

TATVIQ: Evidence-Gated Analysis of Indian IPO Prospectuses with Large Language Models

Rehaan Kumar

October 2026

Abstract

Indian initial public offering prospectuses combine business descriptions, financial statements, ownership disclosures and material risks in documents that are difficult to review systematically. Large language models can make this information more accessible, but fluent reports are insufficient unless their evidence, numerical interpretation and limitations remain inspectable. This paper presents TATVIQ, a prospectus-analysis system for retail investors, and specifies its evidence-control methodology. The implemented pipeline combines page-anchored extraction, chapter-aware passage retrieval, conservative financial-table alignment, source-quotation checks and a deterministic scorecard-to-verdict rule. An evidence-confidence indicator records coverage and source matching without treating model self-confidence as a probability of correctness. The paper distinguishes extraction failure, absence from selected excerpts and substantiated non-disclosure, and separates quotation provenance from support for an analytical claim. Functional validation establishes selected software properties; it does not establish investment-analysis accuracy. A reproducible evaluation protocol is specified for financial extraction, retrieval recall, citation support, appropriate abstention, verdict stability and processing efficiency. Semantic retrieval and specialist readers are identified as testable extensions rather than assumed improvements. The resulting contribution is an inspectable system design and evaluation framework for evidence-based reading of Indian IPO filings.

Keywords: *Indian IPOs; prospectus analysis; retrieval-augmented generation; financial information extraction; source verification; evidence confidence.*

1. Introduction

An initial public offering presents investors with a point-in-time disclosure about an issuer seeking access to public capital. A draft red herring prospectus (DRHP), red herring prospectus (RHP) or final prospectus can contain the information needed to examine the business, financial condition, promoters, offer structure and risks. Extracting a coherent investment interpretation from that information requires more than shortening the document. It requires distinguishing historical performance from forward-looking statements, identifying the basis of financial figures, understanding ownership incentives and tracing

material conclusions to supporting passages. Indian IPO datasets and prospectus-rating research demonstrate that this is a distinct financial-language task [1, 2].

TATVIQ is designed to support this reading process for retail investors. It produces a structured research report containing fifteen research sections, a financial snapshot, a six-category scorecard, risk and positive signals, and a retail verdict. Its objective is to make the reasoning inspectable, rather than to substitute an automated label for the investor’s examination of the prospectus. The system does not establish issuer quality merely by producing a well-written answer.

The central research problem is the relationship between evidence selection and report reliability. If a model receives only chapter openings, a material disclosure located deeper in the filing can be excluded before generation begins. If the generated report includes a quotation or page number without a source check, the appearance of traceability can exceed the evidence available. If an arithmetic rule changes a model-written verdict, the report’s conclusion and explanation can also diverge.

This paper describes an implemented response to those problems and a protocol for evaluating its remaining weaknesses. The contribution is the integration of document provenance, bounded retrieval, explicit source-check status, context-sensitive missing-information language and deterministic verdict arbitration. Retrieval itself is an established technique [3]; this paper does not claim to introduce it or to demonstrate superior investment returns. The term *evidence-gated* refers to controls on evidence status and confidence. In the current implementation, unsupported quotations are flagged; every unsupported judgment is not automatically removed.

2. Related work

2.1. Indian IPO analysis

Ghosh and colleagues investigate the use of multimodal information in predicting the success of Indian IPOs [1]. Their work provides a relevant domain context, but predicting listing outcomes is a different target from establishing that a research claim is supported by a prospectus. InFiNITE introduces India-specific financial narrative tasks, including Bharat IPO Rating, which uses the recommendation categories Apply, May Apply, Neutral and Avoid [2]. Recommendation labels offer a useful comparison target, but agreement with an expert or broker label does not establish factual correctness. TATVIQ therefore treats evidence and financial extraction as primary evaluation targets and rating agreement as a separate measurement.

2.2. Retrieval and long financial documents

Retrieval-augmented generation combines an information-retrieval stage with a generative model so that relevant source material can inform the answer [3]. BM25 provides an established lexical relevance function that accounts for term occurrence, document length and corpus frequency [4]. Lexical retrieval is useful for specialised financial terms, but it can miss semantically equivalent wording that shares few query terms.

DocFinQA extends financial question answering to full-document contexts and demonstrates that long-context financial reasoning remains challenging [7]. This motivates evaluating the selection of relevant evidence separately from the ability to generate an answer. A larger context window alone does not establish that the right passages were used or interpreted correctly.

2.3. Numerical reasoning and report generation

FinQA evaluates numerical reasoning over financial reports and supplies annotated reasoning programs [5]. TAT-QA addresses questions requiring evidence from both tables and surrounding text [6]. These benchmarks motivate representing a financial value together with its metric, period, unit and accounting scope, rather than treating a nearby number as sufficient evidence.

Financial report generation and collaborative analysis have also been explored through FinDebate and Earnings2Insights [8, 9, 10]. Their existence motivates evaluating specialist readers, but does not establish that a multi-reader architecture would improve TATVIQ’s accuracy, latency or cost. Such an extension requires a controlled comparison against the existing single-generator pipeline.

3. Research questions and design requirements

The design addresses four questions: can material evidence beyond chapter openings be made available to the generator; can a reader distinguish a source match from an unsupported quotation; can the system represent missing information without asserting non-disclosure; and can the final verdict remain consistent with the stated scorecard rule?

These questions imply five requirements. First, source passages must retain an identifiable location. Second, evidence selection must search the eligible document rather than privilege only chapter openings. Third, numerical interpretation must preserve periods, units and signs. Fourth, the report must expose uncertainty where extraction or source matching fails. Fifth, the final label must be determined by an inspectable rule, while acknowledging that the model-generated scores feeding that rule can still be wrong.

“Inspectability” is used narrowly: a reviewer can examine the selected source, the quotation-check outcome and the scorecard arithmetic. It is not equivalent to independent financial assurance or an audit of the issuer’s disclosures.

4. System architecture

The pipeline separates document preparation, evidence selection, generation and post-processing. The generator receives structured extraction information and labelled source passages. Ordinary code checks quotation provenance and computes the displayed verdict and evidence-confidence indicator.

1. Document extraction
2. Passage selection
3. Structured generation
4. Source checks
5. Verdict arbitration
6. Research report

Figure 1. Conceptual flow. Scorecard arithmetic is computed from the structured model output; the final explanation is aligned after source checks and confidence calculation.

4.1. Input and page provenance

For readable PDFs, text fragments are grouped by their vertical position and ordered horizontally. Physical PDF page markers are retained. A physical page index can differ from the page number printed in the document, so the two should not be treated as interchangeable. Text-only excerpts retained with earlier analyses can support reanalysis, but their page labels are not independently authenticated against the original PDF.

Scanned pages require optical character recognition, which is outside the implemented reading path. Documents that cannot be extracted or exceed processing limits are rejected instead of being silently treated as complete. This policy improves the meaning of a successful extraction, but creates a coverage limitation: some legitimate filings remain ineligible.

4.2. Chapter identification and structured fields

Heading detection identifies relevant chapter regions, including business, industry, financial information, management discussion, capital structure, offer structure, proceeds, promoters, related parties, litigation, governance, Basis for Offer Price and Contingent Liabilities. A

chapter title mentioned inside an ordinary cross-reference is not sufficient to establish that the chapter has been extracted. The chapter index is reused within a request.

Structured extraction records selected issuer and offer fields together with provenance and validation information. A field's nominal presence does not imply that every competing value has been resolved. The system distinguishes face value from issue size and records conflicts rather than treating the first monetary amount near a keyword as authoritative. Chapter detection is heuristic; unfamiliar headings or table-heavy layouts can still reduce coverage.

4.3. Financial-table alignment

Financial information is interpreted as a structured observation with a metric, period, unit, value, sign, accounting scope and source location. Positioned text extraction retains large horizontal separations so that some rows can be read as columns. The conservative table reader recognises selected financial metrics only when explicit period columns, numeric cells and currency units align. Parentheses representing negative values are preserved.

When a consistent three-period table is available, its period order and source units govern the report's financial snapshot. Explicit source labels prevent an older storage-field name from relabelling a figure as a different year. Conflicting repetitions are not silently resolved. Ambiguous layouts may instead leave model-reported figures requiring manual verification. This is limited table support, not a general solution for nested headers, footnotes, mixed accounting scopes or arbitrary financial layouts.

4.4. Passage index and retrieval

The normalised document is partitioned into page- and chapter-aligned passages. A passage contains up to approximately 360 words, with a 300-word stride producing overlap where a page or chapter is sufficiently long. These are word counts, not token counts. Passage boundaries preserve source locations, but long rows or paragraphs can still be divided at the word limit.

Each passage receives an identifier, a physical page where available and a chapter label where detected. Fourteen fixed analyst query sets cover business, financial, governance, offer and risk topics. For query q and passage d , lexical relevance is calculated using the BM25 form [4]:

$$R(q, d) = \sum_{t \in q} \text{IDF}(t) \cdot f(t, d)(k_1 + 1) / [f(t, d) + k_1(1 - b + b|d| / \text{avgdl})] \quad (1)$$

Here, $f(t, d)$ is term frequency; $|d|$ and avgdl are the passage and mean passage lengths. The implementation uses $k_1 = 1.2$ and $b = 0.75$, with inverse document frequency $\log(1 + (N - n_t))$

$+ 0.5)/(n_t + 0.5)$). A passage in the query’s corresponding chapter receives a modest relevance boost. Passages outside that chapter remain eligible, because material evidence can appear in several disclosure contexts.

Selection rotates across the query sets within a bounded passage budget. This balances topical coverage against relevance, but does not guarantee that every material passage is selected. The implemented retriever is lexical; embeddings, semantic reranking and a persistent shared document index are not part of this pipeline.

4.5. Structured generation

A large language model produces the report in a structured format using the selected passages and extracted fields. The ordinary path uses one generator; recovery from malformed output or an unavailable model is bounded. No model is fine-tuned for this study. The input instructions require evidence-grounded assessments, explicit limitations and source quotations.

Source text remains untrusted input. Instructions embedded in a prospectus must not be treated as instructions to the application. Prompt instructions reduce this risk, but their presence is not proof of complete resistance to prompt injection. Generation can also introduce unsupported interpretations even when every selected passage is genuine.

5. Source verification and missing-information semantics

5.1. What the source checker establishes

Evidence quotations are compared with their cited passage or physical page. The checker normalises whitespace, Unicode character forms and typographic quotation marks before testing source inclusion. A quotation must be sufficiently long to make a meaningful check. Invented passage identifiers, conflicting passage and page references, or quotations absent from the named source produce an unsupported status. Incomplete references and missing or very short quotations remain unverified or missing.

Quotation provenance and claim support are separate properties. A provenance match answers whether the source contains the quoted words. Claim support asks whether those words justify the surrounding proposition, preserve qualifications and describe the relevant accounting or disclosure context. The current automated check addresses the former. The latter requires separately labelled evaluation and, where necessary, human review.

The report retains source-check warnings. A warning should influence the reader’s interpretation of the analysis, but it does not currently remove every flagged claim or

automatically alter the existing risk-cap rules. This distinction prevents a source-check pass rate from being misrepresented as a complete factual-accuracy measure.

5.2. Distinguishing forms of absent evidence

Table 1. Evidence-state distinctions.

State	Permitted interpretation
Necessary material was not extracted	Extraction is incomplete; a conclusion requiring that material is unsupported.
A fact was not found in selected excerpts	The reviewed evidence does not establish the fact; the full filing may contain it.
The filing explicitly states non-disclosure or a relevant absence	The stated absence can be reported within the scope of that disclosure and its source.

A successful chapter extraction is not an exhaustive review. The system therefore uses “Unable to extract from document” for an extraction limitation and “Not found in the reviewed excerpts” for an evidential limitation. Neither statement alone establishes non-disclosure by the issuer. Claims about non-disclosure must be supported by the document or by a separately defined exhaustive-review procedure.

6. Verdict arbitration and evidence confidence

6.1. Fixed scorecard rule

The final verdict is calculated from six category scores rather than accepted directly from the model’s prose. Let s_i be a category score between 1 and 10 and w_i its weight. The weighted score is:

$$S = \sum_{i=1 \dots 6} w_i s_i, \text{ where } \sum w_i = 1 \quad (2)$$

Table 2. Fixed category weights.

Category	Weight
Business quality	0.22
Financial strength	0.22
Governance / promoter quality	0.18
Industry position	0.14
IPO structure / proceeds	0.12
Valuation	0.12

The uncapped rating is STRONG APPLY at 8.3 or above; APPLY at 7.4; SELECTIVE APPLY at 6.4; NEUTRAL at 5.3; SELECTIVE AVOID at 4.4; AVOID at 3.3; and STRONG AVOID below 3.3. Two or more high-severity flags cap a positive verdict at NEUTRAL. One high-severity flag caps STRONG APPLY or APPLY at SELECTIVE APPLY. Three or more medium-severity flags cap STRONG APPLY at APPLY.

The final explanation and retail bottom line are composed using the computed rating, scorecard and displayed flags. This removes a label-level disagreement between a discarded model verdict and the final verdict. It does not prove that the model’s category scores, flag severities or reasoning are correct. The weights, thresholds and caps are design choices, not empirically established estimates of future returns.

6.2. Evidence-confidence indicator

Model self-reported confidence is not used as the displayed evidence-confidence score. Instead, the system combines selected-topic coverage C , extraction quality E , and verified-quotation coverage V , each scaled to $[0, 1]$. If U is the number of unsupported quotation checks, the indicator is:

$$Q = \text{clip}_{[0,100]}(\text{round}(100(0.40C + 0.30E + 0.30V) - \min(30, 5U))) \quad (3)$$

C measures representation of twelve critical topic areas in the selected passages. E is the extraction-quality output, incorporating chapter and field completeness, low-confidence values and detected conflicts. V is the fraction of checked quotation fields that match their named source. Empty denominators do not produce a perfect score. The indicator has no artificial minimum confidence floor.

This is an evidence-quality heuristic. Its components depend on the detector and checker definitions, and their weights have not been calibrated against independently labelled correctness. A high value is not an accuracy probability, and a low value does not itself establish poor issuer quality. Calibration would require a specified correctness target, separate calibration and test sets, and sufficient independently checked observations.

7. Functional validation and evidential scope

Functional validation targets properties that can be checked without asserting an investment outcome. Regression checks cover retrieval of evidence beyond chapter openings, deterministic repeated passage selection, document isolation, correct and fabricated quotations, wrong page references, invented identifiers, missing chapters, confidence without a minimum floor, explicit financial periods and units, negative values, conflicting table rows and coordinate-based PDF extraction. Post-processing tests use a deliberately contradictory model-written verdict to check that the computed rating, final assessment and retail bottom line agree.

Small checked filing excerpts are used for extraction regression; they are not an independently labelled rating benchmark. Repeated controlled model responses test post-processing consistency. They do not measure whether a live model produces the same scores from the same filing over repeated requests.

Table 3. Completed software checks and their interpretation.

Check group	Observed outcome	What it establishes
Server automated checks	39 passed	Covered extraction, evidence and request-processing properties behave as asserted in the test cases.
Client automated checks	7 passed	Covered account/report-storage behaviours remain compatible.
External integration checks	3 not executed	No external-integration conclusion follows from these checks.
Independent analysis-accuracy benchmark	Not established	No quantitative claim of general financial-analysis accuracy is made.

The reported counts describe the completed software-validation suite for this version. They should not be interpreted as percentages of correctly analysed IPOs. No new live-model accuracy, rating-agreement, calibration or return-prediction result is claimed. This separation is necessary because a correct implementation of source matching can still accompany an incorrect financial interpretation.

8. Reproducible evaluation protocol

8.1. Corpus and annotation

An independent evaluation should combine eligible public IPO resources [1, 2] with checked official filings. Dataset versions, collection dates, document stages, exclusion criteria and licences must be recorded. Sample coverage should be described by sector, offer type, document length and table complexity. Scanned or otherwise unsupported documents must be counted as exclusions, not silently removed from the reported denominator.

Annotations should record financial observations as metric–period–unit–scope–value tuples, including signs and source locations. Evidence questions should identify relevant source spans and distinguish extraction failure, absent selected evidence and substantiated absence. A second annotator should independently check a subset, with agreement and adjudication procedures reported. Expert recommendation labels remain a distinct target and must not be treated as ground truth for every factual claim.

Splits should be grouped by issuer and document family to prevent closely related DRHPs, RHPs and later revisions from appearing on both sides of a comparison. A chronological hold-out can improve the study design, but recent publication alone does not prove absence from a model’s training data. Possible contamination and uncertainty about training cut-offs must remain explicit.

8.2. Measures

- **Financial extraction accuracy.** Score complete observation tuples, not naked numbers. Report exact matches and any pre-specified tolerance after explicit unit conversion. Period, scope and sign errors remain errors even when the numerical magnitude appears close.
- **Evidence recall.** Measure whether an annotated supporting passage occurs among selected evidence and within the reported top-ranked results. Report coverage by question type and chapter.
- **Quotation provenance.** Measure agreement between automatic source-check status and checked source quotations. Report false matches and false rejections separately.

- **Claim support.** Independently assess whether the cited evidence entails the claim, preserves qualifications and refers to the correct accounting context. Do not merge this result with quotation provenance.
- **Abstention correctness.** Evaluate the evidence-state distinction in Table 1, using its own confusion matrix. A cautious answer is not automatically a correct abstention.
- **Rating agreement.** If comparison with a four-class recommendation dataset is required, pre-specify the mapping from TATVIQ’s seven categories and report macro-F1 and the full confusion matrix. Avoid selecting the mapping after viewing outcomes.
- **Verdict stability.** Repeat live analysis under the same documented model and generation settings. Report category distributions, modal agreement and category-score dispersion. Stable repeated errors must not be described as accurate.
- **Processing efficiency.** Record extraction, retrieval and generation time, model-call count, and provider-reported input and output tokens. Monetary cost requires stated pricing and a pricing date; absent usage data must remain missing.

8.3. Comparisons and uncertainty

The primary baseline is chapter-opening selection with the same structured report format and verdict rule. Compare it with lexical passage retrieval under matched model settings and context allowance. Separate comparisons can add semantic retrieval, improved table handling or specialist readers. Verification should also be evaluated independently: turning a warning system on can improve visibility of errors without improving the underlying generated claims.

Use paired per-document comparisons and report uncertainty intervals. Aggregate metrics should be accompanied by errors stratified by financial-table complexity, chapter-detection success and input length. Generation failures, rejected documents and missing results must remain visible in the evaluation accounting. Changes in model version, prompt, retrieval configuration or category mapping define a new experimental condition.

9. Discussion and limitations

The pipeline supports traceability, but its components introduce different failure modes. Heading detection can miss a chapter or assign a passage to an incorrect region. Lexical ranking can favour repeated terminology over a decisive disclosure expressed differently. Selection can omit an important exception. Table extraction can confuse entity scope or fail on a multi-level header. A matched quotation can be genuine while an interpretation omits a qualification or draws an unwarranted conclusion.

Several IPO characteristics require contextual judgment. Promoter ownership changes, related-party transactions, an offer for sale and general corporate use of proceeds are not uniformly negative in every issuer or sector. Their significance depends on materiality, governance, financial condition and the source disclosure. A fixed scorecard promotes consistency of presentation, but cannot replace this contextual reasoning.

The existing severity caps are applied to model-reported flags, including those with source-check warnings. Consequently, an unsupported flag can affect the verdict. This is an explicit limitation of the current policy, and should be examined through labelled evaluations before a different arbitration rule is adopted.

Evidence confidence is likewise sensitive to what its component checks recognise. Chapter coverage is a proxy for available evidence; it is not evidence completeness. Quotation matching is a proxy for source provenance; it is not claim truth. The aggregate indicator can help reveal weaknesses only when these meanings remain clear to the reader.

The report is a point-in-time interpretation of supplied material. Amendments, a later price band, revised offer terms or new disclosures can change the relevant evidence. This research does not validate a forecast of listing gains or long-run returns.

10. Extensions for comparative evaluation

Semantic retrieval could improve recall when financial or governance concepts are expressed using different terminology. A richer table representation could distinguish entity scope, annual and interim periods, and linked footnotes. Both extensions require direct comparison with the current method, including any new extraction errors.

Specialist readers could separately examine business, financial quality, governance, offer structure and valuation before a synthesis stage. Prior collaborative financial-analysis work motivates testing this architecture [8, 10]. It does not justify assuming that more readers will improve quality. Evaluation must include disagreement propagation, source reuse, failed calls, latency and token usage, with the verdict rule held fixed where possible.

Persistent document indexes and calibrated confidence are further extensions. A persistent index must preserve the scope of the source document and its authorised users. Confidence calibration requires independently checked outcomes and an explicit prediction target. These are evaluation directions, not capabilities established by the present software tests.

11. Conclusion

TATVIQ operationalises a structured reading of Indian IPO prospectuses through page-aware extraction, bounded passage retrieval, financial alignment, source-quotation checks and

deterministic verdict arbitration. Its central design distinction is that a readable report, a matched quotation, a stable verdict and a correct financial interpretation are separate properties. Each requires its own evidence.

The implemented controls improve the inspectability of selected evidence and the consistency of the report's final label. Functional validation supports those bounded software claims. General analysis accuracy, confidence calibration and investment outcomes remain outside the established results. The evaluation protocol specifies how those questions can be examined without substituting fluent prose or broker-label agreement for checked financial evidence.

Use and availability. This paper describes an educational research system. It does not constitute financial advice. The methodology research document and its accompanying overview describe the same TATVIQ project and are intended to be read alongside the relevant official prospectus.

References

1. Sohom Ghosh, Arnab Maji, N Harsha Vardhan, and Sudip Kumar Naskar. 2024. *Experimenting with Multi-modal Information to Predict Success of Indian IPOs*. arXiv:2412.16174.
2. Sohom Ghosh, Arnab Maji, and Sudip Kumar Naskar. 2025. *InFiNITE (∞): Indian Financial Narrative Inference Tasks & Evaluations*. Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems, pp. 149–166. Association for Computational Linguistics. DOI: 10.18653/v1/2025.eval4nlp-1.13.
3. Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. Advances in Neural Information Processing Systems, 33, pp. 9459–9474. arXiv:2005.11401.
4. Stephen Robertson and Hugo Zaragoza. 2009. *The Probabilistic Relevance Framework: BM25 and Beyond*. Foundations and Trends in Information Retrieval, 3(4), pp. 333–389. DOI: 10.1561/15000000019.
5. Zhiyu Chen, Wenhui Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. *FinQA: A Dataset of Numerical Reasoning over Financial Data*. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3697–3711. Association for Computational Linguistics. DOI: 10.18653/v1/2021.emnlp-main.300.
6. Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. *TAT-QA: A Question Answering Benchmark on a Hybrid of Tabular and Textual Content in Finance*. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 3277–3287. Association for Computational Linguistics. DOI: 10.18653/v1/2021.acl-long.254.
7. Varshini Reddy, Rik Koncel-Kedziorski, Viet Dac Lai, Michael Krumdick, Charles Lovering, and Chris Tanner. 2024. *DocFinQA: A Long-Context Financial Reasoning Dataset*. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 445–458. Association for Computational Linguistics. DOI: 10.18653/v1/2024.acl-short.42.

8. Tianshi Cai, Guanxu Li, Nijia Han, Ce Huang, Zimu Wang, Changyu Zeng, Yuqi Wang, Jingshi Zhou, Haiyang Zhang, Qi Chen, Yushan Pan, Shuihua Wang, and Wei Wang. 2025. *FinDebate: Multi-Agent Collaborative Intelligence for Financial Analysis*. Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing, pp. 268–282. Association for Computational Linguistics. DOI: 10.18653/v1/2025.finnlp-2.20.
9. Takehiro Takayanagi, Tomas Goldsack, Kiyoshi Izumi, Chenghua Lin, Hiroya Takamura, and Chung-Chi Chen. 2025. *Earnings2Insights: Analyst Report Generation for Investment Guidance*. Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing, pp. 246–251. Association for Computational Linguistics. DOI: 10.18653/v1/2025.finnlp-2.17.
10. Mingrui Tan, Yang Liu, Kun Gao, Fei Gao, and Yuting Song. 2025. *Multi-Agent Collaboration for Investment Guidance: Earnings2Insights Report Generation*. Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing, pp. 328–334. Association for Computational Linguistics. DOI: 10.18653/v1/2025.finnlp-2.26.