fosu ltw @ FinMMEval 2026 Task 1: A Traceable Autoresearch Workflow for Multilingual Financial Multiple-Choice Answer Review

fosu ltw @ FinMMEval 2026 Task 1

Tingwei Liang[1], Kaichuan Lin[1], Yueran Qi[1] and Zhongyuan Han*[1]

[1]Foshan University, Foshan, China

Abstract

This paper presents Traceable Autoresearch Review, a traceable answer-review workflow for multilingual financial multiple-choice evaluation in CLEF 2026 FinMMEval Task 1. The workflow is designed for a shared-task setting in which final-test labels are hidden during submission, but submitted answers still need to be complete, consistent, and auditable. GPT-5.5 is used to generate final-test predictions, and Doubao Expert Mode is used for expert-style review of financial formulas, accounting standards, policy timelines, and confidence tiers. The reviewed outputs are organized into submission JSON files, reasoning sheets, and language-specific reports, and each revision is kept only when it strengthens the supporting evidence and preserves cross-file consistency. On the official final test set, the current method obtains 93.50% for English (187/200, rank 2), 92.00% for Chinese (184/200, rank 2), 94.00% for Arabic (188/200, rank 3), and 92.00% for Hindi (184/200, rank 1). For the 800 final-test questions, all four language submissions pass readiness and consistency checks. The integrated review reports mark 742 answers as high-confidence, 32 as medium-confidence, and 26 as requiring attention. These confidence tiers are used for audit and risk disclosure, while the official test-set scores provide the final competition performance reference.

Keywords

Multilingual Financial Question Answering, Answer Verification, Autoresearch, CLEF, FinMMEval, Traceability

1. Introduction

Multilingual financial multiple-choice question answering is a practical evaluation task for financial NLP. Its goal is to select the correct option for questions involving accounting, finance, economics, taxation, ESG, government accounting, and policy knowledge across different languages. Compared with general multiple-choice question answering, this task requires not only semantic understanding, but also exact formula calculation, accounting-standard interpretation, policy-timeline matching, and robust handling of multilingual terminology.

CLEF 2026 FinMMEval Task 1further increases this difficulty [1, 2]. The final-test set contains 800 questions in English, Chinese, Arabic, and Hindi. During the submission phase, the organizers had released only the training set and the public final-test questions, while the labels and scores remained hidden. In addition, some public final-test questions contain imperfect surfaces, such as missing tables, truncated formulas, absent statement blocks, and damaged options. Under this setting, a conventional one-shot prediction system cannot fully explain whether a submitted answer is reliable, why a risky item is kept, or which questions require later source recovery.

To address this problem, we propose Traceable Autoresearch Review, a traceable autoresearch-style workflow for multilingual financial answer review. Unlike methods that focus only on producing final options, our workflow separates answer generation, expert-style review, structural validation, cross-

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

qq1248054712@126.com (T. Liang); 1441584221bruan@gmail.com (K. Lin); 1206585163@qq.com (Y. Qi); hanzhongyuan@gmail.com (Z. Han*)

0009-0005-8200-8536 (T. Liang); 0009-0001-7843-2888 (K. Lin); 0009-0009-3990-2726 (Y. Qi); 0000-0001-8960-9872 (Z. Han*)

file consistency checking, and risk-tier assignment. GPT-5.5 is used to generate initial final-test predictions, Doubao Expert Mode is used to review financial formulas, accounting standards, policy timelines, and confidence tiers, and the resulting submission files, reasoning sheets, and review reports are checked for completeness and consistency.

The main difference between our method and previous answer-generation pipelines is that we treat final-test preparation as an auditable evidence-triage process rather than a black-box prediction process. Each iteration selects a narrow risk set, checks the source question, current answer, reasoning row, and review report, and keeps a change only when the evidence becomes stronger and all effective artifacts remain consistent. This design is especially suitable for shared-task scenarios where official labels are unavailable but submission validity, traceability, and risk disclosure are still required.

This working notes paper makes three practical contributions. First, it reports a four-language review pipeline for final-test submission JSON, reasoning CSV, and review reports. Second, it integrates language-specific expert review findings into a unified risk taxonomy. Third, it reports the official four-language test-set results together with the underlying submission-readiness and risk-disclosure process. Figure 1 shows the complete traceable autoresearch workflow used to connect answer generation, expert review, validation, risk triage, and evidence-based revision.

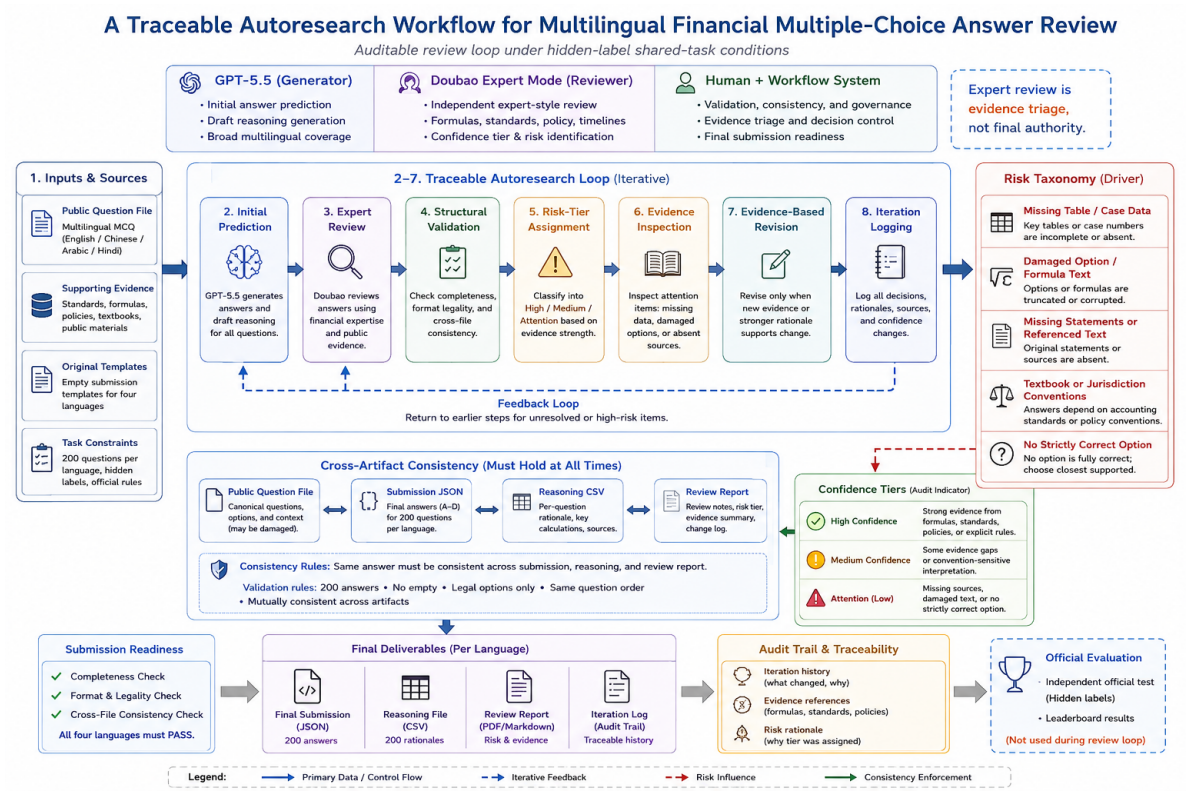

Figure 1: Overview of the traceable autoresearch workflow for FinMMEval Task 1. The workflow connects answer generation, expert review, validation, risk triage, and evidence-based revision.

2. Related Work

Financial question answering has traditionally been handled as a domain-specific reading, retrieval, or reasoning problem. Earlier systems usually combine question understanding, document or knowledge retrieval, and option selection. These methods are useful when the required evidence is available in the input or external resources, but they often provide limited support for final-test auditing when labels are unavailable and public question files contain damaged tables, missing statements, or incomplete options.

Recent LLM-based methods improve coverage by using large language models to generate answers and explanations. Chain-of-thought prompting and self-consistency can elicit intermediate reasoning paths or compare multiple generated outputs [6, 7], while retrieval-augmented generation improves factual grounding for knowledge-intensive tasks [8]. However, these methods still mainly optimize answer production. They do not, by themselves, guarantee that submission files are structurally valid, that reasoning sheets and review reports agree with the final JSON answers, or that uncertain questions are separated from high-confidence ones.

In contrast, our work focuses on the post-generation review layer for a multilingual financial shared task. We do not claim a new base model or a new training algorithm. Instead, we design a traceable workflow that combines LLM-generated predictions, independent expert-style review, fixed validators, and language-specific risk logging. The method therefore differs from previous work in its output target: it produces not only final options, but also an auditable map of which answers are stable, which answers depend on recoverable evidence, and which questions should remain marked as attention items.

3. Task Materials

The project contains four language tracks with 200 questions each. The English and Chinese tracks are associated with RealFin [3], the Hindi track is associated with BhashaBench V1 [4], and the Arabic track is associated with SAHM [5]. Every language is represented by a public JSONL question file, an effective submission JSON, a reasoning CSV, and one or more review reports. Table 1 summarizes the effective artifacts used in this paper.

Table 1
Effective task artifacts after project reorganization. Chinese uses the filled JSON as the authoritative submission source.

Language	Qs	Effective submission	Integrated report evidence
English	200	English template	Key fixes on 044, 053, 163; attention set 044, 053, 118, 140, 163, 179, 190.
Chinese	200	Chinese filled	Filled submission is authoritative; low-confidence set 053, 119, 163, 179; option/formula damage in 026, 070.
Arabic	200	Arabic template	Saudi government accounting and Arabic accounting terminology; attention set 056, 060, 065, 066, 067, 068, 070, 086, 089, 108, 142, 147.
Hindi	200	Hindi template	Economics, Indian policy, GST, exchange-rate conventions; attention set 119, 157, 170; medium-risk set 012, 017, 082, 097, 192.

4. Method

Our method is designed as a review-and-validation framework rather than a model-training framework. A standard answer-generation pipeline usually produces one option per question and then reports the final file. By contrast, Traceable Autoresearch Reviewtreats every answer as a record that must remain consistent across four views: the public question, the submitted option, the reasoning sheet, and the review report. This makes the method closer to an auditable production workflow than to a single prediction model.

4.1. Traceable Autoresearch Loop

The workflow consists of seven steps. First, GPT-5.5 produces initial predictions for the final-test questions using the released training set, public test questions, and domain reasoning prompts. Second, Doubao Expert Mode performs an independent expert-style review, focusing on financial formulas, accounting standards, policy timelines, option validity, and confidence-tier assignment. Third, the

submission is checked against a fixed structural baseline, including question coverage, answer order, option legality, and empty-answer detection. Fourth, the review loop selects a target set, prioritizing attention questions, medium-confidence items, damaged options, missing tables, and cross-file inconsistencies. Fifth, it reads the source question, options, current prediction, reasoning row, and report entry. Sixth, it keeps a change only if evidence is stronger and the submission remains complete and consistent. Finally, the iteration outcome is recorded for later audit.

Compared with majority voting, self-consistency, or direct LLM prompting, the loop does not simply choose the most frequent or most fluent answer. It explicitly records why a question is high-confidence, medium-confidence, or still risky. Therefore, the method can preserve a submitted option while still warning that the original source table, statement text, or complete option set is needed before the answer can be treated as ground truth.

4.2. Evidence Sources

The final decision for each risk item is based on five evidence layers: released training data patterns, public final-test question text, current GPT-5.5 prediction, Doubao Expert Mode review, and project-local reasoning/report artifacts. When these layers conflict, runtime validation and current effective artifacts are prioritized over earlier templates or draft notes. Expert-model review is treated as a secondary evidence source: it can strengthen a rationale, identify risk priority, or suggest confidence-tier movement, but the final answer remains tied to the current effective submission.

4.3. Validation Protocol

The validation protocol has two levels. The first level checks whether each language submission is complete and well-formed: every track should contain 200 answers, preserve the question order, avoid empty predictions, and use only legal option labels. The second level checks whether the submitted option, the reasoning sheet, and the language-specific review report describe the same answer. These checks do not measure correctness against hidden labels directly. Their purpose is to ensure that the submitted files were internally consistent before and during official evaluation.

4.4. Reproducibility and Audit Artifacts

The project is reproducible at the artifact and audit-trail level. Table 2 lists the files that define the workflow state: public question files, effective submissions, reasoning ledgers, review reports, validation scripts, and the experiment ledger. A third party can rerun the structural checks, inspect the reasoning rows and report entries behind a submitted option, and verify whether a documented revision preserves cross-file consistency. This level of reproducibility is different from exact token-level replay. Because the initial answer generation and expert review used closed interactive LLM services, sampler-level parameters such as internal random seeds are not available, and exact model outputs may change over time. We therefore report the auditable artifacts, validation commands, and documented review decisions rather than claiming deterministic regeneration of every model token.

4.5. Documented Review Comparisons

The retained project artifacts do not contain a complete frozen dump of all raw GPT-5.5 outputs before review. Therefore, we do not report a full 800-question initial-versus-final ablation. Instead, Table 3 reports the documented cases for which the project records either an answer change or an explicit keep decision after expert review. These samples show the intended traceability unit: each reviewed item links an initial or current option, a final submitted option, the decisive evidence, and the resulting risk treatment. This partial comparison is sufficient to audit representative revisions, but it should not be interpreted as a full controlled experiment over every initial prediction.

Table 2
Reproducibility artifacts used to audit the workflow.

Artifact type	Project path	Audit role
Public questions	data/public/*.jsonl	Source question text, IDs, and legal option labels.
Effective submissions	submissions/*.json	Final answer labels submitted for each language; Chinese uses the filled JSON as the effective source.
Reasoning ledgers	reasoning/*.csv	Per-question final prediction, confidence tier, and compact rationale.
Review reports	reports/*.md	Human-readable evidence, risk tiers, and language-specific review findings.
Validation scripts	review_scripts/*.py	Rechecks count, ID order, empty answers, illegal labels, and cross-file consistency.
Experiment ledger	logs/results.tsv	Records validation status, kept iterations, and documented target sets.
Prompt summary	docs/user_prompt_summary.md	Summarizes user instructions, expert-review findings, and documented correction evidence.

Table 3
Documented examples of review effects on final submissions.

ID	Before	Final	Decisive evidence	Treatment
044	B	C	Current assets were reconstructed as cash 8,000, net receivables 12,000, inventory 10,000, and prepaid rent 2,800, giving 32,800.	Answer changed; evidence strengthened.
053	A	B	Stock A is a large order better suited to a crossing system, while Stocks B and C fit participation-style execution.	Answer changed; strategy interpretation corrected.
118	B	B	VWAP cost calculation gives $(25.18 - 25.1569) \times 10000 = 230.95$.	Kept; calculation verified.
140	C	C	Expected cash flow is $0.2 \times 0 + 0.5 \times 100000 + 0.3 \times 300000 = 140000$.	Kept; calculation verified.
163	C	B	Recovered statement-pattern evidence supports B, but the public file lacks the original statements.	Changed but kept as high-risk.
179	C	C	The answer matches recovered pattern evidence, but the original statement text is absent.	Kept as high-risk.
190	B	B	IFRS inventory valuation uses the lower of total cost 90,000 and total NRV 85,000.	Kept; accounting rule verified.

5. Integrated Report Findings

5.1. English Track

The English report records a set of concrete corrections supported by recovered original context. Question 044 changes from B to C after current assets are recomputed as cash 8,000, net accounts receivable 12,000, inventory 10,000, and prepaid rent 2,800, totaling 32,800. Question 053 changes from A to B because Stock A is a large order better suited to a crossing system, whereas Stocks B and C fit a logical participation strategy. Question 163 changes from C to B according to recovered statement-pattern evidence.

The English attention set contains 044, 053, 118, 140, 163, 179, and 190. Among them, 163 and 179 are the highest-risk items because the public file lacks the original statement texts. Several medium-confidence items are kept because their damaged options still preserve enough structure: 070 relies on positive homogeneity of coherent risk measures, 155 on the logarithmic CDF transformation of $V = 100000e^R$, and 170 on cumulative default-rate conversion despite percentage formatting errors. Table 4 lists the English attention items and the concrete review issue for each item.

Table 4
English attention items and concrete review issues.

ID	Concrete issue
044	Current assets depend on recovered table values; the answer is supported after reconstructing cash, receivables, inventory, and prepaid rent.
053	Order-execution strategy depends on identifying a large order versus participation-style execution.
118	Missing or incomplete case data weakens the calculation evidence.
140	Missing table/case information prevents a fully closed verification.
163	Original statement text is absent from the public file; the answer is kept as a valid submission but not treated as ground truth.
179	Original statement text is absent, making this one of the highest-priority source-recovery items.
190	Missing source data limits verification beyond the available residual evidence.

5.2. Chinese Track

The Chinese report uses the public Chinese question file as its source and explicitly does not use English as a meaning-alignment reference. Its filled submission is the effective file; the template is kept only as an original empty template. Expert review confirms the overall classification and key calculations, upgrades 064 and 095 to high confidence, downgrades 179 to low confidence, and strengthens the rationale for goodwill recognition in 004 and coherent risk measurement in 070.

The Chinese low-confidence set is 053, 119, 163, and 179. The main failure mode is missing context: 053 lacks order-management details, 119 lacks account-balance tables, and 163/179 lack statement texts. Medium-confidence items include option truncation in 026, damaged formulas in 070, missing case context in 112/116/142/148/177/194, and textbook-dependent transfer pricing in 124. Table 5 shows the Chinese attention and medium-risk items together with their review issues.

Table 5
Chinese attention and medium-risk items with concrete review issues.

ID	Concrete issue
026	Option text is truncated, so the intended option must be inferred from the preserved financial meaning.
053	Order-management details are missing, weakening the basis for selecting the execution strategy.
070	Formula text is damaged; the answer relies on the preserved coherent-risk-measure structure.
119	Account-balance table information is missing, making the calculation weaker than ordinary formula-based items.
124	Transfer-pricing answer depends on textbook or teaching convention.
163	Statement text is absent from the public file.
179	Statement text is absent and was downgraded during expert review.

5.3. Arabic Track

The Arabic report is dominated by accounting, cost accounting, accounting information systems, and Saudi government accounting. Expert review finds no core answer error and moves 075, 125, and 126 to the high-confidence tier. The remaining medium-confidence items are 048, 053, 064, 071, and 120; they depend on traditional asset-exchange treatment, Saudi budget classification, al-uhad classification, synchronized production-stage terminology, and conceptual-framework wording.

The Arabic attention set is 056, 060, 065, 066, 067, 068, 070, 086, 089, 108, 142, and 147. The highest-risk group is 086, 089, 147, and 142: 086/089 lack the referenced source text; 147 depends on whether employees are classified as external users in a specific textbook; 142 contrasts full-disclosure teaching wording with professional materiality. The second group involves Saudi government accounting details such as cash basis, payment evidence, documentary credits, inter-ministry services, unclaimed salaries, municipal bank accounts, and document-folder records. Table 6 summarizes the Arabic attention items and their concrete review issues.

Table 6
Arabic attention items and concrete review issues.

ID	Concrete issue
056–070	Several items depend on Saudi government accounting details, including cash basis, payment evidence, documentary credits, inter-ministry services, unclaimed salaries, and municipal bank accounts.
086	Referenced source text is absent, so expert review cannot close the evidence gap.
089	Referenced source text is absent and should be revisited if the original passage is recovered.
108	Government-accounting terminology requires convention-specific interpretation.
142	Full-disclosure teaching wording conflicts with professional materiality interpretation.
147	The answer depends on whether employees are treated as external users in the textbook convention.

5.4. Hindi Track

The Hindi report covers macroeconomics, microeconomics, Indian economic policy, GST, trade policy, and financial mathematics. Expert review verifies the main formula-based questions and upgrades 057 to high confidence. Medium-confidence items are 012, 017, 082, 097, and 192. Their risks come from business-cycle wording, direct/indirect real-exchange-rate conventions, a missing matching-table column, the timeline distinction between SEBI establishment and legal empowerment, and GST boundary wording.

The Hindi attention set is compact but important: 119, 157, and 170. Question 119 has almost algebraically equivalent arc-elasticity options, so the submitted answer marks the option most likely intended as a non-standard expression. Question 157 produces an open-economy multiplier result of 0.85, which is absent from the options; the submitted answer chooses the closest residual choice under a likely flawed option set. Question 170 has no strictly correct second blank for a negative producer surplus or loss, so the answer keeps the only option beginning with producer surplus. Table 7 presents the Hindi attention and medium-risk items and the corresponding review issues.

Table 7
Hindi attention and medium-risk items with concrete review issues.

ID	Concrete issue
012	Business-cycle wording is convention-sensitive.
017	Real-exchange-rate convention may differ between direct and indirect quotation.
082	Matching-table information is incomplete.
097	The answer depends on the distinction between SEBI establishment and legal empowerment.
119	Arc-elasticity options are nearly algebraically equivalent, so the intended option is not fully stable.
157	The calculated open-economy multiplier is not present in the options; the closest supported option is submitted.
170	No option gives a strictly correct second blank for negative producer surplus or loss.
192	GST boundary wording is policy-convention sensitive.

6. Results

6.1. Official Test-set Performance

The official leaderboard results for CLEF 2026 FinMMEval Task 1are now available. Table 8 reports the final test-set scores, ranks, and completion status for the four language tracks.

6.2. Submission Readiness

All four effective submissions pass the readiness checks used before official submission. Each language contains 200 questions and 200 answers, with no empty or illegal prediction. The submitted options, reasoning sheets, and language-specific reports are mutually consistent for Arabic, Chinese, English, and Hindi. These checks indicate that the files were ready for official evaluation. They are reported

Table 8
Official final-test scores and ranks of the current method across four languages.

Language	Correct/Total	Score	Rank	Status
English	187/200	93.50%	2	Officially evaluated
Chinese	184/200	92.00%	2	Officially evaluated
Arabic	188/200	94.00%	3	Officially evaluated
Hindi	184/200	92.00%	1	Officially evaluated

together with the final test-set scores, but they do not by themselves measure task accuracy. Table 9 shows the submission-readiness summary for the four effective submissions.

Table 9
Submission-readiness summary for effective submissions. This table reports file validity, not task accuracy.

Language	Count	Empty	Illegal	Consistency
Arabic	200	0	0	Pass
Chinese	200	0	0	Pass
English	200	0	0	Pass
Hindi	200	0	0	Pass

6.3. Process-level Confidence Distribution

Table 10 reports the confidence distribution extracted from the reports and reasoning files. Figure 2 visualizes the same distribution by language and confidence tier. These values describe the review process rather than final competition performance. The overall high-confidence ratio is $742/800 = 92.75\%$, which means that most submitted answers are supported by formulas, definitions, standards, or explicit policy rules. The remaining 58 items are not counted as official errors; they are the questions where evidence is weaker and additional source recovery would be most valuable before making a ground-truth claim.

Table 10
Process-level confidence distribution across four languages. These tiers are used for audit and risk disclosure, not as official performance metrics.

Language	High	Medium	Attention	Total
Arabic	180	8	12	200
Chinese	186	10	4	200
English	184	9	7	200
Hindi	192	5	3	200
Total	742	32	26	800

6.4. Risk Taxonomy

The track-level issue tables above list the concrete question IDs and their corresponding review issues. Based on these items, the integrated reports reveal five recurring risk classes. First, missing tables or case data affect calculation tasks, such as English 044, 118, 140, 190 and Chinese 119. Second, truncated or damaged options affect CFA/finance questions such as Chinese 026, English/Chinese 070, English 155, and English 170. Third, absent statement or source text affects English/Chinese 163/179 and Arabic 086/089. Fourth, textbook or jurisdictional conventions affect Arabic Saudi government accounting, Hindi real-exchange-rate formulas, and transfer-pricing questions. Fifth, flawed option

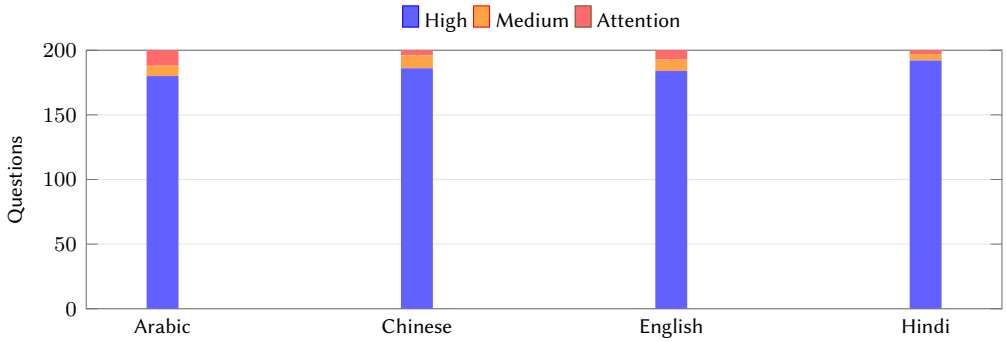

Figure 2: Process-level confidence-tier distribution by language. The figure visualizes review priority only; it is not an accuracy or leaderboard chart.

sets affect Hindi 157 and 170, where the mathematically strict result is not present. Table 11 integrates these recurring risks into representative categories and review actions.

Table 11
Representative risk categories integrated from the four language reports.

Risk class	Representative items	Review action
Missing table/case data	English 044, 118, 140, 190; Chinese 119, 148, 194	Keep answer only when recovered data or stable residual evidence supports it; mark for source recovery.
Damaged option/formula text	Chinese 026, 070; English 155, 170; Hindi 119	Use preserved mathematical structure; do not upgrade confidence unless the intended option is uniquely recoverable.
Missing statements or referenced text	English/Chinese 163, 179; Arabic 086, 089	Highest-priority manual review; current answers remain valid submissions but not final ground-truth claims.
Textbook or jurisdiction conventions	Arabic 056–070, 142, 147; Hindi 017, 097, 192; transfer-pricing items	Record the adopted accounting, policy, or examination convention explicitly.
No strictly correct option	Hindi 157, 170	Submit closest supported option and preserve limitation in report.

6.5. Autoresearch Iterations

The logged iterations show that the workflow was not a single-pass draft. It first validated Hindi template and filled submissions, then rechecked Hindi high-risk and medium-confidence sets without changing answers, added cross-file consistency checks, expanded validation to four languages, generated the English reasoning CSV, and updated the paper to treat Chinese filled JSON as the source of truth. All recorded validation statuses are pass and all kept iterations maintain consistency.

7. Discussion

7.1. Why Report Integration Matters

The four reports contain details that a short result table cannot preserve. English provides concrete answer corrections backed by recovered calculation data. Chinese clarifies that its effective source is the filled file and separates medium confidence from low confidence. Arabic contributes a jurisdiction-specific Saudi government accounting risk hierarchy and term explanations. Hindi identifies policy-timeline and formula-convention issues that are invisible to a structural validator. Integrating these reports into the paper prevents the method section from becoming detached from actual error modes.

7.2. Expert Review as Evidence Triage

The expert model's role is best understood as evidence triage rather than final authority. GPT-5.5 is used for initial answer generation, while Doubao Expert Mode is used as an independent reviewer for formulas, standards, policy timelines, and confidence tiers. The expert-review step upgrades items when formula or rule support is explicit, such as Arabic 075, 125, 126, Chinese 064, 095, English 064, 095, 106, 112, and Hindi 057. It also downgrades or flags items when missing context prevents stable proof, such as English/Chinese 163, 179. This use is compatible with the autoresearch loop because every recommendation must still pass submission-level completeness and consistency checks.

7.3. Relation to LLM Reasoning and Verification

The workflow is related to chain-of-thought prompting and self-consistency methods, which improve reasoning reliability by eliciting intermediate reasoning paths or comparing multiple outputs [6, 7]. It also shares a motivation with retrieval-augmented generation for knowledge-intensive NLP [8]: financial and policy answers need external evidence rather than unsupported generation. Our emphasis is narrower and more operational: in a shared-task setting without final labels, we use independent expert review, structured validation, and risk logging to make the submitted predictions auditable.

7.4. Limits of Accuracy Claims

The paper now reports the official final-test results rather than interim development-set evidence. The current measured results are 93.50% for English with rank 2, 92.00% for Chinese with rank 2, 94.00% for Arabic with rank 3, and 92.00% for Hindi with rank 1. These scores provide the formal competition outcome, while the confidence tiers and issue logs still serve a different role: they explain which answers were strongly supported before release and which answers remained dependent on damaged context, missing source text, or convention-sensitive interpretation. For attention items, the correct scientific stance is to preserve the answer, document why it was the best available submission, and state what original information would be needed to fully close the uncertainty.

8. Conclusion

We presented a CLEF working notes version of a multilingual financial answer-review workflow for CLEF 2026 FinMMEval Task 1. GPT-5.5 generated the final-test predictions and Doubao Expert Mode performed expert-style review. The method turns final-test answer preparation into a traceable autoresearch loop over submissions, reasoning sheets, and review reports. On the official test set, the current method obtains 93.50% for English (187/200, rank 2), 92.00% for Chinese (184/200, rank 2), 94.00% for Arabic (188/200, rank 3), and 92.00% for Hindi (184/200, rank 1). Across 800 final-test questions in four languages, all effective submissions pass submission-readiness and cross-file consistency checks. The integrated reports show that 92.75% of answers are high-confidence, while the remaining risks concentrate in missing source data, damaged options, absent statement text, jurisdiction-specific conventions, and flawed option sets. The final leaderboard results and the process-level audit trail together show that the workflow is both competitive and traceable.

Acknowledgments

This work is supported by the Guangdong Province Basic and Applied Basic Research Fund (Guangdong-Foshan Joint Funds, 2024A1515140026).

Declaration on Generative AI

During the preparation of this work, the authors utilized the DeepSeek language model and GPT-4.1 for grammar and spelling checks. Following the use of this tool, the authors carefully reviewed and edited the content as necessary and assume full responsibility for the content of this publication.

References

[1] Zhuohan Xie, Yuyang Dai, Rania Elbadry, Vanshikaa Jani, Xueqing Peng, Lingfei Qian, Georgi Georgiev, Dimitar Dimitrov, Fan Zhang, Jimin Huang, Jiahui Geng, Yankai Chen, Ye Yuan, Haolun Wu, Yuxia Wang, Ivan Koychev, Veselin Stoyanov, Mingzi Song, Yu Chen, Xue Liu, Preslav Nakov. Overview of FinMMEval 2026: Multilingual and Multimodal Financial Evaluation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer Lecture Notes in Computer Science LNCS, Jena, Germany, September 21–24, 2026.

[2] Zhuohan Xie, Yuyang Dai, Rania Elbadry, Vanshikaa Jani, Georgi Georgiev, Dimitar Dimitrov, Fan Zhang, Xueqing Peng, Lingfei Qian, Jimin Huang, Jiahui Geng, Yankai Chen, Ye Yuan, Haolun Wu, Yuxia Wang, Ivan Koychev, Veselin Stoyanov, Mingzi Song, Yu Chen, Xue Liu, Preslav Nakov. Overview of the FinMMEval 2026 Task 1: Multilingual Financial Multiple-Choice Question Answering. In *CLEF 2026 Working Notes*, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, September 21–24, 2026.

[3] Yuyang Dai, Yan Lin, Zhuohan Xie, Yuxia Wang. RealFin: How Well Do LLMs Reason About Finance When Users Leave Things Unsaid? *arXiv preprint arXiv:2602.07096*, 2026. https://arxiv.org/abs/2602.07096.

[4] Vijay Devane, Mohd Nauman, Bhargav Patel, Aniket Mahendra Wakchoure, Yogeshkumar Sant, Shyam Pawar, Viraj Thakur, Ananya Godse, Sunil Patra, Neha Maurya, Suraj Racha, Nitish Kamal Singh, Ajay Nagpal, Piyush Sawarkar, Kundeshwar Vijayrao Pundalik, Rohit Saluja, Ganesh Ramakrishnan. BhashaBench V1: A Comprehensive Benchmark for the Quadrant of Indic Domains. *arXiv preprint arXiv:2510.25409*, 2025. https://arxiv.org/abs/2510.25409.

[5] Rania Elbadry, Sarfraz Ahmad, Ahmed Heakl, Dani Bouch, Momina Ahsan, Muhra AlMahri, Marwa Elsaid khalil, Yuxia Wang, Salem Lahlou, Sophia Ananiadou, Veselin Stoyanov, Jimin Huang, Xueqing Peng, Preslav Nakov, Zhuohan Xie. SAHM: A Benchmark for Arabic Financial and Shari'ah-Compliant Reasoning. *arXiv preprint arXiv:2604.19098*, 2026. https://arxiv.org/abs/2604.19098.

[6] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *arXiv preprint arXiv:2201.11903*, 2022.

[7] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint arXiv:2203.11171*, 2022.

[8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[9] CEUR-WS.org. Policy on the Use of Generative AI Tools, 2025. https://ceur-ws.org/GenAI/Policy.html.