

Finaira at the FinEval 2026: Dense Retrieval and Validated Prompt Engineering for Cross-Lingual Financial Question Answering

Yara Nasser

Finaira

yara.nasser@finaira.ai

Ashraf Mohamed

Finaira

ashraf.mohamed@finaira.ai

Abstract

We describe our submission to FinEval 2026 Subtask 3. We compare four retrieval strategies and find dense embedding-based retrieval strongest. We then ablate further prompt-level and inference-time techniques, validating each with cross-validation; a category-conditioned router that appeared promising in-sample failed to generalize and was discarded, and we report every design decision alongside the validation method used to accept or reject it, to support reproducibility of both our results and our process. Our final development-set configuration length-constrained, few-shot, and corrected for a measured English-only citation bias achieves 0.412 ROUGE-1 under 5-fold cross-validation, and our separately tuned test-set submission achieves 0.3275 F1 on the public leaderboard (0.3593 private).

1 Introduction

Financial question answering grows harder once a system can no longer rely on a single language or document type: a model may correctly extract a revenue figure from an English filing, yet still fail if the question requires connecting it to commentary published only in Japanese or Greek. FinQA (Chen et al., 2021) and TAT-QA (Zhu et al., 2021) show numerical reasoning over financial tables is hard even monolingually; PolyFiQA (Peng et al., 2026) extends this into the multilingual, cross-document setting we address here as FinEval 2026 Subtask 3.

Given a company’s financial statements and multilingual news, a system must produce a concise, evidence-grounded answer, abstaining with *None* when no evidence supports one, since over-eager citation is penalized as heavily as missed evidence. We treat retrieval strategy as an empirical question rather than an architectural assumption, comparing no retrieval, sparse, dense, and hybrid methods and finding dense strongest. We then layer further techniques on top, holding each to one standard:

adopted only if it survives evaluation on data it was not tuned on. This surfaces a negative result of independent interest: a category-conditioned router that helped in-sample but failed under cross-validation reported here rather than discarded silently. We also track *faithfulness* and *answer groundedness* via RAGAS (Es et al., 2024) as reference-free diagnostics alongside ROUGE-1.

Our contributions: (1) a validated comparison of retrieval strategies for cross-lingual financial QA; (2) an ablation of prompt-engineering and inference-time techniques under a fully disclosed validation protocol, each checked against held-out data before adoption; and (3) a diagnosis and partial correction of measured English-only citation bias via a lightweight prompt intervention.

2 Related Work

Financial QA foundations. FinQA (Chen et al., 2021), TAT-QA (Zhu et al., 2021), and FinanceBench (Islam et al., 2023) show numerical reasoning over financial reports is hard even monolingually; none evaluate multilingual or cross-document synthesis. MultiFinBen (Peng et al., 2026) introduced PolyFiQA to close this gap, and the CLEF-2026 FinMMEval Lab (Xie et al., 2026) formalizes an equivalent evaluation of the same resource as its Task 2 distinct from our FinEval submission but scored under a comparable ROUGE-1 protocol, used here as a methodological comparison point.

Prior systems. Table 1 summarizes the top-5 FinMMEval 2026 Task 2 systems; four were prompt-only. Calibrated Signals (Jain et al., 2026) and Rastogi (Rastogi, 2026) used no retrieval, relying on prompt design and tier-aware/two-pass extraction respectively; Rastogi argues retrieval is unnecessary without testing that claim. IGT (Chiu, 2026) abandoned dense retrieval after diagnosing that fixed-size chunking returned table

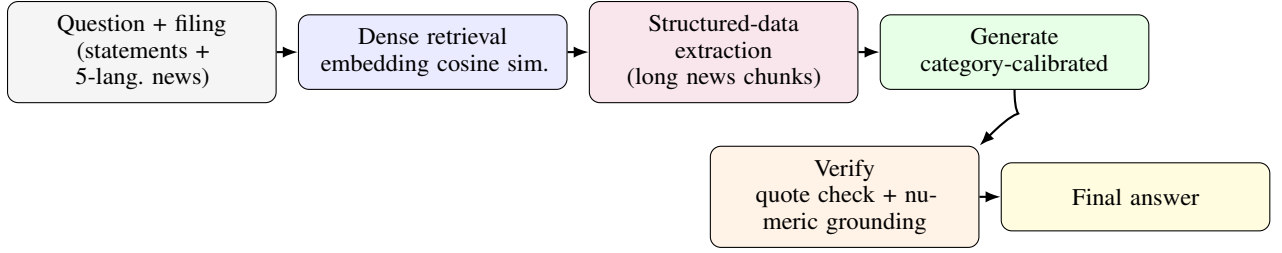


Figure 1: System pipeline: dense retrieval over the multilingual filing/news context, a structured-data extraction pass for long news chunks, generation under our validated prompt-level techniques, and a verification step before the final answer. Development-set and test-set configurations share this architecture but differ in generation-prompt calibration (Sections 4, 5).

headers rather than data, routing instead to keyword extraction. AI_TLfanClub (FiSCO-RAG) was the sole system retaining retrieval, showing lexical, CJK-aware retrieval substantially improves cross-lingual coverage (0.198→0.345 development ROUGE-1) consistent with our own finding, though we use an embedding-based mechanism. pjm mathematician attributed much of their score ceiling to ROUGE-1’s phrasing sensitivity rather than factual error, a limitation we return to in Section 6.

Positioning. We differ in selecting retrieval empirically rather than by assumption, using dense embeddings over lexical/no retrieval, and validating every design decision by cross-validation a discipline that surfaced a negative result (a category-conditioned router) otherwise indistinguishable from an improvement. We also reach comparable scores with a smaller, open-weight, and substantially cheaper model versus the frontier closed models in Table 1 (cost detail in Appendix A.4), though this spans two related but separate competitions and is not fully controlled.

#	System	R1	RAG?	Model
1	Calibrated Signals	31.18%	No	GPT-5.1
2	Rastogi	30.96%	No	Gemini 3.1 Pro
3	IGT	30.71%	No [†]	Claude Sonnet 4
4	AI_TLfanClub	30.39%	Yes	[‡]
5	pjmathematician	29.72%	No	Gemini

Table 1: Top-5 FinMMEval 2026 Task 2 systems.

[†]Abandoned retrieval after a diagnosed chunking failure.

[‡]Backbone not confirmed.

3 Task and Data

FinEval 2026 Subtask 3 asks a system to answer a question about a public company using its core financial statements (English) and news in five languages (English, Chinese, Japanese, Spanish, Greek). A valid answer fol-

lows a fixed schema, `Answer: {...}` then `News Evidence: {...}`, with *None* valid when news provides no support; which source fills which role, however, is not fixed development questions are factual, answered from the statements with news cited as evidence, while test questions are analytical, synthesized from news with statement figures cited as evidence instead (full contrast in Section 5). The data splits into a 76-row public development set with gold answers and a same-sized blind test set. Figure 2 illustrates a development-set row and its use at inference time. We analysed the development set along five axes (Table 2). The `query` field averages ~7,600 tokens (max under 11,000), comfortably within any modern LLM’s context window. 51.3% of gold answers contain `News Evidence: None`, so we treat calibrated abstention as a first-class requirement. Among gold answers citing news, 17.1% cite two or more non-English languages and 65.8% cite none, with Spanish and Chinese appearing roughly twice as often as Japanese or Greek. Every filing receives one question from each of four fixed question types revenue trend, balance sheet ratio, cash flow, and R&D ratio (the test set uses a different, four-category set). Our category router (Section 4) recognizes only three prompt categories and classifies R&D-ratio questions as `revenue_trend` on shared vocabulary, which is why Table 2 reports 38/19/19 rather than the raw 19/19/19/19; this operational grouping also defines our cross-validation strata (Appendix A.5). The set spans four companies across 19 filing dates. Gold-answer formatting is not fully uniform: 31.6% omit the leading `Answer: prefix`, and the evidence label varies across three conventions, identically across categories authoring inconsistency rather than an exploitable signal.

These findings motivate the system design in

Statistic	Value
Development rows	76
Mean context length	~7,622 tokens
News Evidence: None rate	51.3%
Cites 2+ languages	17.1%
Question categories [†]	3 (38/19/19)
Companies / filing dates	4 / 19
Answer: prefix present	68.4%

Table 2: Summary statistics, 76-row development set.

[†]Pipeline-recognized prompt categories; the raw dataset has 4 question types, 19 each (Section 3).

Section 4: retrieval that matches evidence across languages with no shared vocabulary, calibration toward abstention over forced citation, and category-aware prompting.

4 System Description

Given a company’s financial statements and multilingual news (five languages), the system answers analytical questions with a synthesized answer plus an evidence citation, allowing *None* when no evidence supports one.

The architecture is a four-stage pipeline built with DSPy: **Parsing** splits the input into statement and news blocks and categorizes each question with a target answer length; **Retrieval** uses dense embedding similarity to select relevant news chunks, always including the full financial statements; **Structured-data extraction** is a dedicated pass that pulls embedded tables out of long news chunks, added because prompting alone could not reliably surface them; and **Generation and verification** produces the answer and citation via a one-shot-per-category signature, then deterministically checks that evidence is verbatim-present and cited dollar figures are grounded in the statements though this check does not cover derived values such as percentages or growth rates. Full design rationale is given in Appendix A.2.

5 Experimental Setup

5.1 Dataset

We use PolyFiQA (Peng et al., 2026) (development set described in Section 3). The test set differs in question shape, not just label availability: development questions are factual, answered from statements with news as evidence; test questions are analytical, synthesized from news with statement figures as evidence so development cross-validation does not directly estimate test-set perfor-

mance. The test set spans four categories (19 each), distinct from the development set’s three.

5.2 Model Selection

We evaluated four models; for the final submission, Qwen2.5-72B-Instruct (dense, not MoE) via OpenRouter, satisfying the task’s 70B-parameter, OpenRouter-only constraint for API-based solutions. Qwen’s 72.7B count was confirmed acceptable in writing by the organizers. Beyond the parameter constraint, this choice tests whether a smaller open-weight model can match proprietary frontier models without their training-cost, licensing, or per-token pricing advantages (Appendix A.4). Full details in Appendix A.3.

5.3 Infrastructure

Retrieval runs on symmetric embeddings (the OpenRouter endpoint does not support asymmetric mode). The development-tier config (top-3, 40% threshold) was 5-fold cross-validated; the test-set config was separately re-tuned via a grid sweep against development-set gold answers (not cross-validation, since it selects one configuration against fixed labels). Subsequent test-set changes were each validated as a separate leaderboard submission. We disclose full sweep parameters and CV grouping in Appendix A.5 to support independent reproduction.

5.4 Diagnostic Methodology

We compute four reference-free diagnostics: numeric-grounding rate, evidence-verification rate, abstention rate, and a context-utilization review the last of which surfaced the long-context extraction gap motivating our extraction pass (Section 4). We also decompose ROUGE-1 F1 into precision/recall.

6 Results

6.1 Retrieval Method Comparison

Dense retrieval outperforms sparse (BM25), no retrieval, and a hybrid combination (Table 3), consistent with its ability to match semantically related cross-lingual content that sparse retrieval cannot.

6.2 Category-Conditioned Routing

A category-to-method router looked promising naively (0.3751 vs. 0.3732 for dense alone) but scored worse (0.3611) under honest 5-fold cross-validation, with the selected mapping unstable

Method	ROUGE-1
BM25 (sparse)	0.3427
No retrieval	0.3654
Hybrid (sparse + dense)	0.3653
Dense (adopted)	0.3732

Table 3: Retrieval method comparison on the development set.

across folds evidence of insufficient per-category data rather than a genuine signal. We report the router as rejected.

6.3 Development-Tier Result

Our final development-tier system scores **0.4120** ROUGE-1 under 5-fold cross-validation (0.4226 in-sample).

6.4 Test-Set Result

The test-set submission scores **0.3275** ROUGE-1 F1 on the public leaderboard (confirmed at **0.3593** private), reached via a development-set retrieval sweep plus three changes each validated on a separate leaderboard submission (Table 4). Full ablation history and model comparison are in Appendix A and A.6.

Change	F1	P	R
Per-category length calibration	0.3189	0.336	0.352
+ retrieval breadth re-tuning	0.3241	0.362	0.345
+ structured-data extraction	0.3275	0.350	0.355

Table 4: Test-set public leaderboard progression (private-split score: 0.3593).

7 Analysis

Our diagnostic methodology (Section 5) surfaces six findings, each from the model’s own output/verification signals rather than a reference answer; the three most incremental are given in Appendix A.8.

The system’s abstention rate is 0% on both the test set and internal evaluation, despite *None* being an explicitly valid, sometimes-correct answer a calibration gap: the instruction to answer from either source whenever either is remotely relevant is likely too permissive.

A long-context extraction failure, not a knowledge or prompting gap, specifically limited the revenue-driver category. Earnings-release news frequently embeds structured segment-results tables within long prose that the model failed to ex-

tract from, even under direct instruction a genuine long-context attention limitation, not a phrasing one. Separating “find the structured data” into its own focused pass (Section 4) did change the output on the targeted cases and produced a real improvement (F1 0.3241 to 0.3275) for this class of failure, restructuring what the model has to search through mattered more than instructing it to search harder.

8 Conclusion

This work presents two related systems for FinEval 2026 Subtask 3: a development-set system validated at 0.4120 ROUGE-1 F1 (5-fold CV), and a separately tuned test-set system scoring **0.3275** on the public leaderboard (**0.3593** private, confirming the result on unseen data). Three changes to the test-set system length calibration, retrieval re-tuning, and structured-data extraction were each confirmed as real improvements via separate leaderboard submissions. Every change, on either set, was adopted only after evidence it generalized beyond the data used to select it, a standard that also surfaced a negative result: a category-conditioned router that helped in-sample but failed under cross-validation. Main contributions: a retrieval-then-verify architecture with independent evidence checking, category-calibrated length control, a reference-free diagnostic methodology that surfaced a long-context extraction failure, and evidence that restructuring the pipeline resolved that failure where prompting alone did not.

Limitations

Model comparability. Only Qwen2.5-72B and Llama-3.3-70B have confirmed parameter counts; scale claims hold only for that pair, and even it is not fully controlled since Llama was evaluated at an earlier, less-tuned configuration.

Retrieval and verification gaps. Retrieval runs on symmetric embeddings only (OpenRouter lacks asymmetric support). Verification confirms literal statement values but not derived figures (margins, growth rates), concentrated in two of four test-tier categories.

Calibration and evaluation rigor. Abstention rate is zero, suggesting possible over-eagerness to answer; untested fixes carry regression risk. Test-tier changes are validated against single leaderboard submissions rather than repeated runs, a weaker standard than the cross-validated development tier. Because development and test questions

differ in shape and evidence direction (Section 3), the development-tier’s 0.412 cross-validated score should not be read as predictive of test-tier performance the two are validated and reported as distinct results for exactly this reason.

Coverage. The extraction pass helps only when relevant data is actually present in retrieved news; Greek-language performance was not independently benchmarked.

Reproducibility. Full pipeline code, including rejected-candidate implementations from the appendix, is available in our Final Solution Code submission.

References

- Zhiyu Chen, Wenhu 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. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.
- Yuwen Chiu. 2026. IGT @ FinMMEval 2026 task 2: Question-type prompting with targeted extraction for multilingual financial QA. In *CLEF 2026 Working Notes, CEUR Workshop Proceedings*.
- Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*.
- Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.
- Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. FinanceBench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.
- Gaurav Jain, Hitesh Hitesh, Sparsh Sundriyal, Sachit Bhardwaj, Himanshu Behl, Prabal Gautam, and Steven Zimmerman. 2026. Calibrated signals @ FinMMEval 2026 task 2: Inference-only strategies for cross-lingual financial question answering. In *CLEF 2026 Working Notes, CEUR Workshop Proceedings*.
- Xueqing Peng, Lingfei Qian, Yan Wang, Ruoyu Xiang, Yueru He, Yang Ren, Mingyang Jiang, Vincent Jim Zhang, Yueqi Guo, Jiahao Zhao, et al. 2026. MultiFinBen: Benchmarking large language models for multilingual and multimodal financial application. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Pranshu Rastogi. 2026. Pranshu rastogi @ FinMMEval 2026: Systems for tasks 1 and 2 – prompt-engineered Gemini for multilingual and cross-lingual financial reports QA. In *CLEF 2026 Working Notes, CEUR Workshop Proceedings*.
- Zhuohan Xie, Yuyang Dai, Rania Elbadry, Vanshika Jani, Xueqing Peng, Lingfei Qian, Georgi Georgiev, Dimitar Dimitrov, Fan Zhang, Jimin Huang, et al. 2026. The CLEF-2026 FinMMEval lab: Multilingual and multimodal evaluation of financial AI systems. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of CLEF 2026*.
- 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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.

A Supplementary Material

This appendix provides supplementary details on the dataset, system design, experimental setup, results, and analysis. Ablation scores are internal diagnostic measurements against a local reference and are reported for transparency. They are distinct from the leaderboard-validated progression in Table 4 (0.3189 $\rightarrow$ 0.3241 $\rightarrow$ 0.3275 public leaderboard, confirmed at 0.3593 on the private split), which remains our sole source of out-of-sample evidence for the deployed system.

A.1 Dataset Structure (Task and Data)

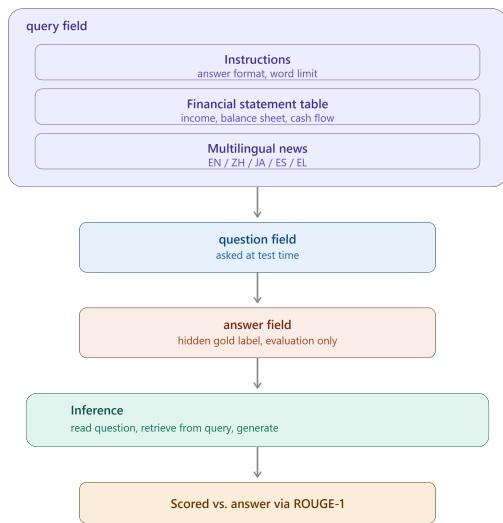


Figure 2: Structure of a FinEval 2026 Subtask 3 development-set instance and its use at inference time.

A.2 System Design Rationale (System Description)

Retrieval is precision-oriented, not exhaustive. A filtered set of relevant news outperformed supplying all available news unfiltered.

The generation instruction permits grounding in the financial statements as well as the news, despite the task framing centering the answer on news synthesis. An earlier, more restrictive version produced spurious refusals on questions particularly capital expenditures where the relevant figure was present in the statements but not stated explicitly in the news.

One worked example per category, hand-authored rather than drawn from any dataset row, illustrates the expected answer/evidence shape for that question type.

Evidence is verified independently of the model’s own claim, checked directly against source text rather than trusted from the model’s report.

Answer length is capped per category, calibrated to each category’s typical target length rather than a single value applied uniformly, motivated by a precision/recall imbalance identified during evaluation (Appendix A.8).

Retrieval breadth was re-tuned specifically for this system rather than inherited unchanged from the development-tier configuration, using the methodology described in Appendix A.5.

A.3 Model Selection Detail (Experimental Setup)

Model	Access	Role
Gemini Flash	Direct API	Development
Claude Sonnet 5	Direct API	Paired comparison
Qwen2.5-72B	OpenRouter	Final submission
Llama-3.3-70B	OpenRouter	Fallback check

Table 5: Candidate models evaluated during development.

A.4 Cost Comparison (Experimental Setup)

Table 6 compares per-token OpenRouter pricing (retrieved 2026-08-26) for Qwen2.5-72B-Instruct against the three frontier closed models appearing in Table 1, sourced from the same provider for a like-for-like comparison rather than each vendor’s own listed rate. Qwen2.5-72B is roughly 3–8 $\times$ cheaper per input token and 25–37 $\times$ cheaper per output token than any of the three, output tokens being the more consequential axis since generation, not the fixed prompt, dominates per-query cost in our pipeline. This is a list-price comparison, not a controlled measurement of our own inference spend, and does not account for provider-side discounts, batching, or caching available to any of the four models.

Model	Input \$/M	Output \$/M
Qwen2.5-72B-Instruct	\$0.36	\$0.40
GPT-5.1	\$1.25	\$10.00
Gemini 3.1 Pro	\$2.00	\$12.00
Claude Sonnet 4	\$3.00	\$15.00

Table 6: OpenRouter list pricing per million tokens, retrieved 2026-08-26.

A.5 Retrieval Sweep and Cross-Validation Grouping (Experimental Setup)

The test-set retrieval configuration was selected via a grid sweep over $\text{top-}k \in \{1, 2, 3, 5\}$ and threshold $\in \{0.3, 0.4, 0.5, 0.6\}$, scored against public development-set gold answers; top-2 chunks at a 50% threshold scored highest (0.3806 mean ROUGE-1) and was adopted.

The development-tier’s 5-fold cross-validation stratifies by question category only, not company or filing date since each filing contributes one row per category, a filing can appear in a training fold for one category while held out for another. The category used for stratification is itself the pipeline’s 3-way keyword router (Section 3), which folds R&D-ratio questions into `revenue_trend`; combined with the filing overlap above, we read the reported 0.412 as a plausibly optimistic upper estimate of development-tier generalization rather than a fully independent validation.

A.6 Model Comparison (Results)

Table 7 reports development-time model exploration, not a controlled comparison: Gemini and Claude (undisclosed parameter counts) were evaluated on the development set, Qwen and Llama on the test set, and Llama’s run predates the length, retrieval, and extraction refinements later added for Qwen. The only supported claim is that Qwen was selected as primary and continued to receive tuning, not that it strictly outperforms Llama under identical settings.

Model	Params	Access	Score
Gemini	undisclosed	Direct	0.4120 [†]
Claude Sonnet 5	undisclosed	Direct	0.3032 [†]
Qwen2.5-72B	72.7B	OpenRouter	0.3275 [‡]
Llama-3.3-70B	70.6B	OpenRouter	0.2859 [‡]

Table 7: Development-time model exploration. [†]Development set. [‡]Test set, public leaderboard; Llama at an earlier configuration.

A.7 Full Ablation Study (Results)

We report the complete record of techniques evaluated during system development, including candidates that were tested and rejected.

A.7.1 Development-Phase Sweep

Table 8 reports the sequence of techniques tried during initial system development, in the order evaluated. Each row’s Δ is relative to the preceding

adopted configuration, so the table reads as a running trace of the design process rather than a set of independent comparisons.

Technique	R-1	Δ	Verdict
First baseline (no retrieval, plain signature)	0.3052	-	-
60-word length cap	0.3043	−0.0009	Rejected
ROUGE-aware framing	0.3059	+0.0007	Rejected
Verbatim-quote evidence instruction	0.2968	−0.0084	Rejected
No few-shot	0.3023	−0.0029	Rejected
Dense retrieval (vs. unfiltered)	0.3159	+0.0107	Adopted
Context hints (company + filing date)	0.3151	+0.0099	Candidate
Dense + context hints combined	0.3076	+0.0024	Rejected
Dense + allow-statements signature ¹	0.3300	+0.0141	Adopted
Period captions on ambiguous column headers	0.3482	−0.0065	Rejected
Structured row-pointer evidence extraction	0.3550	+0.0003	Flat
Self-consistency / MBR ($N=3$, $T=0.7$)	0.3501	−0.0046	Rejected
Per-category word budgets (3555w)	0.3711	+0.0164	Adopted

Table 8: Development-phase technique sweep using an internal diagnostic reference. R-1 denotes ROUGE-1.

A.7.2 Retrieval- and Sampling-Side Interventions

Since the shipped system led on precision but trailed on recall, we tested five recall-targeting techniques HyDE retrieval (Gao et al., 2023), multi-query expansion with RRF, each combined with a length cap, and self-consistency/MBR selection (Table 9). Retrieval reformulations raised recall to 0.410.42 but inflated output to ~ 65 words, costing more precision than the recall gain offset; adding a length cap overcorrected to ~ 30.5 words, losing recall instead. Self-consistency showed the same length-inflation pattern, attributed to ROUGE-1-based candidate agreement favoring longer, more generic candidates. None of the five techniques improved on the shipped system, and none were submitted to the leaderboard.

¹Explicitly permits the model to pull figures from the statements when retrieved news does not state them; addressed a high false-abstention rate observed specifically on *capex_significance* questions.

²Gao et al. (2023): embeds an LLM-generated hypothetical answer to the question, rather than the question itself, as the dense-retrieval query vector.

³Generates three LLM paraphrases of the question, retrieves separately for each, and fuses the resulting rankings via Reciprocal Rank Fusion.

⁴Samples five candidates at $T=0.7$, verifies each independently, and selects the candidate with the highest mean pairwise ROUGE-1 agreement with the other four; a reference-free MBR proxy since gold answers are unavailable at inference time on the blind test set.

Config	F1	P	R	W
Shipped system	0.3483	0.3689	0.3642	41.4
HyDE retrieval ²	0.3190	0.2803	0.4185	65.4
Multi-query + RRF ³	0.3160	0.2789	0.4109	65.1
HyDE + per-category cap	0.3300	0.3823	0.3142	30.6
Multi-query + length cap	0.3223	0.3759	0.3088	30.5
Self-consistency / MBR ⁴	0.3286	0.2908	0.4230	65.0

Table 9: Retrieval- and sampling-side experiments targeting the precision/recall trade-off, against the internal diagnostic reference.

A.8 Additional Analysis Findings (Analysis)

Precision trailed recall in the test-tier system

(0.325 vs 0.469 internally): answers ran longer than the target style rewards, which unigram precision penalizes directly the opposite imbalance from an earlier, too-terse Gemini-based version, illustrating that failure modes shift across generation models even with the rest of the pipeline fixed.

Evidence-verification rate varies sharply by category: near ceiling for capex and capital-allocation questions, markedly lower for profit-margin and revenue-driver questions, which more often cite a *derived* metric (a margin, a growth rate) rather than a literal statement line item verbatim-quote verification cannot confirm a computed value regardless of whether the computation is correct. A prompt fix instructing citation of the raw figures behind a computed claim was tested and did not close this gap, indicating a structural verification limitation rather than a phrasing one.

Per-category length calibration was the one technique with a clear, isolated positive effect (+0.0058 test-tier ROUGE-1). An evidence-citation phrasing fix for the finding above was rejected as net negative; a self-consistency/resampling approach was tested at more than one stage and rejected each time as flat-to-negative.

Retrieval breadth inherited from the development-tier configuration was not optimal for the test-tier system: re-tuning to a narrower setting produced a real leaderboard improvement (F1 0.3189 to 0.3241; Table 4).