform document-level evidence retrieval (Min et al., 2023). Recent advances in knowledge graph-based fact-checking (Chen et al., 2025) and multi-modal analysis (Wang et al., 2024b) have shown promise in handling complex document structures. However, these methods often struggle with the multi-modal nature of sustainability reports, where critical information is embedded not only in text but also in charts, graphs, and infographics. This limitation results in incomplete evidence gathering and compromises compliance verification accuracy.

To overcome these challenges, we present ESG-KG, a system that automates ESG compliance assessment through multi-modal information extraction and knowledge graph construction. Our approach extends beyond fact-level retrieval by incorporating layout analysis and visual data extraction to construct a comprehensive evidence base, capturing quantitative and contextual information from charts, tables, and infographics that text-only systems typically miss.

The extracted multi-modal data is structured into a knowledge graph that semantically aligns ESG disclosures with standard requirements, particularly the GRI Standards. This KG serves as a retrieval-augmented knowledge base, enabling accurate evidence extraction and providing a trusted foundation for LLM-based compliance assessment with traceable reasoning chains.

ESG-KG enables scalable, automated compli-

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 3: System Demonstrations, pages 602–608

ance evaluation while substantially reducing manual verification effort. Our demonstration showcases how the system processes real-world sustainability reports end-to-end, bridging multi-modal content with structured ESG criteria to deliver interpretable compliance assessments. By integrating multi-modal extraction with knowledge graph reasoning, ESG-KG promotes greater transparency, reliability, and efficiency in corporate sustainability reporting and compliance verification.

2 Related Work

LLM-based Compliance Assessment. Automated compliance assessment has evolved from rule-based systems to approaches leveraging LLMs (Radford and Narasimhan, 2018), (Radford et al., 2019), (Brown et al., 2020), (Lewis et al., 2020), (Raffel et al., 2020). Early LLM applications leveraged semantic capabilities to interpret regulatory texts and summarize requirements directly from their parametric knowledge (Min et al., 2023). However, pure LLM approaches prove insufficient for high-stakes auditing due to their susceptibility to hallucination and inability to systematically cross-reference lengthy corporate disclosures against complex regulatory frameworks without external grounding.

Retrieval-augmented Generation for Fact Verification. To address these limitations, retrieval-augmented generation (RAG) approaches have emerged that ground LLM reasoning in retrieved evidence. Methods like FActScore (Min et al., 2023) decompose claims into atomic facts and retrieve supporting evidence, while knowledge graph-based fact-checking systems (Chen et al., 2025) provide structured reasoning paths. However, these approaches primarily focus on textual evidence and struggle with the multi-modal nature of corporate reports.

Multi-modal Document Understanding. Recent work has begun addressing multi-modal information extraction from complex documents. However, a critical gap remains: corporate sustainability reports often present crucial quantitative metrics—such as emissions and resource use—not in continuous text but embedded in charts, infographics, and complex tables (Gupta et al., 2025). Existing ESG benchmarks and text-centric systems explicitly acknowledge the "exclusion of visual elements" as a major limitation that directly impacts evidence coverage (He et al., 2025).

Systems like SubstationAI (Wang et al., 2024b) demonstrate the value of processing visual elements alongside text, while recent document understanding approaches (Zhang et al., 2024) show progress in table structure recognition and chart-to-structured-data conversion. However, these multi-modal capabilities have not been systematically integrated with knowledge graph architectures for compliance assessment.

ESG-KG addresses this gap by combining specialized multi-modal extraction—including table structure recognition and chart data conversion—with knowledge graph construction, creating a unified framework where visual and textual evidence are jointly represented and retrievable for compliance verification.

3 The Proposed ESG-KG System

We present ESG-KG, a system for automated ESG compliance assessment that bridges unstructured regulatory documents with verifiable evidence retrieval from multi-modal corporate reports.

The system operates through three core components: (1) construction of a regulatory knowledge graph (KG) from GRI standards, (2) semantic refinement to ensure entity uniqueness and relational consistency, and (3) an online scoring pipeline that processes multimodal reports for evidence retrieval and compliance evaluation.

Figure 1 illustrates the system architecture.

3.1 Regulatory Knowledge Graph Construction

To transform the dense, unstructured text of GRI standards into a machine-interpretable format, we employ an LLM-driven extraction pipeline. We formalize the GRI standards as a Knowledge Graph (KG) $\mathcal{G} = (\mathcal{E}, \mathcal{R})$, where $\mathcal{E}$ represents the set of entities and $\mathcal{R}$ represents semantic relations between them.

The LLM parses raw standard documents into structured subgraphs following a strict schema. For each standard $\mathcal{S}$, extracted entities are decomposed into three categories:

$$\mathcal{E}_S = \mathcal{T}_{\text{disc}} \cup \mathcal{R}_{\text{req}} \cup \mathcal{C}_{\text{score}}, \tag{1}$$

where:

- $\mathcal{T}_{\text{disc}}$ (**Disclosure Targets**): Specific metrics organizations must disclose (e.g., "Scope 1 GHG emissions").

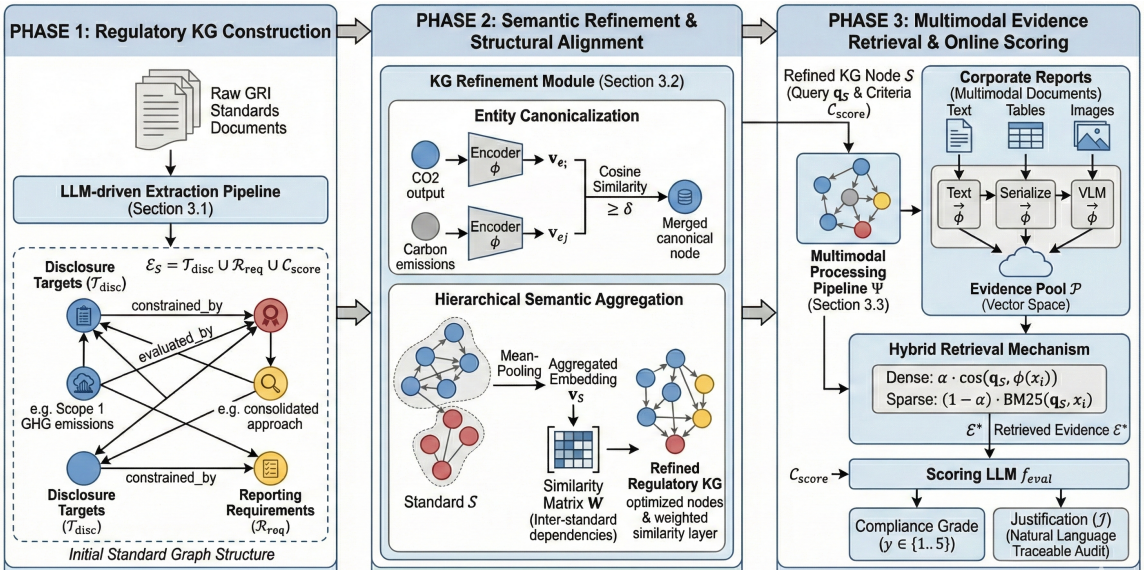

Figure 1: Multimodal Evidence Retrieval and Online Scoring Pipeline

- $\mathcal{R}_{\text{req}}$ (**Reporting Requirements**): Contextual constraints defining how disclosures should be prepared (e.g., "consolidated approach").

- $\mathcal{C}_{\text{score}}$ (**Scoring Criteria**): Evaluative rubrics for assessing disclosure completeness and quality .

Relations $r \in \mathcal{R}$ preserve the logical structure by linking these components. Requirement constrain targets via triples $(e_t, \texttt{constrained\_by}, e_r)$ where $e_t \in \mathcal{T}_{disc}$ and $e_r \in \mathcal{R}_{req}$. Evaluation logic is captured as $(e_t, \texttt{evaluated\_by}, e_c)$ where $e_c \in \mathcal{C}_{\text{score}}$. This structured representation creates a "standard graph" that serves as the foundation for subsequent compliance verification.

3.2 Semantic Refinement and Structural Alignment

Raw triples extracted from natural language often contain redundancy and ambiguity. To construct a robust reasoning foundation, we implement a KG refinement module that addresses two key challenges: entity canonicalization and hierarchical semantic aggregation.

Entity Canonicalization. We employ a semantic encoder ϕ (e.g., SBERT) to map the textual description of an entity to a dense vector representation $\mathbf{v}_e = \phi(e)$. To address synonymous concepts (e.g., "CO2 output" vs. "Carbon emissions"), we perform pairwise comparisons. Two entities e_i and e_j are merged into a single canonical node if their cosine similarity exceeds the threshold δ:

$$\text{merge}(e_i, e_j) \iff \frac{\mathbf{v}_{e_i} \cdot \mathbf{v}_{e_j}}{\|\mathbf{v}_{e_i}\| \|\mathbf{v}_{e_j}\|} \geq \delta.$$

Hierarchical Semantic Aggregation. Following node-level refinement, we compute holistic representations for each GRI Standard to capture inter-standard dependencies. For standard S with constituent nodes $\mathcal{E}_S$, we calculate the aggregated embedding $\mathbf{v}_S$ via mean-pooling:

$$\mathbf{v}_S = \frac{1}{|\mathcal{E}_S|} \sum_{e \in \mathcal{E}_S} \mathbf{v}_e$$

This aggregation computes the semantic centroid of the standard's requirements. We then construct a weighted similarity matrix $\mathbf{W} \in \mathcal{R}^{K \times K}$ (where K is the total number of standards), with each entry $W_{ij} = \cos(\mathbf{v}_{S_i}, \mathbf{v}_{S_j})$ represents semantic affinity between standards. This matrix serves as a retrieval prior, enabling the system to identify structurally related requirements beyond exact keyword matching, thereby enhancing evidence recall robustness..

3.3 Multimodal Evidence Retrieval and Online Scoring

The final component processes corporate reports and executes compliance audits, as illustrated in Figure 1(3). To address the "multimodal gap" where crucial evidence resides in non-textual formats, we implement a specialized preprocessing pipeline that projects visual and tabular data into a unified semantic space for retrieval.

Multi-modal Document Processing. An uploaded document $\mathcal{D}$ is parsed into segments $\mathcal{D} = \{u_1, u_2, \ldots, u_N\}$, where each segment u_i belongs to one of three modalities: text ($\mathcal{U}_{\text{text}}$), tables ($\mathcal{U}_{\text{tab}}$), or images ($\mathcal{U}_{\text{img}}$). We apply modality-specific transformation Ψ to convert all segments into textual representations:

$$x_i = \Psi(u_i) = \begin{cases} u_i & \text{if } u_i \in \mathcal{U}_{\text{text}} \\ \text{Serialize}(u_i) & \text{if } u_i \in \mathcal{U}_{\text{tab}} \\ \text{VLM}(u_i) & \text{if } u_i \in \mathcal{U}_{\text{img}}, \end{cases}$$

where $\text{VLM}(\cdot)$ denotes a vision-language model to generate descriptive captions for charts and infographics (preserving quantitative values), and $\text{Serialize}(\cdot)$ linearizes table structures. Processed segments are then embedded using the sementic encoder ϕ (defined in Sec. 3.2) to form an evidence pool $\mathcal{P} = \{(\phi(x_i), x_i)\}_{i=1}^{N}$.

Hybrid Retrieval Mechanism. For a given GRI standard node $\mathcal{S}$, the system formulates a query vector $\mathbf{q}_S$ and employs hybrid retrieval to identify relevant evidence $\mathcal{E}^* \subset \mathcal{P}$. The relevance score for candidate segment x_i combines semantic and lexical matching:

$$\begin{aligned} \text{Score}(\mathcal{S}, x_i) = \alpha \cdot \cos(\mathbf{q}_S, \phi(x_i)) \\ + (1 - \alpha) \cdot \text{BM25}(\text{q}_S, x_i) \end{aligned}$$

where α balances dense semantic similarity with sparse keyword matching via BM25.

LLM-based Compliance Scoring. Retrieved evidence $\mathcal{E}^*$ and scoring criteria $\mathcal{C}_{\text{score}}$ (from Eq. (1)) are provided to the scoring LLM, denoted as f_{eval}, to generate the compliance assessments:

$$(y, \mathcal{J}) = f_{\text{eval}}(\mathcal{E}^*, \mathcal{C}_{score}),$$

where $y \in \{1, 2, 3, 4, 5\}$ represents the compliance grade and $\mathcal{J}$ contains natural language justification. This design ensures a traceable audit trail, explicitly linking evidence to regulatory requirements through the knowledge graph.

4 Implementation and Demonstration

4.1 Python Toolkit

To facilitate reproducibility and foster community engagement, we have released the core components of our system as an open-source Python toolkit, available at our GitHub[1]. This toolkit is designed with a modular architecture, consisting of:

Data Ingestion Module. Leveraging MinerU[2] ((Wang et al., 2024a), (Niu et al., 2025), (He et al., 2024)) for high-fidelity PDF parsing, this module performs advanced layout analysis to accurately extract and serialize multi-modal content—including complex tables, diagrams, and cross-page text flows—from unstructured regulatory documents and corporate reports.

KG Construction Engine. The engine implements the logic for building and refining the regulatory knowledge graph, serving as the foundational layer for structuring unstructured compliance standards. It employs semantic extraction algorithms to transform raw regulatory text (e.g., GRI Standards) into a structured graph format, identifying core entities—such as disclosure requirements and metric definitions—and establishing hierarchical relationships between them. Furthermore, the engine includes a refinement layer that resolves entity ambiguity and enforces schema consistency, ensuring a reliable knowledge base for downstream reasoning.

Evaluation Pipeline. This pipeline encapsulates the comprehensive logic for evidence retrieval and LLM-based compliance scoring. Crucially, it utilizes the explicit mapping between specific GRI standards and page numbers—voluntarily disclosed by companies in their sustainability reports (typically within the *GRI Content Index*)—as the primary reference for evidence localization. By anchoring the retrieval process to these self-disclosed page references, the system employs a Retrieval-Augmented Generation (RAG) workflow to precisely align corporate data with regulatory nodes. Beyond simple scoring, the pipeline manages the generation of natural language justifications and citation mapping, ensuring that every compliance assessment is transparent, auditable, and grounded in the specific document segments identified by the reporting entity.

Researchers and developers can utilize this toolkit to deploy their own local instances or extend the framework to support additional sustainability standards beyond GRI.

4.2 System Demonstration

Building upon the proposed toolkit, we developed a web-based system to demonstrate ESG-KG's capabilities in a real-world scenario. We present a compliance assessment case study using the *Delta Electronics 2023 ESG Report*. Figure 2 illustrates

[1] https://github.com/cnclabs/website.kg.esg.demo.git

[2] https://github.com/opendatalab/MinerU

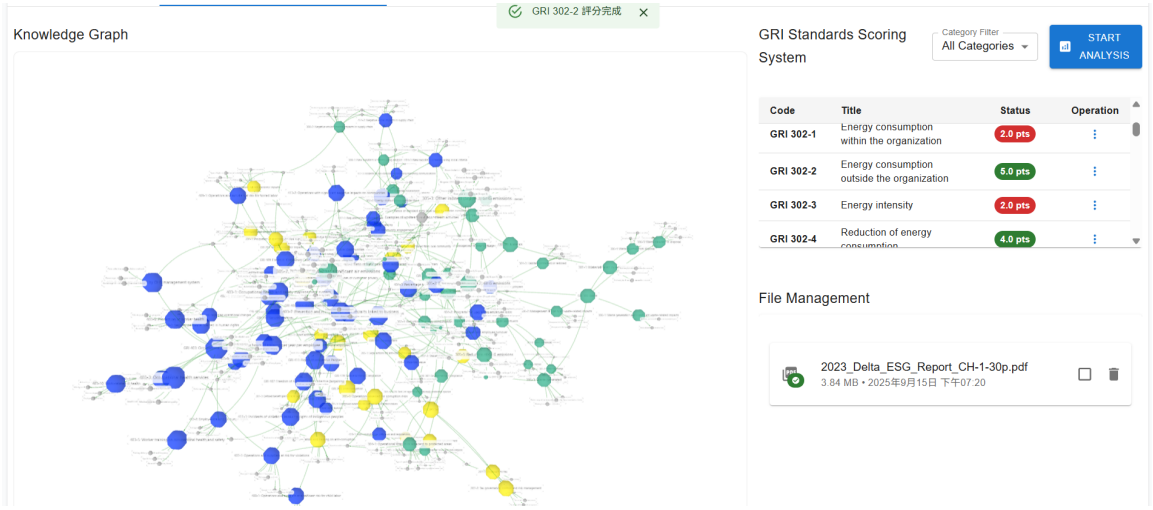

Figure 2: The ESG-KG system interface, featuring knowledge graph visualization for regulatory navigation and the automated compliance scoring panel for GRI standards

the system interface during the audit process, featuring a split-screen layout with the knowledge graph visualization on the left and the scoring control panel on the right.

Document Upload and Processing. As shown in the bottom-right "File Management" module, users upload reports such as the `2023_Delta_ESG_Report_CH-1-30p.pdf`. The system automatically executes the multi-modal pipeline, decomposing the report into semantic units—text segments, tables, and infographics. These units are serialized, embedded, and indexed in a vector database to form the evidence pool.

Knowledge Graph Visualization and Retrieval. The left panel displays an interactive visualization of the regulatory Knowledge Graph, where nodes represent specific standards and compliance concepts. For compliance assessment against **GRI 302: Energy** standards, this graph serves as the navigational backbone. The system performs grounded retrieval to extract relevant evidence from the document while simultaneously querying the visualized GRI regulatory KG to retrieve corresponding logic, such as emission thresholds. This dual-graph alignment ensures that extracted evidence semantically maps to precise regulatory requirements.

Automated Scoring with Traceable Evidence. The top-right "GRI Standards Scoring System" presents the evaluation results. The model evaluates disclosure completeness and accuracy, assigning granular compliance scores as visible in the "Status" column: **2.0 pts** for GRI 302-1 (partial compliance), **5.0 pts** for GRI 302-2 (full compliance), and **4.0 pts** for GRI 302-4. Real-time feedback is provided, as seen in the top overlay notification confirming the completion of the GRI 302-2 assessment. Users can filter results by category or initiate new evaluations using the "START ANALYSIS" button, enabling an efficient and interactive verification process.

Interactive Exploration. The interface allows users to explore the knowledge graph structure, trace evidence provenance, and review the reasoning chain connecting regulatory requirements to corporate disclosures. This transparency enables auditors to validate system decisions and identify areas requiring human expert review.

5 System Evaluation

5.1 Experimental Setup

To rigorously evaluate the retrieval performance of ESG-KG, we constructed a custom dataset named **ESG-50**. This dataset was collected by the authors and consists of publicly available sustainability reports from 50 representative companies in Taiwan. Published between 2024 and 2025, these reports cover a wide range of industrial sectors—including technology, finance, and manufacturing—and strictly adhere to the Global Reporting Initiative (GRI) standards, ensuring a diverse and standardized testbed for our experiments.

Ground Truth Construction. We leverage the *GRI Content Index* typically included in compliant reports, which explicitly maps each GRI disclosure item (e.g., GRI 302-1) to specific page numbers or sections. For a given standard $\mathcal{S}$, the ground

truth evidence set $\mathcal{E}_{\text{gt}}$ consists of all text segments, tables, and charts located on the referenced pages.

Baselines. We compare ESG-KG against two established retrieval approaches:

- **BM25**: Keyword-based retrieval using exact term matching between GRI standard descriptions and document segments.
- **Dense retrieval**: Semantic search using a pretrained embeddings (OpenAI text-embedding-3) without KG guidance or multi-modal processing.

5.2 Retrieval Performance

Table 1 presents the comparative results using Recall@K and NDCG@K metrics. BM25 achieves the lowest performance (Recall@5: 42.3%) due to the vocabulary mismatch between regulatory terminology and corporate reporting language. Dense retrieval improves substantially (Recall@5: 61.5%) by capturing semantic similarities, butstruggles with quantitative data in tables and charts.

ESG-KG significantly outperforms both baselines, achieving Recall@5 of 84.1%. This improvement stems from two key capabilities: (1) **Multi-modal parsing** successfully retrieve evidence from tables, charts and infographics, which constitute a large portion of GRI data, and (2) **KG-guided hybrid retrieval** utilizes the structured "Standard Graph" to expand queries with semantically related requirements, ensuring retrieval captures logically relevant compliance evidence rather than merely semantically similar text.

Table 1: Retrieval Performance on ESG-50 Dataset.

Metric	BM25	Dense	ESG-KG (Ours)
Recall@5	0.423	0.615	**0.841**
Recall@10	0.518	0.702	**0.915**
NDCG@10	0.387	0.594	**0.812**

6 Conclusion

We presented ESG-KG, a system that automates ESG compliance assessment through multi-modal information extraction and knowledge graph construction. By integrating specialized multi-modal document processing with a structured GRI-based knowledge graph, ESG-KG enables precise, evidence-based compliance verification across textual, tabular, and visual content. Our demonstration showcases how the system reduces manual assessment effort while enhancing transparency, accuracy, and auditability in ESG reporting. The system provides a practical foundation for scalable, standardized compliance validation aligned with global sustainability frameworks. Future work will extend coverage to additional ESG standards beyond GRI and incorporate user feedback mechanisms for iterative model refinement.

6.1 Limitations

While ESG-KG demonstrates effective automated compliance assessment, several limitations warrant consideration. First, the knowledge graph construction currently focuses on GRI Standards and does not cover all regional or industry-specific ESG frameworks (e.g., SASB, TCFD). Second, the multi-modal extraction pipeline may face challenges with highly unconventional document layouts or proprietary infographic formats. Third, although the LLM-based scoring provides generally consistent assessments, it may require human oversight for edge cases involving complex regulatory interpretations. Finally, the system has been primarily developed and evaluated on English-language reports; multilingual support remains an area for future development.

Future work will focus on expanding standard coverage, improving robustness to diverse document formats, and incorporating mechanisms for human-in-the-loop validation in ambiguous cases.

6.2 Ethics Statement

ESG-KG is designed as a decision-support tool to assist auditors and stakeholders in evaluating ESG disclosures, serving to augment rather than replace human judgment or definitive compliance determinations. We emphasize that automated compliance systems should support expert analysis, particularly in contexts involving complex regulatory interpretations or significant stakeholder impact.

We acknowledge the responsibility to mitigate potential biases inherited from training data or reporting standards. To this end, the system's transparency features—including traceable evidence chains and explicit reasoning—are intentionally designed to enable human reviewers to validate automated assessments and identify potential errors. Ultimately, users must critically evaluate system outputs and retain full accountability for final compliance decisions.

Acknowledgements

This research was supported in part by an industrial collaboration project with Delta Electronics, Inc.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 1877–1901.

Yingjie Chen, Hao Liu, Yulu Liu, Jiawei Xie, Ruikang Yang, Hanming Yuan, Yanjun Fu, Peter Y. Zhou, Qi Chen, James Caverlee, and Irwin Li. 2025. GraphCheck: Breaking long-term text barriers with extracted knowledge graph-powered fact-checking. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 14976–14995.

Tanay Gupta, Tushar Goel, and Ishan Verma. 2025. Exploring multimodal language models for sustainability disclosure extraction: A comparative study. In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 141–149, Albuquerque, New Mexico. Association for Computational Linguistics.

Chaoyue He, Xin Zhou, Yi Wu, Xinjia Yu, Yan Zhang, Lei Zhang, Di Wang, Shengfei Lyu, Hong Xu, Wang Xiaoqiao, Wei Liu, and Chunyan Miao. 2025. ESGenius: Benchmarking LLMs on environmental, social, and governance (ESG) and sustainability knowledge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 14623–14664, Suzhou, China. Association for Computational Linguistics.

Conghui He, Wei Li, Zhenjiang Jin, Chao Xu, Bin Wang, and Dahua Lin. 2024. Opendatalab: Empowering general artificial intelligence with open datasets. *arXiv preprint arXiv:2407.13773*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training fornatural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FACTScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Junbo Niu, Zheng Liu, Zhuangcheng Gu, Bin Wang, Linke Ouyang, Zhiyuan Zhao, Tao Chu, Tianyao He, Fan Wu, Qintong Zhang, Zhenjiang Jin, Guang Liang, Rui Zhang, Wenzheng Zhang, Yuan Qu, Zhifei Ren, Yuefeng Sun, Yuanhong Zheng, Dongsheng Ma, Zirui Tang, Boyu Niu, Ziyang Miao, Hejun Dong, Siyi Qian, Junyuan Zhang, Jingzhou Chen, Fangdong Wang, Xiaomeng Zhao, Liqun Wei, Wei Li, Shasha Wang, Ruiliang Xu, Yuanyuan Cao, Lu Chen, Qianqian Wu, Huaiyu Gu, Lindong Lu, Keming Wang, Dechen Lin, Guanlin Shen, Xuanhe Zhou, Linfeng Zhang, Yuhang Zang, Xiaoyi Dong, Jiaqi Wang, Bo Zhang, Lei Bai, Pei Chu, Weijia Li, Jiang Wu, Lijun Wu, Zhenxiang Li, Guangyu Wang, Zhongying Tu, Chao Xu, Kai Chen, Yu Qiao, Bowen Zhou, Dahua Lin, Wentao Zhang, and Conghui He. 2025. Mineru2.5: A decoupled vision-language model for efficient high-resolution document parsing.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training. In *OpenAI Technical Report*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. pages 8–9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. In *Journal of Machine Learning Research*, pages 1–67.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024a. Mineru: An open-source solution for precise document content extraction.

Jinhong Wang, Qixiu Song, Li Qian, Hang Li, Qinghua Peng, and Jianming Zhang. 2024b. SubstationAI: Multimodal large model-based approaches for analyzing substation equipment faults. *CoRR*, abs/2412.17077.

Qintong Zhang, Bin Wang, Victor Shea-Jay Huang, Junyuan Zhang, Zhengren Wang, Hao Liang, Conghui He, and Wentao Zhang. 2024. Document parsing unveiled: Techniques, challenges, and prospects for structured information extraction. *arXiv preprint arXiv:2410.21169*.