TCFDNAV: BENCHMARKING INDEX-GUIDED NAVIGATION FROM ANNUAL REPORTS

Anonymous authors
Paper under double-blind review

ABSTRACT

When a PDF provides a human-readable index, its mapping table should be used to locate information. Disclosure indexes based on the Task Force on Climate-related Financial Disclosures (TCFD) provide a representative example. However, existing approaches largely rely on similarity-based retrieval, while index-guided navigation remains underexplored. We therefore introduce TCFDNav, a benchmark for evaluating the recovery and execution of index-declared navigation relations in the TCFD setting. Rather than focusing on an isolated step, we address the complete navigation process, from an input annual report to the TCFD-related pages referenced by its index. Starting with 100 UK annual reports and extending to 17 external PDFs, we construct an annotated resource covering the full navigation chain. To support benchmark evaluation, we propose a three-stage reference framework that discovers complete indexes, recovers their declared relations, and resolves target addresses to produce traceable navigation maps. From a deployment perspective, Qwen3.5-27B performs competitively with leading proprietary models, showing the practical potential of open-weight models for this task. An anonymous review artifact is available at https://anonymous.4open.science/r/tcfdnav-review-artifact-28B4/.

1 INTRODUCTION

The Task Force on Climate-related Financial Disclosures (TCFD) provides a framework for reporting decision-useful information about how organizations govern, assess, and manage climate-related risks and opportunities (Task Force on Climate-related Financial Disclosures, 2017). Using these disclosures, however, first requires locating the relevant evidence, which companies may distribute across different sections of annual reports and companion documents. After locating it, TCFD-aligned evidence can inform capital-allocation decisions, analysts' earnings forecasts, assessments of financing outcomes, and evaluations of firms' strategic resilience to climate risk (Hewa et al., 2025; Ding et al., 2025; Wang et al., 2024). These downstream uses therefore depend on accurately determining where each reporting pillar is addressed and navigating to the corresponding pages.

As Figure 1 illustrates, index-guided navigation and similarity-based retrieval provide different routes to TCFD-related evidence. Similarity-based retrieval asks which content is relevant to a query, whereas index-guided navigation follows topic–location relations explicitly declared by the company. These company-declared links can provide page-level ground truth (Ahmed et al., 2025). In practice, the UK Financial Conduct Authority observed widespread use of TCFD grids or indexes and cross-references to disclosures elsewhere in annual reports or in standalone reports (Financial Conduct Authority, 2022). We collected 100 annual reports from the 2025 reporting cycle for companies in a fixed FTSE 100 roster, of which 70% contain a TCFD index. This motivates our core question: **when an annual report already provides a TCFD index that tells readers where to look, can an automated pipeline follow it as a human reader would to locate the specified pages?**

Although disclosure indexes are understandable to humans, their instructions are difficult for a model to interpret and execute. To read and operationalize them, a model must identify qualifying indexes, recover their pages, reconstruct topic–reference relations, preserve cross-page and cross-document references, and map printed page labels to PDF positions. These requirements go beyond index de-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

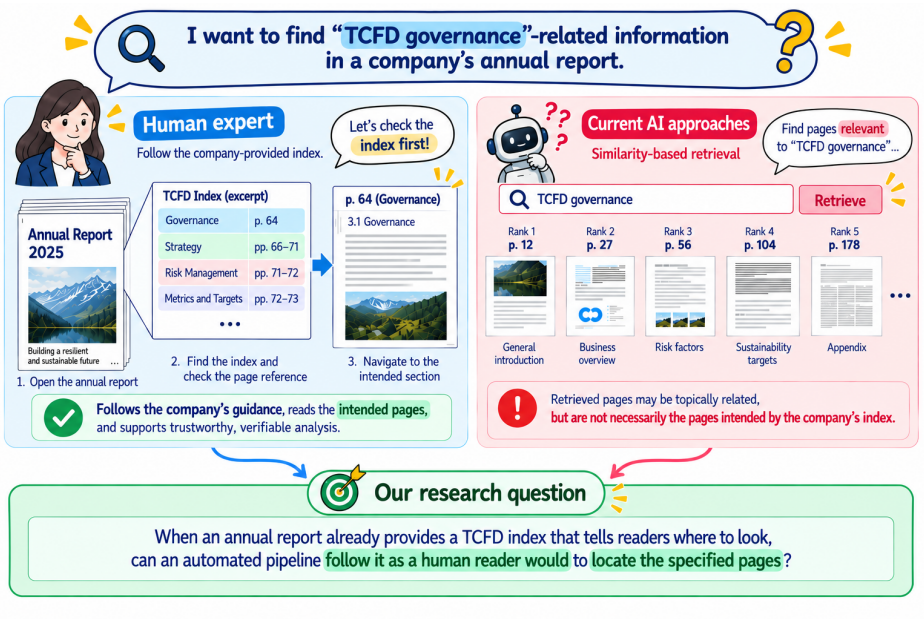

Figure 1: **Motivation for TCFDNav.** An analyst seeking TCFD Governance information can follow the disclosure index supplied by the company, whereas similarity-based retrieval ranks pages by relevance. The two routes use different location criteria and need not return the same page set.

tection or page-number extraction. Multimodal language models are promising because they jointly process text, layout, and document structure.

We therefore introduce **TCFDNav (TCFD Navigation)**, a benchmark for index-guided navigation that uses TCFD disclosures as its study setting. Alongside the benchmark, we implement an end-to-end framework that takes an annual report and available target PDFs as input. The framework comprises three modules: index discovery to locate the complete index, mapping recovery to extract its declared topic–reference relations, and target resolution to identify the corresponding pages in the target documents. The resulting navigation map retains the original references and source evidence, making each resolved destination traceable to the company's index.

Composing saved stage outputs matches all annotated navigation relations for 47/70 index-positive reports and correctly rejects all 30 reports without an index, with 93.01% physical-address relation F_1. The evaluation supplies the target registry, document aliases, and pagination probe positions.

Our contributions are twofold:

1. **TCFDNav benchmark.** We construct and publicly release TCFDNav, a dataset with layered annotations and evaluation resources for testing whether systems can interpret and follow human-readable disclosure indexes in annual reports. We evaluate leading open-source and proprietary multimodal language models across its component tasks and report a conditional composition of their saved outputs.

2. **End-to-end navigation framework.** We implement an end-to-end framework that locates disclosure indexes, extracts their topic–reference relations, resolves references to target documents and physical PDF pages, and produces traceable navigation maps linked to the original index evidence.

2 Related Work

Table 1 summarizes how prior work relates to the three stages studied in TCFDNav.

2.1 Disclosure Index and Structure Discovery

Prior work has studied navigation structures in long financial and ESG documents. FinTOC detects section titles and reconstructs hierarchical tables of contents in financial documents (Bentabet et al., 2020), while ESGDoc recovers the table-of-contents hierarchy of long ESG reports (Wang et al., 2023). Related work also models heading hierarchies from document images (Hu et al., 2022). These tasks primarily reconstruct a document's global organization.

Disclosure indexes pose a more specific discovery problem: the navigation structure already exists in the report but must be located before it can be interpreted. Gupta et al. (2025), for example, identify reports containing GRI indexes and manually annotate the index page ranges before extracting disclosure–page pairs. TCFDNav instead evaluates index discovery directly, requiring systems to identify the first index page and recover the complete index span.

2.2 Recovering Declared Disclosure Mappings

Disclosure indexes have also been used as sources of location information. Polignano et al. (2022) associate GRI identifiers with their page and context, while Gupta et al. (2025) extract disclosure–page pairs for downstream multimodal extraction. ESG-CID similarly converts requirement–page mappings from disclosure indexes into retrieval supervision (Ahmed et al., 2025).

These mappings differ from semantic relevance judgments. ClimRetrieve, for example, evaluates whether retrieved evidence is useful for expert-defined information needs (Schimanski et al., 2024); an indexed location, however, reflects what the company declared, not necessarily every passage that is semantically relevant. Related TCFD work likewise studies classification after candidate disclosure text has already been identified (Li et al., 2026). TCFDNav instead treats the complete declared relation set itself as the target, including the binding between topic, target document, and printed location.

2.3 Reference Resolution and Executable Navigation

A recovered page reference is not yet an executable address. Prior pagination work distinguishes visible or logical page labels from physical PDF pages (Déjean & Meunier, 2008). Daniel@FinTOC-2021 operationalizes this distinction by linking table-of-contents entries to document pages, converting them to PDF pages through pagination alignment, and locating the corresponding headings (Giguet & Lejeune, 2021). FinTOC-2022 further treats PDFs as bundles of documents that may have distinct structures and pagination sequences (Giguet & Lucas, 2022).

More generally, Sannier et al. (2017) separate cross-reference detection, contextual interpretation, and resolution to target provisions in legal documents. AEC-Bench extends reference tracing and navigation to multi-document engineering packages (Mankodiya et al., 2026). TCFDNav combines these concerns for disclosure indexes: references must be bound to the correct document, interpreted as printed locations, and resolved to physical PDF addresses where possible. Shared, ranged, non-contiguous, and unexecutable references are preserved, allowing index discovery, mapping recovery, and target resolution to be evaluated separately.

3 Framework

Starting from a source annual report, TCFDNav seeks the pages that its disclosure index designates for each TCFD topic, whether in that report or an available target PDF. Our framework finds the index, extracts its declared topic–document–location relations, and resolves them to document-specific addresses. The output is either `NO_INDEX` or a navigation map of company-designated pages grouped by topic and document, with original references, source evidence, and unresolved states retained. Figure 2 shows the three modules and their intermediate outputs.

3.1 Stage 1: Index Discovery

This module determines whether a qualifying disclosure index exists and recovers its complete *index span*, including continuation pages. A qualifying index explicitly links TCFD pillars or recom-

Table 1: Related navigation tasks. ✓ denotes a directly studied component; ◐ a related problem; – outside the primary scope.

Work	Stage 1 Index discovery	Stage 2 Mapping recovery	Stage 3 Target resolution
FinTOC (Bentabet et al., 2020)	◐	–	–
ESGDoc (Wang et al., 2023)	◐	–	–
Gupta et al. (Gupta et al., 2025)	◐	✓	–
ESG-CID (Ahmed et al., 2025)	–	✓	–
Daniel@FinTOC-2021 (Giguet & Lejeune, 2021)	◐	◐	✓
Sannier et al. (Sannier et al., 2017)	–	–	✓
AEC-Bench (Mankodiya et al., 2026)	–	–	✓
TCFDNav	✓	✓	✓

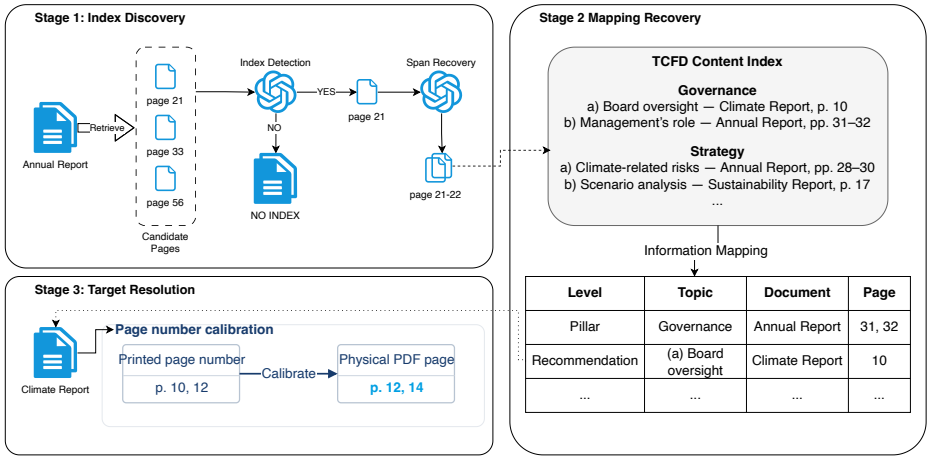

Figure 2: **Overview of TCFDNav.** Starting from an annual report, the framework discovers the complete TCFD index, recovers company-declared topic–document–page mappings, and resolves printed-page references to physical PDF pages. Documents and page numbers are illustrative.

mended disclosures to locations in a consolidated table, list, or reporting crosswalk; merely mentioning TCFD does not qualify. The output is the full index span or `NO_INDEX` when no qualifying index is found.

We rank pages by their relevance to TCFD-related queries to reduce the number of pages the model must read. Section 5.2 compares ranking methods that use extracted text or page images. The model checks candidate pages in physical-PDF order to identify the index start, then examines subsequent pages for continuing entries. Depending on the configuration, it receives text, page images, or both. The recovered index span is passed to mapping recovery.

3.2 Stage 2: Mapping Recovery

Given the index span, a multimodal language model extracts which document and location the index specifies for each TCFD topic. It reads the index pages together, using row and column alignment, headings, and entries that continue across pages. Each output record contains a topic unit, the referenced document as described in the index, and the cited location. A *topic unit* is a pillar, a specific recommended disclosure, or a group of disclosures that share a reference. Appendix B.5 summarizes the four TCFD pillars and their recommended disclosures.

The model records only the topic associations stated in the index. A reference shared by several disclosures remains shared; a pillar-level reference is not assigned to individual disclosures. Document names and locations stay together, so references to the same page number in different documents remain distinct. Stage 2 produces topic–reference records that associate each pillar- or

recommendation-level topic with the target document and printed-page reference stated in the index. Stage 3 then resolves these references to physical PDF locations.

3.3 Stage 3: Target Resolution

Stage 3 converts the references recovered in Stage 2 into addressable locations. This requires two steps: identifying the target document and resolving the cited printed page to its physical PDF page.

If a reference points to the annual report itself, we use the source annual-report PDF; if it points to a separate report, we identify the target PDF from the available report files using the report description in the index. For each target PDF, a multimodal language model maps printed page labels to physical PDF pages and identifies two-up layouts when present. The resulting locations are then assigned to both pillar- and recommendation-level topics recovered in Stage 2.

For example, a reference to printed page 35 of Aviva's *Climate-related Financial Disclosure 2025* is resolved as

$$address = (\text{Climate-related Financial Disclosure 2025}, 37, \text{FULL})$$

This resolved address can then be attached to the corresponding pillar- and recommendation-level topics.

4 Benchmark Construction

4.1 Corpus Selection and Acquisition

The United Kingdom provides a policy-relevant and concentrated setting for studying disclosure navigation. The FCA's TCFD-aligned rules require in-scope listed companies to state in their annual financial reports whether the recommended disclosures are consistent with TCFD, where those disclosures are located, and, where applicable, why disclosures are omitted and what steps are planned (Financial Conduct Authority, n.d.). UK company regulations additionally require climate-related financial disclosures from in-scope listed companies, large private companies, and limited liability partnerships for financial years beginning on or after 6 April 2022 (Department for Business, Energy & Industrial Strategy, 2022). The FCA's review of early reporting documented widespread use of TCFD indexes and cross-references within annual reports and to separate documents (Financial Conduct Authority, 2022). These conditions make the UK suitable for studying company-declared navigation; they do not imply that every company fully complies with every TCFD recommendation.

We fixed the company scope using the FTSE 100 constituent roster acquired on 15 June 2026 (FTSE Russell, 2026). For each of its 100 constituents, we collected one latest annual document from the 2025 reporting cycle (FY2025 or FY2024/25), yielding 25,463 physical PDF pages. We also collected 17 available external PDFs cited by the indexes (1,408 pages). One spreadsheet and two webpage destinations are recorded separately and excluded from the PDF totals. Reports were selected from the fixed roster rather than by whether an index was found, so reports without a qualifying index remain in the corpus.

4.2 Annotation Procedure and Quality Control

For each annual report, we record where its TCFD index is, what directions it gives, and where those directions lead in the available PDFs. These linked records form the benchmark's reference answers.

We first search the extracted PDF text with a fixed set of TCFD-related terms. Annotators then inspect the matching pages and their neighbors to decide whether they contain a qualifying index. They record all pages containing the index, including continuation pages, counting from the first page of the PDF. If no qualifying index is found through this review, the report receives `NO_INDEX`. This label does not mean that the report lacks climate disclosures.

For reports with a confirmed index, a multimodal extractor drafts records from its pages. We compare these records with the original pages, correct errors, and add missing entries found during review. Each record links a TCFD topic to the document and location named in the index, while preserving the original wording and source page. We record specific recommendations or shared groups

only when the index explicitly identifies them. Explicit page ranges are expanded into individual pages; for example, "pp. 10–12" becomes pages 10, 11, and 12 for the same topic and document.

For each target PDF covered by calibration, AI proposes a mapping between visible printed page numbers and their positions in the file. We then inspect the PDF pages to verify and, where necessary, correct the proposed mapping. We record the target file, a checked page, and the conversion rule. For example, printed pages 4–7 in Diageo's external methodologies report correspond to PDF pages 6–9. A program applies the checked rule to the recorded references to generate target addresses. Each address includes the target file and PDF page; two-page layouts also specify the left or right half.

As a quality check, reports provisionally labeled `NO_INDEX` receive an additional review with Qwen3.5-27B to check for missed indexes. The authors resolve uncertain records and manually check generated target addresses against the target PDFs. Automated checks flag conflicting index labels, missing source pages, duplicate record identifiers, and inconsistent document or page references. Reports with an index must have recorded index pages; `NO_INDEX` reports must have none. We did not conduct an independent double-annotation agreement study.

4.3 Dataset Characteristics

Table 2 summarizes the fixed 100-report corpus and the source-checked annotations used for evaluation, which restore verified document names and a six-page omission in the earlier normalized references. Because reports were selected independently of index presence, the 30 `NO_INDEX` cases support a separate rejection test. The revision has 3,407 resolved topic-bound PDF relations, equivalent to 2,411 relations when collapsed to pillars; all current address scores use the topic-bound form.

Table 2: Dataset hierarchy. Indentation shows the scope or derivation of each count.

Scope and counting unit	Count
Collected documents	
Source annual reports	100
Physical pages	25,463
With a qualifying index	70
Annotated index pages	114
Reports with a multi-page index	18
`NO_INDEX`	30
External PDFs	17
Physical pages	1,408
Within the 70 index-positive reports	
Annotated index records	622
Report–pillar groups	279
Report–recommendation groups	536
Declared relations	
Topic–document–printed-page relations	3,552
Resolved to PDF addresses	3,407
Distinct PDF addresses	1,847
Without a resolved PDF address	145
Document-only	23
Pagination support (subset of 117 collected PDFs)	
PDFs with verified page-number mapping	83
Annual reports	70
External PDFs	13

5 Results and Analysis

We first evaluate whether complete reports are navigated correctly, then examine index discovery, mapping recovery, and target resolution. All results use the same 100-report development cohort, containing 70 index-positive and 30 index-negative reports. Stage 2 and address-level scores are de-

rived from one source-checked annotation revision. Failed or invalid outputs remain in the denominator as empty predictions. Model coverage and supplied inputs are specified for each experiment; implementation details appear in Appendix D.

5.1 Overall Performance

We evaluate whether the framework recovers every annotated navigation relation declared by an annual report's TCFD index. This conditional composition uses saved Gemini discovery and mapping outputs and Qwen3.5-27B calibration outputs, with the target registry, document aliases, and pagination probe positions supplied; mapping uses predicted index spans. Table 3 shows report-level exact match after index discovery, mapping recovery, and final navigation. All annotated navigation relations match for 47/70 reports with an index, and the system correctly returns `NO_INDEX` for all 30 reports without one. Although Stage 2 achieves a relation F_1 above 90% and the index span is correct for 65/70 positive reports, final navigation is entirely correct for only 47/70.

Table 3: Complete reports by stage. The 30/30 entry denotes correct `NO_INDEX` rejection.

Reports	Complete index	Complete mapping	Complete navigation map
With an index (70)	65/70	47/70	47/70
Without an index (30)	30/30	–	–

5.2 Stage 1: Index Discovery

Candidate selection. The 100 annual reports contain more than 25,000 pages, making it costly to send every page to a vision-language model. Table 4 compares candidate selection using BM25 (Robertson & Zaragoza, 2009), ColQwen2 (Faysse et al., 2025), reciprocal rank fusion (RRF; Cormack et al., 2009), and keyword baselines. BM25 and ColQwen2 each place 67/70 index starts among their top three pages.

Table 4: Index starts found among the top-ranked pages (70 reports).

Page ranking	Top 1	Top 3	Top 5
Keyword matches in page order	8.57%	47.14%	74.29%
Keyword heuristic	65.71%	91.43%	94.29%
Text-ranking RRF	74.29%	94.29%	95.71%
BM25	**81.43%**	**95.71%**	95.71%
ColQwen2	80.00%	**95.71%**	**98.57%**

We observe that a highly ranked page can lie next to the index start. We therefore test a one-page window on either side of the top-ranked pages and remove duplicates. **Expanding the BM25 top-three pages in this way covers all 70 index starts with 797 candidate pages across the 100 reports** (Figure 3). At a similar budget, selecting the top eight pages without expansion uses 800 candidates but covers only 68 starts. We use the expanded BM25 candidate set for subsequent index discovery.

Index discovery. We present the selected candidate pages in physical-PDF order and ask each model to identify the first qualifying index page. Figure 4(a) reports each model's best observed index-start accuracy, together with correct rejection of reports without an index. Since an index may continue across several pages, detecting its first page is followed by reading subsequent pages to recover the complete span.

To examine the importance of the index start, Figure 4(b) compares complete-span recovery from predicted and Gold starts. For each model, we select its best observed predicted-start span configuration and keep that configuration fixed when supplying the Gold start. **With Gold starts, Gemini improves from 65/70 to 70/70 exact spans**, whereas Qwen3.5-9B recovers 31/70 even with Gold starts. Thus, a correct entry point can remove all observed span errors for one configuration, while

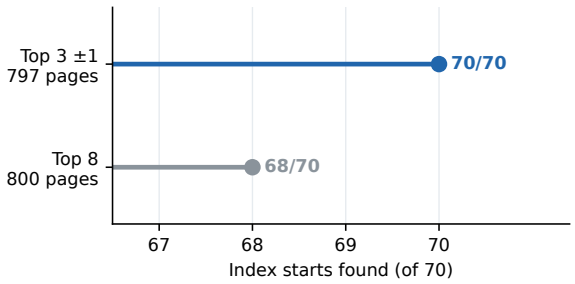

Figure 3: BM25 page selection. Points show index starts found among 70 positive reports; candidate-page counts cover all 100 reports.

continuation decisions remain a limitation for another. Configuration details and full results appear in Appendix C.2.

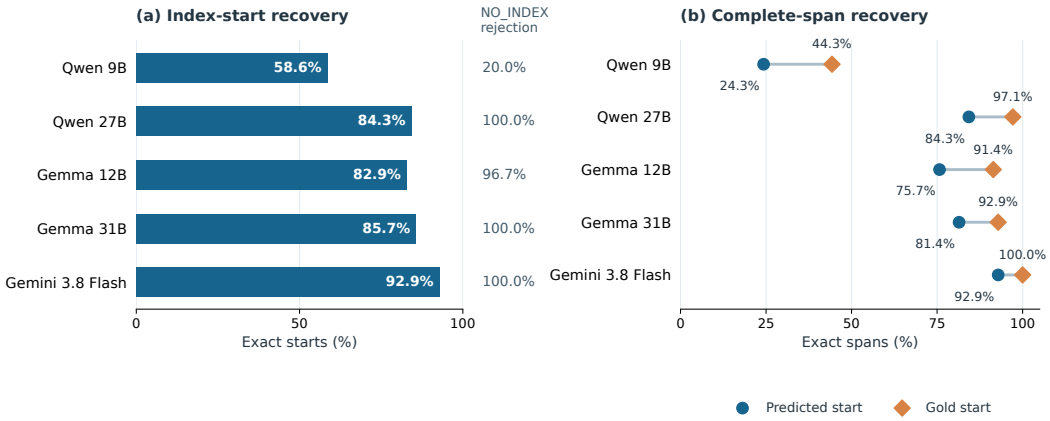

Figure 4: Index discovery and span recovery across five models. (a) Best observed index-start accuracy on 70 positive reports and correct rejection on 30 negative reports. (b) Complete-span exact match from predicted and Gold starts, holding each model's selected span configuration fixed. Configurations were selected on this development cohort; full results appear in Appendix Figure 6.

5.3 Stage 2: Mapping Recovery

We first supply Gold index spans to isolate mapping recovery from index localization. Each scored relation binds a topic unit to a document and a printed page label, or to a document-only destination when no page is stated. The source-checked annotation revision contains 3,552 numeric relations and 23 document-only relations across 70 reports. Table 5 reports complete-binding F_1, mapping report EM, and the diagnostic pillar–page and recommendation–page projections.

High relation-level F_1 does not guarantee high report-level accuracy. Qwen3.5-27B and Gemini achieve complete-binding F_1 of 94.28% and 94.14%, respectively, yet fully recover 44/70 and 49/70 reports. Mapping report EM therefore distinguishes systems that recover comparable numbers of relations but distribute their remaining errors differently across reports. The two page projections omit document identity and cannot establish complete binding. Gemini's external-document F_1 is 87.50%, and its document-only F_1 is 69.57%; these diagnostics expose errors hidden by page-only evaluation (Appendix C.3).

Upstream index errors also reduce report completeness. With the same Gemini extractor, mapping report EM falls from 49/70 with Gold spans to 47/70 with the selected Gemini text-only spans and 43/70 with Qwen3.5-27B text-and-image spans. The earlier Gemini text-and-image spans also yield 47/70, despite recovering three fewer complete indexes. Improved localization can recover

Table 5: Mapping recovery with Gold spans and text-and-image input (70 reports). Page projections omit document identity; recommendation scores cover 56 reports. Scoring counts appear in Appendix C.3.

Model	Mapping report EM	Complete-binding F_1	Pillar–page F_1	Rec.–page F_1
Qwen3.5-9B	32.86%	72.54%	81.58%	79.22%
Qwen3.5-27B	62.86%	**94.28%**	96.49%	**94.65%**
Gemma 4 12B	44.29%	83.34%	91.24%	87.11%
Gemma 4 31B	60.00%	92.86%	96.42%	93.11%
Gemini 3.8 Flash	**70.00%**	94.14%	97.37%	94.46%
GPT-6 Sol	65.71%	92.61%	95.58%	92.79%
GPT-6 Luna	42.86%	92.81%	**98.17%**	94.05%

additional relations without making an otherwise imperfect report exact. All rejected positive reports remain empty predictions; full results and report subsets appear in Appendix C.3.

5.4 Stage 3: Target Resolution

Stage 3 converts the printed page references recovered in Stage 2 into physical PDF addresses by first identifying the rule that maps printed page labels to physical PDF positions. Front matter and page layout can change this mapping, while embedded page labels may be missing or incorrect. Table 6 compares five methods on the same 83 target PDFs, measuring whether each recovers the correct mapping rule at the supplied probe position.

Table 6: Printed-to-PDF page calibration on 83 target PDFs.

Method	How it works	Correct local rules
Physical = printed	Treat the printed number as the physical PDF page.	7.2%
PDF PageLabels	Use embedded PDF page-number metadata.	61.4%
Text-layer parser	Extract a page number from the PDF text layer.	92.8%
Single-page VLM	Read the label and layout on one page image.	98.8%
Sequential VLM	Check neighboring page images; read a third if they disagree.	**100.0%**

Note. All methods use supplied PDFs and probe positions. Both VLM rows use Qwen3.5-27B; the single-page row scores its first reading. Accuracy measures the local rule, which need not hold throughout a PDF.

Single-page VLM improves on the metadata and text-layer methods, but a single image can still mislead layout interpretation. In National Grid's annual report, an unnumbered illustrated panel beside printed page 40 led it to infer two printed pages per PDF page (Figure 7). Sequential VLM checks the mapping inferred from one page against its neighbor and reads a third page only when the first two disagree. In this case, neighboring pages confirmed one printed page per PDF page, correcting the single-page prediction. **Visual reading combined with neighboring-page checks recovers the correct local mapping rule for all 83 PDFs.** Results for five models appear in Appendix C.4 (Table 12); Qwen3.5-9B also achieves this result with Sequential VLM.

6 Conclusion

We introduced TCFDNav, a benchmark and three-stage reference framework for interpreting and following human-readable disclosure indexes. Its layered annotations connect company-declared relations to executable PDF addresses while preserving their source evidence. The experiments show that strong relation-level recovery does not guarantee complete report-level navigation, and that correct index localization alone does not eliminate mapping errors. These findings motivate evaluating both individual relations and the completeness of the resulting navigation map. The reported conditional evaluation does not establish autonomous target-file selection. TCFDNav nevertheless offers an evaluation resource and a reference design for studying similar index-guided navigation tasks beyond TCFD.

AI USE STATEMENT

Generative AI tools assisted with preliminary index and mapping annotations, printed-to-physical page mapping, and additional checks of reports initially labeled `NO_INDEX`. They also supported experimental design, prompt development, code implementation and debugging, scoring checks, literature review, result interpretation, manuscript drafting and language editing, and the creation and revision of schematic figures. The authors checked preliminary annotations and target addresses against the source PDFs, resolved uncertain cases, and made the final labeling decisions. Experimental results were checked using scoring scripts and consistency tests. The authors reviewed the AI-assisted text, references, figures, code, and analyses incorporated into this work and take responsibility for the final manuscript and released artifacts.

REFERENCES

Shafiuddin Rehan Ahmed, Ankit Shah, Quan Hung Tran, Vivek Khetan, Sukryool Kang, Ankit Mehta, Yujia Bao, and Wei Wei. Enhancing retrieval for ESGLLM via ESG-CID: A disclosure content index finetuning dataset for mapping GRI and ESRS. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pp. 1–16, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.climatenlp-1.1. URL https://aclanthology.org/2025.climatenlp-1.1/.

Najah-Imane Bentabet, Rémi Juge, Ismail El Maarouf, Virginie Mouilleron, Dialekti Valsamou-Stanislawski, and Mahmoud El-Haj. The financial document structure extraction shared task (FinTOC 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pp. 13–22, Barcelona, Spain (Online), December 2020. COLING. URL https://aclanthology.org/2020.fnp-1.2/.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 758–759, 2009. doi: 10.1145/1571941.1572114.

Hervé Déjean and Jean-Luc Meunier. Versatile page numbering analysis. In *Document Recognition and Retrieval XV*, volume 6815, pp. 68150K. SPIE, 2008. doi: 10.1117/12.764839. URL https://doi.org/10.1117/12.764839.

Department for Business, Energy & Industrial Strategy. Climate-related financial disclosures for companies and limited liability partnerships (LLPs), February 2022. URL https://www.gov.uk/government/publications/climate-related-financial-disclosures-for-companies-and-limited-liability-partnerships-llps. Published 21 February 2022; accessed 11 September 2026.

Dong Ding, Bin Liu, Jing Yu, and Millicent Chang. TCFD-aligned climate change disclosures and analyst forecasts, 2025. URL https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5177862. Working paper, SSRN.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/99e9e141aafc314f76b0ca3dd66898b3-Abstract-Conference.html.

Financial Conduct Authority. Review of TCFD-aligned disclosures by premium listed commercial companies, July 2022. URL https://www.fca.org.uk/publications/multi-firm-reviews/tcfd-aligned-disclosures-premium-listed-commercial-companies. Accessed 14 August 2026.

Financial Conduct Authority. Climate change and sustainable finance: Reporting requirements, n.d. URL https://www.fca.org.uk/firms/climate-change-sustainable-finance/reporting-requirements. Accessed 22 September 2026.

FTSE Russell. FTSE 100 Index, 2026. URL https://www.lseg.com/en/ftse-russell/indices/ftse100. Constituent roster acquired 15 June 2026; accessed 22 September 2026.

Emmanuel Giguet and Gaël Lejeune. Daniel@FinTOC-2021: Taking advantage of images and vectorial shapes in native PDF document analysis. In *Proceedings of the 3rd Financial Narrative Processing Workshop*, pp. 70–74, Lancaster, United Kingdom, 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.fnp-1.13/.

Emmanuel Giguet and Nadine Lucas. GREYC@FinTOC-2022: Handling document layout and structure in native PDF bundle of documents. In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pp. 100–104. European Language Resources Association, 2022. URL https://aclanthology.org/2022.fnp-1.15/.

Google DeepMind. Gemini 3.8 Flash model card, 2026a. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/. Accessed 24 September 2026.

Google DeepMind. Gemma 4 model card, 2026b. URL https://ai.google.dev/gemma/docs/core/model_card_4. Accessed 24 September 2026.

Tanay Gupta, Tushar Goel, and Ishan Verma. Exploring multimodal language models for sustainability disclosure extraction: A comparative study. In *The Sixth Workshop on Insights from Negative Results in NLP*, pp. 141–149. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.insights-1.13. URL https://aclanthology.org/2025.insights-1.13/.

Samindi Ishara Hewa, Rajni Mala, Jinhua Chen, and John Dumay. Climate related disclosures and investor behaviour: An australian study. *Advances in Accounting*, 68:100809, 2025. doi: 10.1016/j.adiac.2025.100809.

Pengfei Hu, Zhenrong Zhang, Jianshu Zhang, Jun Du, and Jiajia Wu. Multimodal tree decoder for table of contents extraction in document images. In *Proceedings of the 26th International Conference on Pattern Recognition*, pp. 1756–1762. IEEE, 2022. doi: 10.1109/ICPR56361.2022.9956301. URL https://doi.org/10.1109/ICPR56361.2022.9956301.

Guosheng Li, Fenghui Ren, Bin Liu, Chuan Yu, Kaiying Ji, Lin Yue, Jun Shen, and Sasa Qian. What transfers under source shift? definitions, examples, and fine-tuning for climate disclosure classification. arXiv preprint arXiv:2607.17952, 2026. URL https://arxiv.org/abs/2607.17952. Accepted to Findings of EMNLP 2026.

Harsh Mankodiya, Chase Gallik, Theodoros Galanos, and Andriy Mulyar. AEC-Bench: A multimodal benchmark for agentic systems in architecture, engineering, and construction. arXiv preprint arXiv:2603.29199, 2026. URL https://arxiv.org/abs/2603.29199.

National Grid plc. Annual report and accounts 2025/26. https://www.nationalgrid.com/document/579676/download, 2026. Source PDF, physical page 42.

OpenAI. GPT-6 Luna model documentation, 2026a. URL https://developers.openai.com/api/docs/models/gpt-6-luna. Accessed 24 September 2026.

OpenAI. GPT-6 Sol model documentation, 2026b. URL https://developers.openai.com/api/docs/models/gpt-6-sol. Accessed 24 September 2026.

Marco Polignano, Nicola Bellantuono, Francesco Paolo Lagrasta, Sergio Caputo, Pierpaolo Pontrandolfo, and Giovanni Semeraro. An NLP approach for the analysis of global reporting initiative indexes from corporate sustainability reports. In *Proceedings of the First Computing Social Responsibility Workshop within the 13th Language Resources and Evaluation Conference*, pp. 1–8. European Language Resources Association, 2022. URL https://aclanthology.org/2022.csrnlp-1.1/.

Qwen Team. Qwen3.5-27B model card, 2026a. URL https://huggingface.co/Qwen/Qwen3.5-27B. Accessed 24 September 2026.

Qwen Team. Qwen3.5-9B model card, 2026b. URL https://huggingface.co/Qwen/Qwen3.5-9B. Accessed 24 September 2026.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Nicolas Sannier, Morayo Adedjouma, Mehrdad Sabetzadeh, and Lionel C. Briand. An automated framework for detection and resolution of cross references in legal texts. *Requirements Engineering*, 22(2):215–237, 2017. doi: 10.1007/s00766-015-0241-3. URL https://doi.org/10.1007/s00766-015-0241-3.

Tobias Schimanski, Jingwei Ni, Roberto Spacey Martín, Nicola Ranger, and Markus Leippold. ClimRetrieve: A benchmarking dataset for information retrieval from corporate climate disclosures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17509–17524. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.969. URL https://aclanthology.org/2024.emnlp-main.969/.

Task Force on Climate-related Financial Disclosures. Recommendations of the task force on climate-related financial disclosures. Technical report, Financial Stability Board, June 2017. URL https://www.fsb-tcfd.org/wp-content/uploads/2017/06/FINAL-TCFD-Report-062817.pdf.

Xinyu Wang, Lin Gui, and Yulan He. A scalable framework for table of contents extraction from complex ESG annual reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13215–13229. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.816. URL https://aclanthology.org/2023.emnlp-main.816/.

Yizhou Wang, Siyu Shen, Jun Xie, Hidemichi Fujii, Alexander Ryota Keeley, and Shunsuke Managi. How corporate climate change mitigation actions affect the cost of capital. *Corporate Social Responsibility and Environmental Management*, 31(6):5139–5154, 2024. doi: 10.1002/csr.2853.

A Limitations

Data coverage and generalization. TCFDNav covers one UK company roster and one reporting period in the TCFD disclosure setting. This cohort also informed system development and error analysis. Generalization to other countries, reporting periods, and disclosure frameworks therefore remains untested and requires evaluation on independent report collections.

Scope of index-guided navigation. TCFDNav evaluates whether systems can follow the directions declared in a qualifying disclosure index or return `NO_INDEX` when no such index is present. It includes reports without a qualifying index, but does not require systems to retrieve their climate disclosures or find additional relevant content beyond the index's declared directions. Its results therefore concern index-guided document navigation rather than general-purpose document retrieval.

B Framework and Evaluation Details

We give implementation details and source examples for Section 3.

B.1 Evaluation Protocol

Relation micro-F_1 measures agreement with annotated relations; report exact match (EM) requires the complete scored set, without additions or omissions. Index discovery scores exact starts and spans on positive reports and `NO_INDEX` rejection on negative reports. Gold starts isolate continuation errors.

Mapping recovery scores complete topic–document–location bindings, keeping printed-page and document-only directions distinct. Mapping report EM checks the entire report; pillar–page, recommendation–page, and document scores are diagnostic projections. Gold and predicted spans separate extraction from upstream errors. Target resolution scores topic-bound addresses (d, q, r), including document, physical PDF page, and region. *Address-set report EM* checks all scored PDF addresses, but not whether every declared destination has an executable address.

Navigation-map report EM additionally checks topic units, document mentions, and destinations. *Fully page-resolved report EM* also requires a PDF page address for every scored destination; correct

document-only or unresolved directions cannot meet this stricter criterion. Scores use one source-checked revision and exclude free-text locator wording and source-evidence accuracy. Failed or invalid outputs count as empty predictions. Supplied Gold inputs and target identities are stated per condition; such scores are conditional, not autonomous end-to-end results.

B.2 Index Discovery and Span Recovery

We rank extractable-text pages with BM25 and retain the top three and their immediate neighbours after deduplication. A model then checks candidates in physical-page order for an explicit topic-to-location structure; British Land's list-style index in Figure 5(a) shows why a grid is not required. Text, images, or both can be supplied.

After a start is found, the model reads forward until the index entries end. The output is a complete span in one-based physical PDF pages. Failed calls, uncertain boundaries, and truncated searches are recorded separately from confirmed `NO_INDEX` decisions.

B.3 Mapping Recovery

The extractor reads the complete span to recover topic–document–location relations. It retains the pillar, any explicit recommendation or shared-ID group, the company's document hint, the original location expression, and the supporting index excerpt. This keeps references in adjacent document columns distinct; Aviva's external Governance reference in Figure 5(b) is one example.

Normalization expands explicit ranges but preserves gaps: Prudential's Strategy pages $\{119, 120, 122, 123, 125\}$ are not the interval $[119, 125]$. Text-only locations remain textual.

B.4 Target Resolution and Map Construction

Resolution matches each document hint to an available file before locating its printed page. The reviewed-input evaluation supplies the external-target registry. Missing, ambiguous, and non-PDF targets retain their references and resolution states.

For a matched PDF, a document-specific rule maps printed label p to (d, q, r): target document d, one-based physical PDF page q, and region r. One-up segments may use $q = p + \Delta_d$ and $r = \texttt{FULL}$; two-up segments require left/right regions, as Berkeley illustrates in Figure 5(c). A vision-language reader derives a candidate rule from visible labels on supplied probe images. The sequential check is specified in Appendix D.2.

Map construction applies the rule to normalized references at their declared topic granularity. Unresolved directions remain in the navigation map but have no executable PDF address. The resulting address is traceable to the source index.

B.5 TCFD Pillars and Recommended Disclosures

TCFD organizes its recommendations under four pillars, with eleven recommended disclosures in total (Task Force on Climate-related Financial Disclosures, 2017). The letters (a)–(c) restart within each pillar: Governance (a), for example, concerns board oversight, whereas Strategy (a) concerns climate-related risks and opportunities across time horizons. The hierarchy is summarized below.

(a) A list can be an index

British Land | physical PDF 66

- Governance – recommended disclosures (a) and (b). Pages 64 to 65
- Strategy – recommended disclosures (a) to (c). Pages 66 to 71
- Risk Management – recommended disclosures (a) to (c). Pages 71 to 72
- Metrics and Targets – recommended disclosures (a) to (c). Pages 72 to 73

Difficulty
No table grid

Recognise the explicit pillar-to-location links within disclosure prose.

(b) A reference can lead to another document

Aviva annual report | index: physical PDF 76

Source index: external-document column

Section of the Climate-related Financial Disclosure with further details, in compliance with the Listing Rules
• Governance - Our management's climate roles and responsibilities (see page 35)

↓

Target: separate disclosure PDF

OUR MANAGEMENT'S CLIMATE ROLES AND RESPONSIBILITIES 35

Printed 35 -> PDF 37 (FULL); reviewed +2 rule

Stage 2: read the direction
Governance + target document hint + p. 35

Stage 3: execute it
Identify the separate PDF and resolve its page 35.
Do not open page 37 of the annual report.

(c) One PDF page can contain two printed pages

Berkeley | physical PDF 31



Original footer labels

58 59

LEFT **RIGHT**

Retain the region
58 -> (31, LEFT)
59 -> (31, RIGHT)
Same PDF, different sides.

Figure 5: Examples from source reports. (a) British Land's list-style index qualifies without a grid. (b) Aviva's index points to printed page 35 of a separate climate disclosure PDF, located at physical page 37. (c) Berkeley's printed pages 58 and 59 occupy the left and right regions of one physical PDF page.

Table 7: TCFD pillars and recommended disclosures, paraphrased from Task Force on Climate-related Financial Disclosures (2017). Each letter is scoped to its pillar.

Pillar	Recommended disclosures
Governance	(a) Board oversight of climate-related risks and opportunities; (b) management's role in assessing and managing them.
Strategy	(a) Risks and opportunities identified over the short, medium, and long term; (b) effects on business, strategy, and financial planning; (c) resilience under climate scenarios, including a 2°C-or-lower scenario.
Risk Management	(a) Processes for identifying and assessing climate-related risks; (b) processes for managing them; (c) integration into overall risk management.
Metrics and Targets	(a) Metrics used to assess climate-related risks and opportunities; (b) Scope 1 and 2, and where appropriate Scope 3, greenhouse-gas emissions and related risks; (c) targets and performance against them.

An annual-report index may cite an entire pillar, one recommended disclosure, or several disclosures through a shared reference. These are different levels of detail in the company's directions, not additional levels of the TCFD hierarchy. TCFDNav records the level stated in the index and does not infer recommendation-level links from the contents of target pages.

C Supporting Results and Evaluation Conditions

These diagnostics supplement Section 5 under the stated inputs.

C.1 Composed Evaluation: Scoring Objects and Coverage

Stage 2, Stage 3, and composed scores use the same source-checked annotation revision (24 September 2026): 613 topic units, 693 topic–document mentions, 3,552 numeric relations, and 23 document-only relations. Address conversion yields 3,407 topic-bound PDF relations; 145 numeric references remain without a calibrated address. Collapsing topic granularity yields 2,411 pillar-level relations over 1,847 distinct addresses. These are projections of one revision. Eleven reports have document-only or unresolved destinations, limiting fully page-resolved report EM to at most 59/70. Free-text locator wording and source-evidence accuracy are not scored.

Table 8: Matched span-input comparisons with the same Gemini text-and-image mapping extractor and Qwen3.5-27B address calibration. Report percentages use 70 positive reports. Gold spans are diagnostic; bold marks the best predicted-span result.

Index-span input	Navigation-map report EM	Address F_1
Gemini text-only (selected)	**67.14%**	**93.01%**
Gemini text+image	**67.14%**	92.73%
Qwen3.5-27B text+image	61.43%	88.53%
Gold spans	70.00%	94.06%

The target registry, source-scoped aliases, and probe positions are supplied. The selected Gemini text-only spans add exact starts for Entain, Rentokil Initial, and Unilever; their saved Gold-span requests used those same pages, so mapping outputs can be reused without supplying Gold mappings. The 80-document Qwen3.5-27B calibration bank includes Diageo.

C.2 Index Input and Continuation Diagnostics

Unlike the text-based rankings in Table 4, ColQwen2 processes full-report page images before selection.

Figure 4(a) selects each model's configuration by exact start accuracy, breaking ties by correct rejection and then preferring text-and-image input. Panel (b) instead selects by predicted-start span exact match, with start accuracy and correct rejection as tie-breakers, and uses the same configuration for

its Gold-start comparison. The selected span configurations use text for Gemini and Qwen3.5-9B, and text-and-image input for Qwen3.5-27B and both Gemma models. Selection uses the development cohort.

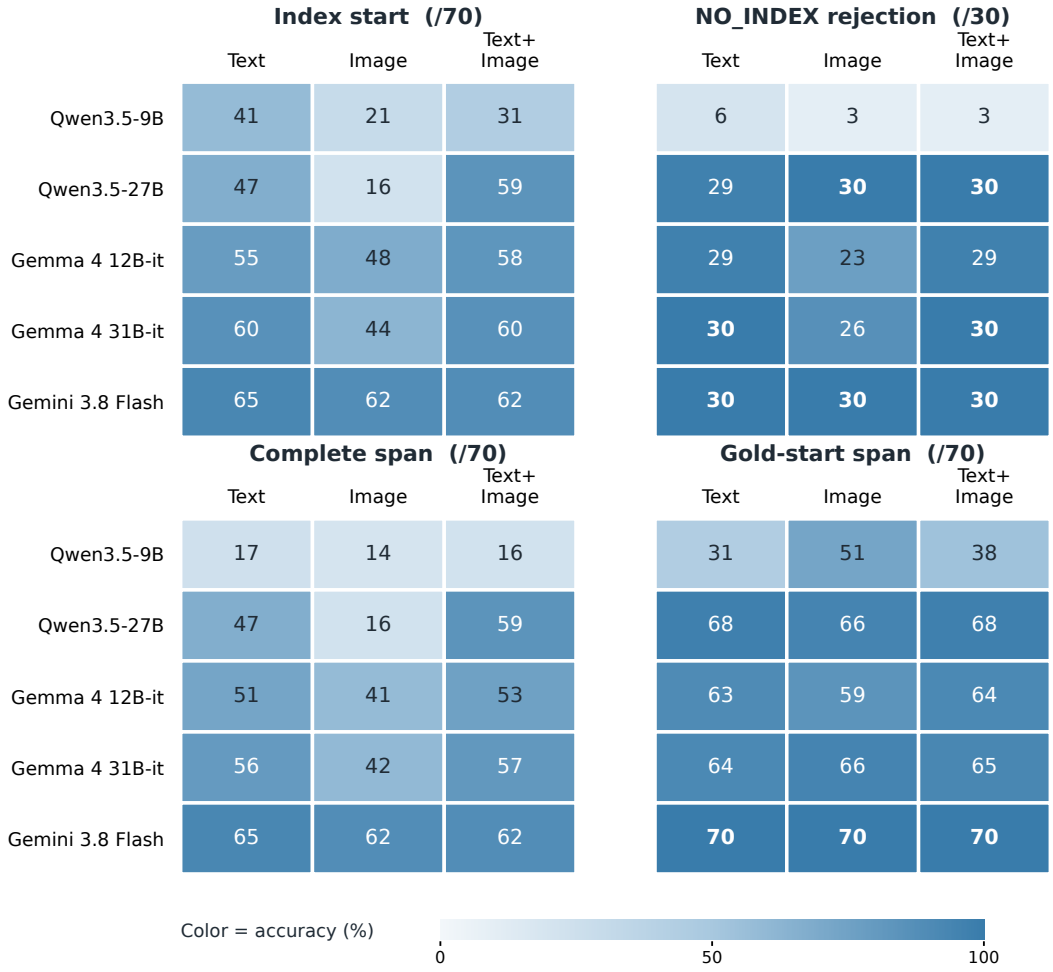

Figure 6: Index discovery across five models with matched input modes. Cells show correct reports; color encodes accuracy. Start and span scores use 70 index-positive reports; `NO_INDEX` rejection uses 30 negative reports. Complete spans use predicted starts. Missing and context-overflow outputs remain in the denominators.

Explicit continuation criteria. In a separate paired prompt experiment, requiring every included page to contain index-mapping entries improves exact-span recovery for all four evaluated open models. Gemma 4 12B rises from 28/70 to 52/70 with saved predicted starts. Start decisions and inference settings are fixed within this comparison; its start predictions differ from the updated main runs.

The utility of visual input depends on the subtask. In the matched modality comparison, text-only input yields higher start accuracy than image-only input for every tested model. Adding images to text raises Qwen3.5-27B from 47/70 to 59/70 exact starts, but does not exceed text-only performance for Gemini or Qwen3.5-9B. With the correct start supplied, however, Qwen3.5-9B recovers 51/70 exact spans from image-only input and 31/70 from text-only input. These are comparisons among implemented system configurations, not a controlled causal estimate of visual input.

C.3 Mapping Configurations and Error Categories

Complete-binding F_1 scores 3,552 numeric and 23 document-only relations. The pillar–page and recommendation–page projections contain 2,460 and 2,933 relations, respectively; the latter covers recommendation/shared-group annotations in 56 reports.

Table 9: Document-reference diagnostics from the same Gold-span predictions as Table 5. External-document F_1 scores 26 named targets across 70 reports; document-only F_1 scores 23 no-page relations. These scores do not test complete topic–document–location bindings.

Model	External document F_1	Document-only F_1
Qwen3.5-9B	48.65%	10.84%
Qwen3.5-27B	71.43%	51.28%
Gemma 4 12B	45.00%	25.00%
Gemma 4 31B	84.44%	30.00%
Gemini 3.8 Flash	**87.50%**	**69.57%**
GPT-6 Sol	83.02%	53.85%
GPT-6 Luna	54.90%	33.33%

Table 10: Complete-binding report exact match on two diagnostic subsets of the Gold-span, text-and-image comparison. The recommendation subset contains 56 reports with explicit recommendation/shared-group annotations; 14 of them also have named external targets. Each report is correct only if all of its scored page-level and document-only relations match. Physical PDF positions are not scored.

Model	Reports with recommendation annotations ($n = 56$)	Of these, reports with external targets ($n = 14$)
Qwen3.5-9B	39.29%	28.57%
Qwen3.5-27B	67.86%	50.00%
Gemma 4 12B	44.64%	28.57%
Gemma 4 31B	62.50%	28.57%
Gemini 3.8 Flash	**73.21%**	**64.29%**
GPT-6 Sol	66.07%	42.86%
GPT-6 Luna	42.86%	28.57%

Table 11: Complete-binding recovery under Gold and predicted index spans. Each extractor keeps the same text-and-image configuration across span inputs. Relation F_1 uses 3,575 typed Gold relations; report exact uses 70 positive reports. Bold identifies the best value within each span-input condition. GPT predicted-span runs are unavailable. Gemini T and T+I denote text-only and text-and-image upstream spans. The selected Gemini T row holds the extractor fixed and is recomposed from exact-span-matched saved requests.

Extractor	Span input	Complete-binding F_1	Mapping report EM
Qwen3.5-9B	Gold	72.54%	32.86%
Qwen3.5-9B	Gemini T+I	72.38%	32.86%
Qwen3.5-9B	Qwen T+I	70.86%	31.43%
Qwen3.5-27B	Gold	**94.28%**	62.86%
Qwen3.5-27B	Gemini T+I	92.13%	60.00%
Qwen3.5-27B	Qwen T+I	88.73%	54.29%
Gemma 4 12B	Gold	83.34%	44.29%
Gemma 4 12B	Gemini T+I	81.65%	41.43%
Gemma 4 12B	Qwen T+I	79.17%	41.43%
Gemma 4 31B	Gold	92.86%	60.00%
Gemma 4 31B	Gemini T+I	90.82%	57.14%
Gemma 4 31B	Qwen T+I	87.45%	52.86%
Gemini 3.8 Flash	Gold	94.14%	**70.00%**
Gemini 3.8 Flash	Gemini T+I	**92.88%**	**67.14%**
Gemini 3.8 Flash	Qwen T+I	**88.82%**	**61.43%**
Gemini 3.8 Flash	Gemini T (selected)	93.15%	67.14%

Table 12: Sequential VLM calibration across the same 83 target PDFs as Table 6. Accuracy requires a correct local page rule at the supplied probe position; it does not evaluate every referenced page. Missing or invalid rules remain in the denominator.

Model	Correct local rules ↑
Qwen3.5-9B	**100.0%**
Qwen3.5-27B	**100.0%**
Gemma 4 12B-it	90.4%
Gemma 4 31B-it	92.8%
Gemini 3.8 Flash	**100.0%**

C.4 Page-Rule Calibration Diagnostics

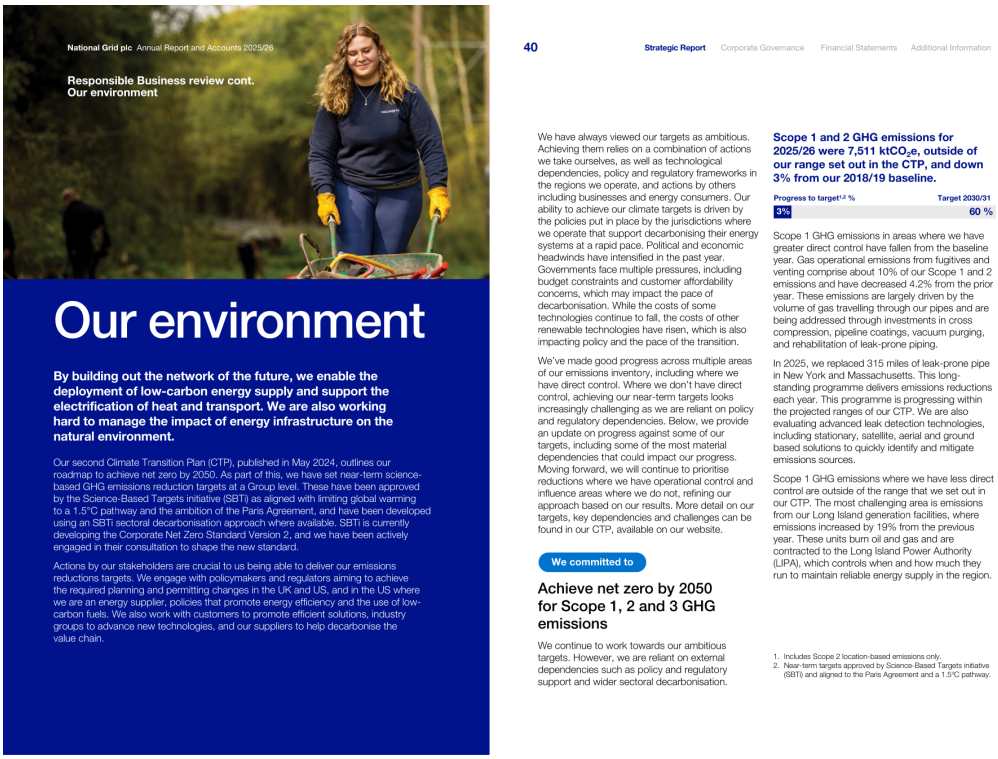

Figure 7: A visually misleading pagination example from National Grid plc's *Annual Report and Accounts 2025/26* (National Grid plc, 2026), physical PDF page 42. An unnumbered illustrated panel appears beside printed page 40, resembling a two-page spread. The first visual reading predicted two-up, while neighboring pages established a one-up numbering rule. The source page is reproduced without added markings.

The single-page row in Table 6 reads the first prediction from the same Qwen3.5-27B sequential run. Table 12 replays saved outputs after page-label normalization; no models were rerun. Local rules are checked at supplied probes, not across entire PDFs.

D Experimental Protocol Details

D.1 Model Coverage and Input Conditions

Table 13 lists the models in the reported comparisons and their sources. The input conditions behind the reported results are described below. The 100-report cohort has 70 index-positive and 30 index-negative source reports. Local Qwen and Gemma checkpoints in these runs used bfloat16 weights, temperature zero, seed zero, and disabled thinking. Cloud providers required separate request settings: Gemini used low thinking where it could not be disabled.

Table 13: Models in the reported comparisons and their sources.

Model	Source
Qwen3.5-9B	(Qwen Team, 2026b)
Qwen3.5-27B	(Qwen Team, 2026a)
Gemma 4 12B-it	(Google DeepMind, 2026b)
Gemma 4 31B-it	(Google DeepMind, 2026b)
Gemini 3.8 Flash	(Google DeepMind, 2026a)
GPT-6 Sol	(OpenAI, 2026b)
GPT-6 Luna	(OpenAI, 2026a)

Stage 1 start and rejection. All models in Table 13 except GPT-6 Sol and Luna are evaluated on the 100 source reports. The input comprises ranked candidate pages checked in physical-page order, PDF text and 1920×2720 JPEG95 page images; the output is one start label.

Stage 1 matched spans. The two Qwen models, two Gemma models, and Gemini 3.8 Flash receive a predicted or Gold start, all confirmed index pages, and the next page; the output is one continuation label.

Stage 2 mapping. The same five models, together with GPT-6 Sol and Luna, are evaluated with one complete Gold span per positive report. The main text-and-image comparison includes all seven; additional image and text conditions are separate implemented configurations. One JSON object contains declared units, document hints, printed references, and source pages. Predicted-span diagnostics hold the downstream extractor fixed.

Stage 3 pagination. The local calibration comparison covers 83 supplied target PDFs for five Qwen, Gemma, and Gemini model configurations. Its 83-document cohort combines the original 79 targets, the Diageo supplement, and the author-reviewed BT, Bunzl, and Admiral PDFs. The previously reported conditional address-execution scores use the earlier 80-document calibration bank and are not re-labeled as 83-document execution results. The visual inputs use 1600-pixel-long-edge JPEG90 probe images to predict visible printed labels and page layout. Gold references, target identities, and probe positions are supplied for conditional address execution.

The reported Stage 1 configurations use a 65,536-token context budget. The logical label budget is 16 tokens; Gemini's transport ceiling is larger to accommodate hidden thinking. Stage 2 image and text inputs are ordered by physical PDF page, use one complete span per call, and allow up to 8,192 output tokens for the joint flat-schema extraction. The text-plus-image condition also changes the prompt, schema, and parser relative to the single-modality runs, so its differences do not isolate the effect of images. Gemini's Vertex requests omit the temperature parameter and add a schema-limit instruction while retaining local validation of the original schema. Stage 3 applies the same image-only page-label task to each saved deployment; rule selection compares two probe pages and requests a third on disagreement. Missing or invalid responses remain in the relevant fixed denominator.

D.2 Sequential Page Calibration

The sequential variant induces page-mapping rules from two consecutive page images and requests a third when the first two do not agree. Agreement between any pair selects the rule. If no pair agrees, the implementation retains the first-page result with low confidence, rather than treating the fallback as verified consensus. A missing valid first-page result supplies no rule. The target PDF and starting probe location are supplied inputs; reviewed page rules constitute a separate condition, not model predictions.