MMESGBench: Benchmarking Multimodal Understanding and Complex Reasoning in ESG Documents

Supplementary Material

A ESG Document Source

To ensure that MMESGBench accurately reflects the real-world diversity and complexity of ESG disclosures, we curate a balanced and representative collection of 45 PDF documents spanning the full spectrum of ESG reporting practices. The selected documents are drawn from three major categories: (1) Corporate ESG Reports, (2) ESG Standards and Frameworks, and (3) Government and International Organization Documents. This taxonomy is informed by both ESG industry practices and widely adopted disclosure frameworks.

For the Corporate ESG Reports category, we include both comprehensive annual sustainability reports and CDP Climate Questionnaire responses from leading global corporations, such as Apple, Microsoft, and Alibaba. These documents capture bottom-up reporting practices and voluntary disclosures, often characterized by dense content spanning hundreds of pages, extensive use of visual layouts, and fine-grained reporting on emissions, governance, and social initiatives. The ESG Standards and Frameworks category encompasses documents from established standard-setting bodies, grouped across four ESG sub-dimensions: environmental (e.g., ISO 14001, TCFD, GHG Protocol), social (e.g., ISO 26000, SA8000), governance (e.g., ISO 37001, OECD), and comprehensive/multi-dimensional frameworks (e.g., GRI, SASB, TNFD, IFRS). These documents contain structured guidelines, metrics, and disclosure templates that are essential for evaluating regulatory alignment and compliance-oriented reasoning. Government and International Organization Documents category includes policy reports and regulatory frameworks issued by authoritative institutions such as the IPCC, UN, SDG Secretariat, and NGFS. These sources represent top-down global sustainability agendas and cover climate risk, transition planning, and sustainable finance regulations.

Figure A1: Taxonomy of ESG Document Sources in MMESGBench

Across all categories, documents were selected based on the following criteria: (1) presence of multimodal content (including tables, charts, layout structures, and images), (2) coverage of all three ESG pillars, and (3) diversity in origin (corporate, regulatory, and institutional). The resulting corpus ranges in length several pages to over 2,000 pages, with an average of 157 pages per document, and

Table A1: Detailed ESG documents in MMESGBench. *Document names in the first column are hyperlinked to publicly accessible sources from which the original PDFs were obtained.*

Doc name	#page	Doc type	#QA
CDP Corporate Scoring 2024	19	Environment-related Standards	13
AR6 Synthesis Report 2023	186	Government & Organization Documents	37
Microsoft CDP Response 2023	169	CDP Climate Responses	31
ipcc-ar6-wg3	2042	Government & Organization Documents	42
GRI Standards 2021	677	Comprehensive Standards	35
WHO GHG	22	Government & Organization Documents	16
WHO Social	32	Government & Organization Documents	11
TNFD Summary	7	Comprehensive Standards	11
TNFD GRI Mapping	15	Comprehensive Standards	11
ISO 26000	20	Social-related Standards	14
GHG Survey	51	Environment-related Standards	21
TCFD Guidence 2020	133	Environment-related Standards	28
SUSTAINALYTICS ESG	27	Comprehensive Standards	17
SDG 2024	51	Government & Organization Documents	21
SDG Agenda	41	Government & Organization Documents	13
SBTi-criteria	23	Environment-related Standards	11
SBTI Zero v2	133	Environment-related Standards	23
Alphabet CDP response 2024	140	CDP Climate Responses	23
Apple CDP Questionnaire 2023	133	CDP Climate Responses	23
Meta CDP Survey 2023	141	CDP Climate Responses	23
SASB Education	18	Comprehensive Standards	11
SASB Agricultural Products	30	Comprehensive Standards	15
SASB Health Care Distributors	17	Comprehensive Standards	10
SASB Software & IT Services	29	Comprehensive Standards	11
SA8000 2016	138	Social-related Standards	20
PRI 2024	79	Government & Organization Documents	25
OECD Employment	293	Governance-related Standards	26
OECD Net Zero+	286	Environment-related Standards	21
LSEG ESG	34	Comprehensive Standards	15
LEED 2025	319	Environment-related Standards	30
NGFS 2024	37	Government & Organization Documents	14
ISO 14001	44	Environment-related Standards	14
ISO 37001	54	Governance-related Standards	11
ISO 37000	19	Governance-related Standards	7
IFRS S2 Vol29	6	Comprehensive Standards	6
IEA 2024	398	Government & Organization Documents	30
GHG Public	112	Environment-related Standards	21
Gender 2024	62	Governance-related Standards	16
Alibaba Group 2023 ESG Report	242	Annual ESG Reports	35
Microsoft 2024 ESG Report	88	Annual ESG Reports	34
Google 2024 environmental report	86	Annual ESG Reports	27
Credo ESG Report 2024	44	Annual ESG Reports	17
Dell ESG Report	116	Annual ESG Reports	30
Amazon sustainability report 2023	98	Annual ESG Reports	29
WELL 2020	366	Environment-related Standards	34

covers seven document types and detailed in Table A1, hyperlinks to all document sources are embedded in the document name column for transparency and reproducibility. The hierarchical structure of our source taxonomy is illustrated in Figure A1. This collection provides a robust foundation for evaluating multimodal document understanding, retrieval, and reasoning across heterogeneous ESG contexts.

B Multimodal LLM-based QA Generation

To complement the QA generation description in the main text, we detail here the prompt design used to guide single-page QA construction via a multimodal MLLM. The prompt is tailored for ESG reasoning and reflects real-world analytical needs in sustainability domains.

We adopt an example-based prompting strategy, where the model is given a small set of carefully designed QA examples to illustrate the desired output style, format, and reasoning depth. These examples cover diverse question types, including factual lookup, quantitative computation, and regulatory compliance checks. The model is expected to mimic these patterns when generating new QA pairs based on the input document page.

The prompt emphasizes high-quality reasoning and imposes the following constraints:

- Reasoning depth: Only generate questions that require calculation, aggregation, comparison, or regulatory interpretation; avoid superficial or copy-based queries.
- Answerability filtering: At least 20% of questions should be marked as Not answerable if sufficient evidence is not visible in the input image.
- Modality awareness: The model should jointly consider visual layout, tabular content, and text, and explicitly label the modality used to answer each question.
- Fidelity to domain-specific language: Terminologies and ESG-specific phrasing must be preserved without paraphrasing or generalization.

To support consistent output formatting, the prompt provides a structured JSON schema for each QA pair, including fields such as doc_id, question, answer, evidence_sources, and answer_format. The complete prompt, along with a representative set of generated examples, is included below to facilitate reproducibility.

The full prompt, shown below, includes structured output instructions and self-consistency checks to guide generation fidelity. Each resulting QA is expected to reflect fine-grained multimodal reasoning grounded in layout-aware ESG content.

Prompt Used for ESG QA Generation

You are an ESG domain expert assistant. Your task is to generate high-quality, reasoning-based ESG-focused QA pairs from the given PDF page image from ESG-related document.

Requirements:

1. Focus on the most relevant Environmental, Social, and Governance questions that a sustainability analyst or ESG auditor would care about.
2. Only generate questions that require reasoning, comparison, calculation, or deeper understanding. Avoid lookup or copy-paste style questions.
3. For each page, generate 1 to 2 high-quality QA pairs. Prioritize those that require inference, reasoning, or calculation (chain-of-thought).
4. At least 20% of your questions should be marked as 'Not answerable' if the answer cannot be found in the image.
5. Avoid altering domain-specific or technical keywords in the content.

Self-Correction Check (Internal Thought Process before outputting): Does this question *genuinely* require multi-step reasoning or deep inference based *only* on the text?

Output Format: For each QA pair, output a JSON object with the following fields:

- `doc_id`: the PDF file name,
- `doc_type`: the document type,
- `question`: the question based on the content of the image,
- `answer`: the answer to the question no more than 8 words (if the answer requires additional pages, set it as "Not answerable"),
- `evidence_sources`: a string representation of a list of sources (e.g., `["Chart", "Table", "Pure-text (Plain-text)", "Generalized-text (Layout)", "Image"]`),
- `answer_format`: a string representing the answer type (e.g., `"Str"`, `"Int"`, `"Float"`, `"List"`, `"None"`).

Output Example:

```
{
  "doc_id": "AR6 Synthesis Report Climate Change 2023.pdf",
  "doc_type": "Government & International Organization Documents",
  "question": "Using the IPCC report, calculate the total additional population exposed
      to coastal flooding events by 2040 under SSP2-4.5 scenario.",
  "answer": "19.62",
  "evidence_pages": "[116]",
  "evidence_sources": "['Image', 'Generalized-text (Layout)']",
  "answer_format": "Float"
},
{
  "doc_id": "SDG Agenda.pdf",
  "doc_type": "Government & International Organization Documents",
  "question": "How many hectares of terrestrial forest must be restored annually in each
      region to achieve SDG Goal 15 by 2030?",
  "answer": "Not answerable",
  "evidence_sources": "[]",
  "answer_format": "None",
  "evidence_pages": "[]"
},
{
  "doc_id": "SASB Health Care Distributors.pdf",
  "doc_type": "Comprehensive Standards",
  "question": "What are the key strategies to reduce health and safety risks of products
      sold? Write the answer in the list format",
  "answer": "['Labeling', 'training', 'education', 'right-sizing packaged dosages']",
  "evidence_sources": "['Pure-text (Plain-text)']",
  "answer_format": "List",
  "evidence_pages": "[10]"
},
{
  "doc_id": "IFRS S2 Vol29.pdf",
  "doc_type": "Comprehensive Standards",
  "question": "If a fleet consumes 50,000 litres of fuel and transports 2,500 revenue
      tonne-kilometres, what is the payload fuel economy? Write the answer in
      Litres/RTK",
  "answer": "20",
  "evidence_sources": "['Table', 'Pure-text (Plain-text)']",
  "answer_format": "Int",
  "evidence_pages": "[4, 5]"
}
```

C Cross-page QA Generation

To support multi-page reasoning over long ESG documents, we construct a semantic clustering pipeline that groups content-relevant pages before QA generation. Each page is rendered as an image and encoded into 128-dimensional semantic embeddings using PaliGemma-3B, which fuses patch-wise visual and token-level semantic features. These embeddings are indexed using a FAISS vector library [**?**], enabling fast similarity retrieval at scale.

We perform nearest-neighbor search and clustering within the embedding space to identify semantically coherent page groups. These clusters typically correspond to ESG reporting themes such as emission metrics, risk assessments, or governance structures. The grouping process is unsupervised and does not require predefined section headers, allowing flexible adaptation across heterogeneous document structures. An illustration of this workflow is provided in Figure A2. Once clusters are formed, a multimodal LLM is prompted to generate questions that require cross-page reasoning within each group. The prompt emphasizes multi-hop question types, including aggregation (e.g., synthesizing KPI categories), temporal comparison (e.g., tracking changes), and causal or referential reasoning (e.g., linking risk disclosures to mitigation actions). This method captures long-range dependencies without exceeding model context limits, enabling scalable QA generation for ESG documents.

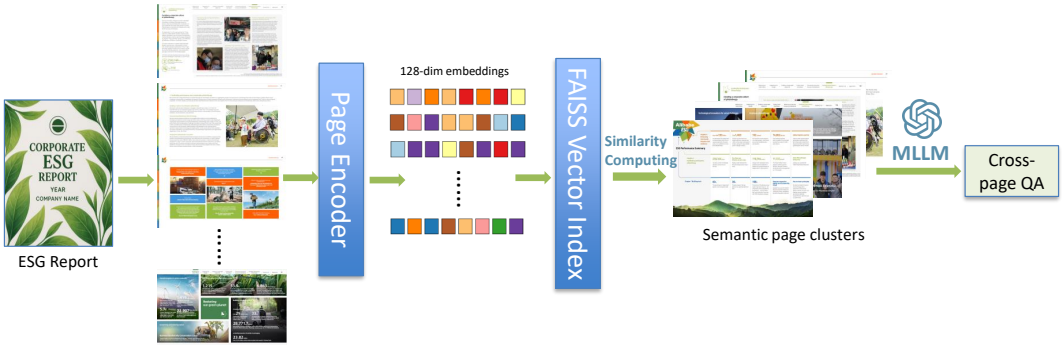


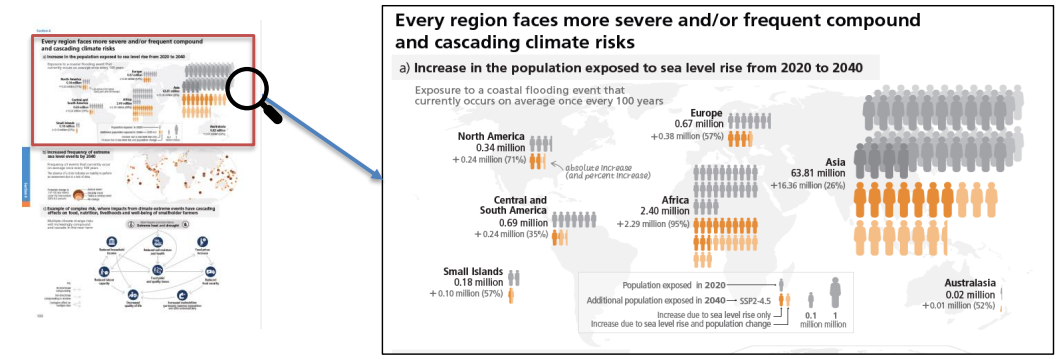
Figure A2: Detailed cross-page QA generation framework.

D QA examples

To highlight the reasoning diversity and multimodal challenges embedded in MMESGBench, we present four representative QA examples that reflect distinct question archetypes frequently encountered in ESG document understanding. These examples span a range of reasoning tasks—from visual computation to multi-page score derivation—and demonstrate the dataset's focus on non-trivial, contextually grounded inference beyond simple text lookup. Each QA is paired with multimodal evidence (e.g., tables, charts, layout-dependent text) and demands precise cross-modal reasoning, often incorporating numerical synthesis, temporal alignment, or policy-driven logic. The examples below also serve to illustrate key reasoning categories captured in MMESGBench.

Visual-based Numerical Reasoning Example: Total flood-exposed population by region (Figure A3) This question asks the model to calculate the total additional population exposed to coastal flooding by 2040 under a given scenario (SSP2-4.5). The relevant information is embedded within a high-density infographic combining layout-anchored population figures, visual grouping by geography, and multiple numerical annotations. Answering correctly requires spatial parsing of all affected regions, semantic filtering of absolute deltas, and aggregation across heterogeneous visual segments. This example highlights MMESGBench's inclusion of fine-grained visual reasoning, a task type rarely supported in existing QA benchmarks.

Multi-page Tabular Score Aggregation Example: CDP band score based on threshold mapping (Figure A4) This example requires determining a company's final CDP band score by aligning individual subdomain scores (Climate, Forests, Water) with a tabular scoring rubric and applying CDP's rule:



QA Example:

```
"doc_id": "AR6 Synthesis Report Climate Change 2023.pdf",
"doc_type": "Government & Organization Documents",
"question": "Calculate the total additional population exposed to coastal flooding by 2040 under SSP2-4.5 scenario.",
"answer": "19.62",
"evidence_pages": "[116]",
"evidence_sources": "['Image', 'Generalized-text (Layout)']",
"answer_format": "Float"
```

Analysis:

To calculate the **total additional population exposed to coastal flooding events by 2040** under the **SSP2-4.5 scenario**, we sum up the additional exposed population values from each region as shown in the image:

- **North America**: +0.24 million
- **Central and South America**: +0.24 million
- **Europe**: +0.38 million
- **Africa**: +2.29 million
- **Asia**: +16.36 million
- **Australasia**: +0.01 million
- **Small Islands**: +0.10 million

Total Additional Population =

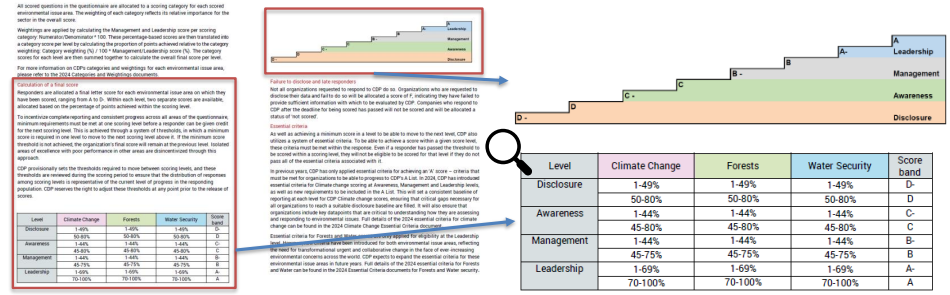
0.24 + 0.24 + 0.38 + 2.29 + 16.36 + 0.01 + 0.10 = **19.62 million**

✅ **Answer: 19.62 million additional people** will be exposed to coastal flooding by 2040 under the SSP2-4.5 scenario.

Figure A3: Visual-based Numerical Reasoning QA.

the lowest performing subscore determines the overall band. The relevant table and policy description are distributed across different visual sections, requiring the model to locate and integrate multimodal evidence across pages.

Answering this question involves accurate table lookup, threshold alignment, and rule-based aggregation, reflecting a core real-world ESG assessment task. It highlights MMESGBench's emphasis on layout-aware reasoning, numerical interpretation, and standard-specific logic application—capabilities essential for automating ESG evaluation and ensuring fidelity to reporting frameworks.



QA Example:

```
"doc_id": "CDP_Full_Corporate_Scoring_Introduction_2024.pdf",
"doc_type": "Environment-related Standards",
"question": "If a company scores 92% in Climate Change, 20% in Forests, and 90% in Water Security, what is the final CDP score? ",
"answer": "D",
"evidence_sources": "['Table','Image']",
"answer_format": "Str",
"evidence_pages": "[10, 11]"
```

Analysis:

Final CDP Score Determination

To determine the final CDP score for the company, we need to evaluate its scores in each category and then determine the overall score based on the provided table.

Company Scores: Climate Change: 92% | Forests: 20% | Water Security: 90%

1. **Climate Change (92%)** – Falls into the *"Leadership"* level → Score band: **A**
2. **Forests (20%)** – Falls into the *"Disclosure"* level → Score band: **D**
3. **Water Security (90%)** – Falls into the *"Leadership"* level → Score band: **A**

Since the **lowest score is D** (Forests), the **final CDP score** is determined by that.

✅ **Final Score: D**

Figure A4: Multi-page Tabular Score Reasoning QA.

Cross-page Numerical Comparison and Arithmetic Example: Computing annual emission reduction needs (Figure A5) This QA requires the model to compute the average annual reduction in Scope 1 and 2 emissions from a 2020 to 2030. Key values appear on non-adjacent pages, requiring inter-page reference resolution and contextual linking of base year and target year figures. The question cannot be answered without performing a multi-step subtraction and division operation. This exam-

ple showcases MMESGBench's capacity to benchmark multi-hop, cross-page arithmetic reasoning—a critical but underexplored ability for models deployed in sustainability analysis.

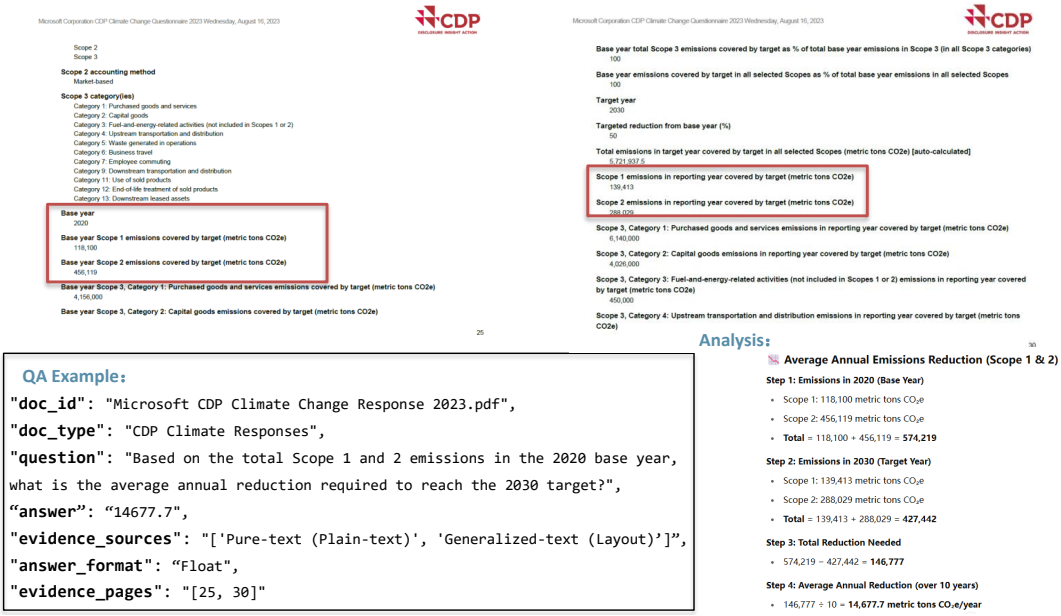


Figure A5: Cross-page Numerical Comparison and Arithmetic QA.

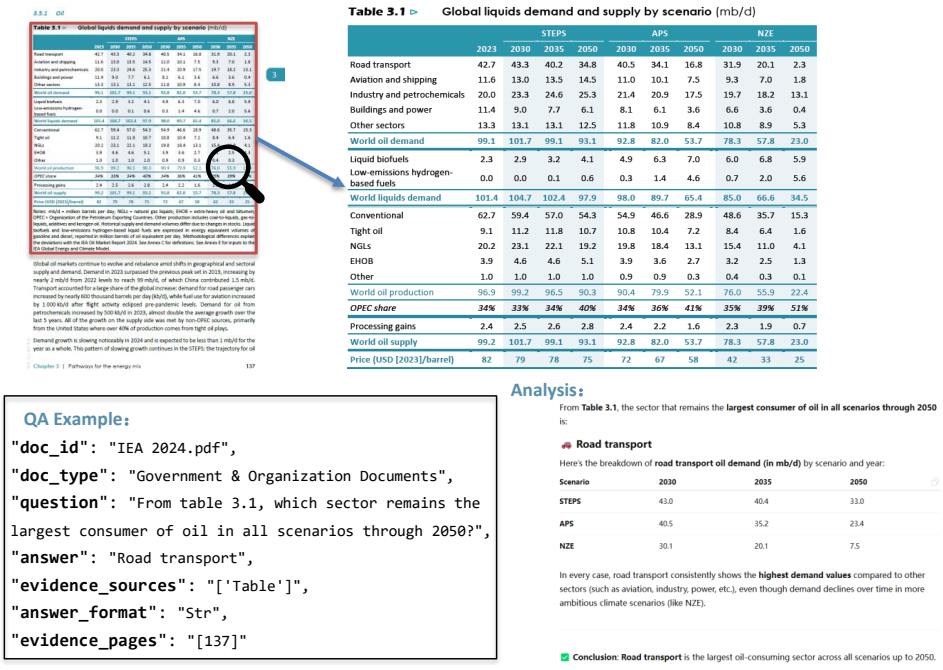
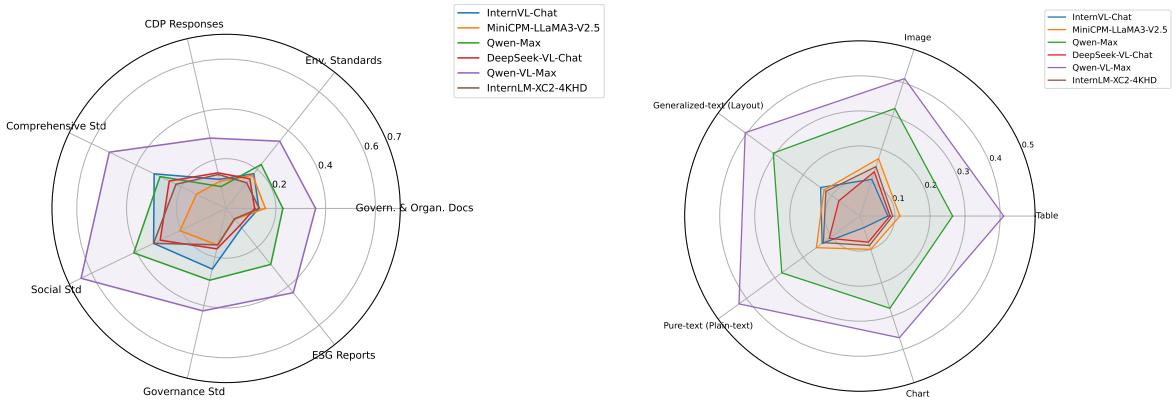


Figure A6: Trend Inference from Structured Tables QA.

Trend Inference from Structured Tables Example: Oil demand dominance across climate scenarios (Figure A6) This QA instance requires identifying the sector that consistently remains the largest consumer of oil across all years and climate scenarios presented in a dense tabular forecast. The model must interpret the table schema, locate relevant numerical rows for "Road transport," and perform comparative aggregation across columns representing multiple years (2030, 2035, 2050) and mitigation pathways (STEPS, APS, NZE). The answer is not a direct lookup but emerges from rec-



(a) Accuracy Distribution by Document Type

(b) Accuracy Distribution by Evidence Source

Figure A7: Comparative analysis of multimodal QA model performance across ESG document types and evidence modalities.

ognizing dominance patterns embedded in numerical distributions. This highlights MMESGBench's emphasis on reasoning over structured, high-dimensional quantitative data. It captures a key competency in ESG analytics—temporal trend interpretation under scenario variation—which is essential for tasks such as climate policy assessment, energy forecasting, and longitudinal sustainability analysis.

E Extended Evaluation Results

To better understand the performance characteristics of multimodal QA models, we conduct a detailed analysis of accuracy by document type and evidence modality, as illustrated in Figure A7.

Models generally perform better on structured documents such as Comprehensive Standards and Social-related Standards, where consistent formatting and layout provide clearer semantic cues. In contrast, Government and International Organization Documents remain challenging due to their high variability in length, dense policy-oriented language, and irregular structure, which require stronger contextual reasoning. Most models show notably lower accuracy in this category, highlighting architectural limitations in handling unstructured content.

Regarding evidence modality, models achieve higher accuracy on Pure-text (Plain-text) and Generalized-text (Layout), reflecting their relative strength in language-based and structured layout reasoning. Performance drops significantly on visual modalities—especially charts—which consistently yield the lowest scores. This suggests that fine-grained visual and numerical reasoning remains an underdeveloped capability. Among all models, Qwen-VL-Max achieves the most consistent and robust results across both document types and modalities, while others like InternVL and DeepSeek-VL perform well on text-heavy questions but struggle with visual elements. These findings emphasize the importance of multimodal grounding and layout-sensitive inference for ESG document understanding.