Calibrated Signals @ FinMMEval 2026 Task 2: Inference-Only Strategies for Cross-Lingual Financial Question Answering

Gaurav Jain[1,*,†], Hitesh Hitesh[1,†], Sparsh Sundriyal[1,†], Sachit Bhardwaj[1,†], Himanshu Behl[1,†], Prabal Gautam[1,†] and Steven Zimmerman[2,†]

[1]SBO Bangalore
[2]SHELL AUSTRIA GESELLSCHAFT M.B.H

Abstract

We describe an inference-only system for cross-lingual financial QA, which requires concise ($\leq$100 words) answers to financial questions grounded in English SEC filings and parallel news in five languages (EN, ZH, JA, ES, EL). Our setup is intentionally inference-only (no model fine-tuning), to match the realistic resource constraints of a shared-task setting. We evaluate eight prompting strategies (P0–P7) over a single Azure AI Foundry agent, four inference-time techniques (Structured IR, Self-Refine, LLM-as-Judge, Cross-lingual Consistency), and a reward-guided Best-of-N pipeline that generates DPO preference pairs as a by-product. P7 (ROUGE-Optimised) is the best single-prompt strategy (R1=0.346/0.341 Easy/Expert); LLM-as-Judge filtering gives the largest single-technique gain (+0.119 R1 on Easy); the Best-of-N oracle reaches R1=0.413/0.411. An English-bias rate of 41.3% and a language-coverage reward of 0.12/3.00 suggests that current prompting alone is insufficient for robust cross-lingual synthesis. For blind-test submission, we use P7 with strict format enforcement as the best deployable configuration.

Keywords

multilingual financial QA, prompting ablation, Reward-Guided Best-of-N, DPO preference data, ROUGE-1, Azure AI Foundry, cross-lingual evaluation

1. Introduction

The benchmark used in this study [1, 2] is a cross-lingual financial QA benchmark, pairing English 10-K/10-Q excerpts with news in Chinese, Japanese, Spanish, and Greek (344 instances across Easy and Expert tiers). Existing English-only benchmarks such as FinQA [3] and TAT-QA [4] do not evaluate cross-lingual synthesis. Because working notes are submitted before blind test labels are released, our objective is to identify the strongest deployable inference strategy on development data and then freeze that strategy for the external evaluation portal. Our work makes three contributions: (1) a systematic prompt-engineering ablation over eight strategies; (2) four inference-time augmentation techniques including a novel reward-guided Best-of-N pipeline that produces DPO preference pairs [5] as a side-product (no policy optimisation is run; DPO fine-tuning is left to future work); and (3) a per-language marginal-contribution ablation that quantifies the information value of each non-English source.

Score context. For cross-lingual financial QA, ROUGE-1 in the range 0.30–0.45 is competitive: English-only FinQA systems report 0.35–0.55 [3] on extractive tasks, but cross-lingual synthesis across five languages over structured tables is substantially harder. Our zero-shot P0 baseline (R1$\approx$0.30) suggests

CLEF 2026: Conference and Labs of the Evaluation Forum, 2026
*Corresponding author.
†These authors contributed equally.
g.jain3@shell.com (G. Jain); hitesh.hitesh@shell.com (H. Hitesh); Sparsh.Sundriyal@shell.com (S. Sundriyal); Sachit.Bhardwaj@shell.com (S. Bhardwaj); Himanshu.Behl@shell.com (H. Behl); Steven.Zimmerman@shell.com (S. Zimmerman)
0009-0006-9451-0061 (G. Jain); 0000-0001-9061-4388 (H. Hitesh); 0000-0003-0644-7500 (S. Bhardwaj); 0009-0000-0727-8185 (H. Behl); 0000-0002-4686-2066 (S. Zimmerman)

the model already extracts relevant financial figures; our best single-prompt system P7 (R1=0.346/0.341) improves by ∼15% over baseline; and the Best-of-N oracle (R1=0.413) achieves a ∼40% relative gain over P0. The primary bottleneck is not scores per se but the 41.3% English bias rate and near-zero language-coverage reward, indicating the model rarely cites non-English sources even when instructed to do so.

2. Dataset

Table 1 summarises the dataset used for development.

Table 1
Dataset statistics.

Tier	Instances	Languages	Source	Ans. limit
Easy	172	EN+ZH+JA+ES+EL	SEC 10-K/10-Q	100 words
Expert	172	EN+ZH+JA+ES+EL	SEC 10-K/10-Q	100 words
Total	344	5		

Each instance provides: (i) structured financial tables (income statement, balance sheet, cash flow), (ii) five parallel news articles, and (iii) a gold answer in the schema `Answer: {...} / [News|Financial Statements] Evidence: {...}`. The task is part of the FinMMEval 2026 lab [1] and the development set is drawn from the MultiFinBen benchmark [6]. We have not used any other external dataset in this study.

3. System Architecture

Figure 1 shows the end-to-end inference pipeline. All experiments share a single Azure AI Foundry agent (OpenAI-compatible Responses API) with exponential back-off retry, backed by model **gpt-5.1** (version `2025-11-13`). Post-processing extracts the `Answer:` section and truncates to 100 words before ROUGE scoring. As the model is accessed via a commercial pay-as-you-go Azure AI Foundry deployment, we are unable to report exact computational costs for any of the experiments described in this paper.

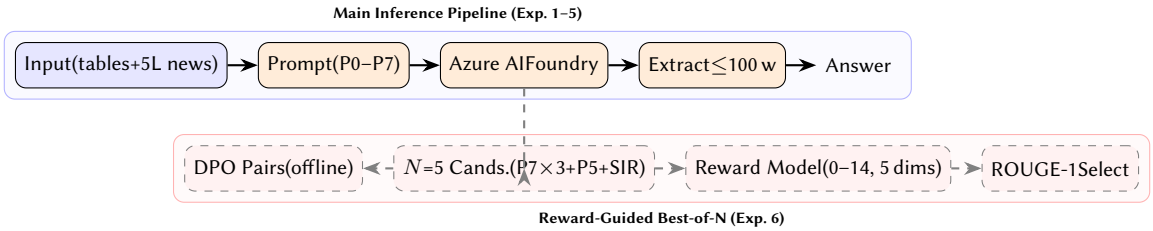

Figure 1: Inference pipeline. **Blue region** (solid arrows): main path for Exp. 1–5 — query and system prompt combined, sent to Azure AI Foundry, output post-processed (extract `Answer:` section, truncate to 100 words). **Red region** (dashed arrows): Exp. 6 Reward-Guided Best-of-N — N=5 candidates scored by a 5-dim reward model; ROUGE-1 winner kept and refined; all (winner, loser) pairs logged as DPO preference data.

4. Prompting Strategies

Table 2 summarises all eight strategies. P7 is the best single-strategy result; P5 was the originally intended primary submission but underperformed P0 on Expert (see Exp. 2 results).

Table 2
Prompting strategies (P0–P7).

ID	Name	Key additions over P0
P0	Zero-shot	No system prompt
P1	Role	Senior multilingual analyst role; grounding constraint
P2	Multilingual	P1 + explicit per-language extraction rules
P3	CoT	8-step evidence extraction chain-of-thought [7]
P4	Tier-Aware	Conditional routing: tables-first (factual) vs news-first (analytical)
P5	Combined	P2 + P3 + P4 unified (primary run)
P6	Few-Shot	2 gold examples prepended [8]
P7	ROUGE-Opt	Explicit instructions to re-use question key-terms and exact numbers (full prompt: Appendix A)

5. Inference-Time Techniques and Reward-Guided Best-of-N

Table 3 gives a concise description of each technique.

Table 3
Inference-time techniques (Exp. 4–6).

Technique	Description
Structured IR (SIR)	Two-step prompt: extract per-source scratchpad → synthesise. Encourages explicit non-English attention.
Self-Refine	Generate (P5) → critique (5 axes) → improve; up to 2 iterations; best ROUGE-1 retained [9].
Judge Filtered	Generate → LLM judge scores on 4 dims: groundedness (0–3), language coverage (0–3), format compliance (0–2), conciseness (0–2); total 0–10; retry $\leq 2\times$ with P5 if total < 6; highest-ROUGE-1 response kept. Full rubric: Table 4; full prompt: Appendix B [10].
Consistency Check	Parallel multilingual vs English-only generation; LLM classifies English bias; reports bias rate.
Reward-Guided Best-of-N	N=5 candidates (P7$\times$3 + P5 + SIR) → reward model (0–14) + ROUGE-1 selection → ROUGE refinement pass; (winner, loser) pairs logged as DPO preference data for future fine-tuning [5].

Reward model dimensions. The reward model scores each candidate on five axes: factual accuracy (0–4), groundedness (0–3), language coverage (0–3), format compliance (0–2), and conciseness (0–2), giving a maximum **LLM reward of 14**. In Exp. 6 a ROUGE-1 bonus (max 12) is added, giving a **combined score of 26** (reported as Reward/26 in Tables 10–12); all reward values in Exp. 4 use the 14-point LLM-only scale. Table 4 gives full per-level anchor descriptions for both the judge (Exp. 4) and the reward model (Exp. 6). Final selection uses raw ROUGE-1 rather than the reward score; therefore, Best-of-N measures the best reference-informed selection available within this candidate pool, but cannot be deployed unchanged on the blind test set.

Blind-test deployability note. Exp. 6 is an oracle-style analysis on development data because candidate selection depends on reference-based ROUGE. This signal is unavailable on the final blind

Table 4
Scoring rubrics for the LLM-as-Judge (Exp. 4, 4 dims, total 0–10) and the Reward Model (Exp. 6, 5 dims, total 0–14). Retry threshold for judge filtering: total $< 6/10$.

Dimension	Judge	Reward	Per-level anchors
Factual accuracy	—	0–4	0 hallucinated facts; 1–2 partial errors; 3 mostly correct; 4 fully accurate
Groundedness	0–3	0–3	0 unsupported claims; 1 mostly grounded; 2 fully grounded; 3 grounded + source cited
Language coverage	0–3	0–3	0 EN only; 1 mostly EN; 2 $\geq$2 non-EN sources; 3 all 5 languages consulted
Format compliance	0–2	0–2	0 missing Answer/Evidence; 1 partial; 2 perfect schema
Conciseness	0–2	0–2	0 $>$100 words; 1 $\leq$100 words but verbose; 2 concise & complete
Total	**0–10**	**0–14**	Retry if judge total < 6; reward bonus: +12 ROUGE-1 $= 26$ combined

test set. Therefore, our official submission strategy is the best deployable method from Exp. 2 (P7), combined with deterministic decoding and strict output-schema post-processing.

6. Experimental Results

6.1. Exp. 1 – Baseline

We first establish a zero-shot baseline by sending each query directly to the model with no explicit system prompt (P0). Easy (R1=0.294, RL=0.226) and Expert (R1=0.308, RL=0.190) score similarly, suggesting the model already extracts relevant financial figures from structured tables even without instructions. The higher ROUGE-2 on Easy (0.119 vs 0.096) reflects that Easy questions are more factual and extractive, producing greater two-gram overlap with gold answers (Table 5).

Table 5
Exp. 1: Zero-shot baseline (P0).

Tier	ROUGE-1	ROUGE-2	ROUGE-L
Easy	0.2937	0.1192	0.2257
Expert	0.3084	0.0964	0.1900
Avg	0.3011	0.1078	0.2079

Note on P0 score variation. The P0 baseline appears with slightly different values across tables (Easy R1: 0.2904–0.2939; Expert R1: 0.3058–0.3108) because each experiment was run as an independent API call without a fixed random seed; run-to-run variance from non-deterministic sampling at the default temperature likely contributes. All comparisons within a single experiment use the same corresponding P0 run.

6.2. Exp. 2 – Prompting Ablation

We evaluate all eight strategies on the full 172-instance development set per tier. Several findings stand out. First, simple role instructions (P1) slightly hurt both tiers (-0.009/-0.001 ΔR1), suggesting the model is already well-grounded without explicit role framing. Second, strategies that mandate per-language reading (P2, P3, P4) consistently improve over P0 on Easy but underperform on Expert.

Third, P5—our originally intended primary submission—*underperforms P0 on Expert by −0.019 ΔR1*: the combined role+CoT+tier-routing prompt adds conflicting constraints that confuse the model on analytical questions. Finally, P7 (ROUGE-Opt) achieves the best cross-tier average (R1=0.344) by instructing the model to reuse question key-terms and exact numeric values; P6 (Few-Shot) wins on Easy (R1=0.350) through gold-answer format calibration but drops on Expert (0.297), giving a lower cross-tier average of 0.324. Table 6 reports the full per-strategy breakdown.

Table 6
Exp. 2: ROUGE-1/L per strategy (Δ vs P0 per tier). **Bold** = best per tier; P7 wins overall by avg R1.

Strategy	Easy R1	Easy RL	ΔR1	Expert R1	Expert RL	ΔR1
P0 Zero-shot	0.2939	0.2265	0.000	0.3058	0.1878	0.000
P1 Role	0.2853	0.2192	−0.009	0.3047	0.1876	−0.001
P2 Multilingual	0.3064	0.2352	+0.013	0.3029	0.1823	−0.003
P3 CoT	0.3007	0.2329	+0.007	0.3007	0.1826	−0.005
P4 Tier-Aware	0.3120	0.2322	+0.018	0.3010	0.1835	−0.005
P5 Combined	0.3041	0.2349	+0.010	0.2864	0.1740	−0.019
P6 Few-Shot	**0.3500**	**0.2644**	+0.056	0.2973	0.1820	−0.009
P7 ROUGE-Opt	0.3461	0.2679	+0.052	**0.3411**	**0.2201**	+0.035

6.3. Exp. 3 – Per-Language Ablation

To quantify each non-English source's marginal contribution we run the best strategy from Exp. 2 six times per tier: once with the full five-language context (baseline), then with each of ZH, JA, ES, EL removed in turn, and finally with English only. A negative Δ indicates the removed language provided useful signal.

Table 7
Exp. 3: Per-language ablation with strategy P7 (best from Exp. 2, avg R1=0.3436) — ROUGE-1 and ROUGE-L (Δ vs full context; negative = language helps).

Tier	Condition	ROUGE-1	ΔR1	ROUGE-L	ΔRL
Easy	Full (all 5 lang)	0.3505	0.000	0.2685	0.000
Easy	−ZH (Chinese)	0.3540	+0.004	0.2715	+0.003
Easy	−JA (Japanese)	0.3484	−0.002	0.2659	−0.003
Easy	−ES (Spanish)	0.3488	−0.002	0.2686	+0.000
Easy	−EL (Greek)	0.3516	+0.001	0.2680	−0.001
Easy	EN only	0.3366	−0.014	0.2569	−0.012
Expert	Full (all 5 lang)	0.3373	0.000	0.2144	0.000
Expert	−ZH (Chinese)	0.3337	−0.004	0.2128	−0.002
Expert	−JA (Japanese)	0.3360	−0.001	0.2168	+0.002
Expert	−ES (Spanish)	0.3325	−0.005	0.2162	+0.002
Expert	**−EL (Greek)**	0.3272	−0.010	0.2093	−0.005
Expert	EN only	0.2923	−0.045	0.1920	−0.022

On Easy, removing any single non-English language has minimal impact ($|\Delta R1| \leq 0.004$), showing no single source dominates. Dropping to English-only reduces Easy R1 by 0.014, indicating the non-English articles collectively add signal. On Expert the pattern is sharper: removing Greek (EL) causes the largest single-language drop (−0.010 R1, −0.005 RL) despite Greek being the lowest-resource language in the set, suggesting the Greek news contains analytical commentary (margins, ratios) absent from other sources. Dropping to English-only produces the most severe degradation (−0.045 R1, −0.022 RL), suggesting that multilingual context is genuinely necessary for analytical reasoning on Expert questions (Table 7).

6.4. Exp. 4 – Beyond Prompting

Having identified P7 as the best single prompt, we tested whether inference-time techniques could push quality further *without* changing the underlying model. Four techniques were evaluated on a full 172-instance run per tier: (i) **Structured IR (SIR)** — a two-step prompt that extracts a per-source scratchpad before synthesising the final answer, forcing explicit non-English attention; (ii) **Self-Refine** — generate with P5, produce a self-critique on five quality axes, and iteratively improve (up to 2 passes), retaining the best ROUGE-1 output [9]; (iii) **Judge Filtered** — generate with P5, score the output with an LLM judge on four dimensions (0–10), and retry up to twice with a stricter prompt if the score falls below 6, keeping the highest-ROUGE-1 response [10]; (iv) **Consistency Check** — run P5 twice in parallel (multilingual vs English-only) and use an LLM classifier to flag responses that ignore non-English sources, logging the overall English bias rate.

LLM-as-Judge filtering is the standout performer (+0.119 R1 on Easy, +0.051 on Expert over P0), demonstrating that automatic quality gating provides a large, reliable signal even without gold labels. Self-Refine also helps on Easy (+0.044) but provides minimal gain on Expert (+0.003), suggesting the self-critique loop is less effective for multi-step analytical questions.

The English bias rate of 41.3% (71/172 Easy responses) indicates that even judge-filtered outputs frequently default to English-centric synthesis (Table 8).

Table 8
Exp. 4: Inference-time techniques vs P0 baseline.

Technique	Tier	R1	R2	RL	ΔR1 vs P0
P0 Baseline	Easy	0.2904	0.1143	0.2223	0.000
Structured IR	Easy	0.3018	0.1309	0.2399	+0.011
Self-Refine	Easy	0.3339	0.1353	0.2543	+0.044
Judge Filtered	Easy	**0.4096**	**0.2302**	**0.3202**	+0.119
P0 Baseline	Expert	0.3108	0.0976	0.1882	0.000
Structured IR	Expert	0.3253	0.1068	0.1971	+0.015
Self-Refine	Expert	0.3136	0.0984	0.1890	+0.003
Judge Filtered	Expert	**0.3614**	**0.1540**	**0.2401**	+0.051
English Bias Rate (172 Easy rows)		41.3% (71/172 biased)			

6.5. Exp. 5 – Error Analysis

To understand *why* the model fails, we manually inspected the 40 predictions with the lowest ROUGE-1 scores (R1 < 0.3) from the P7 full-dataset run and categorised each into one of five error types. This taxonomy directly informs the design of the reward model dimensions used in Exp. 6.

Hallucination is the dominant failure mode, accounting for 24/40 cases (60%): the model produces plausible-sounding financial figures or trends that are absent from, or directly contradict, the provided tables. This is particularly common in responses to comparative questions spanning multiple fiscal periods, where the model conflates different reporting dates. **English bias** (12/40, 30%) occurs when the model produces a valid answer grounded only in the English 10-K/10-Q data while ignoring corroborating or divergent signals in the non-English news, leading to lower vocabulary overlap with gold answers that cite multilingual sources. **Vocabulary mismatch** (4/40, 10%) arises when correct numeric values are expressed with different phrasing than the gold answer (e.g. "up 15%" vs "15% year-over-year growth"), which ROUGE-1 penalises unfairly. Zero format violations show in our experiments that the post-processing pipeline reliably extracts the `Answer:` section.

These findings motivate three reward dimensions for Exp. 6: *factual accuracy* (targets hallucination), *language coverage* (targets English bias), and *groundedness* (targets unsupported claims) (Table 9).

Table 9
Exp. 5: Error taxonomy for low-ROUGE predictions (ROUGE-1 < 0.3).

Error Type	Count	%
Format violation	0	0.0%
English bias	12	30.0%
Vocabulary mismatch	4	10.0%
Wrong tier	0	0.0%
Hallucination	24	60.0%
Total low-ROUGE	40	100%

6.6. Exp. 6 – Reward-Guided Best-of-N and Preference-Pair Harvesting

Exp. 6 asks: *what is the empirical ceiling of this system if we can select the best of several candidate answers?* No policy optimisation (DPO or PPO) is run in this experiment; the reward model is used solely for *reranking*, and the logged preference pairs are an asset for future fine-tuning (see Section 8.3). We generate N=5 candidates per question using a diverse pool of strategies (P7$\times$3 with different seeds, P5, and SIR), score them with a five-dimensional LLM reward model (max 14 pts), add a ROUGE-1 bonus (max 12 pts) for a combined score of 26, and select the highest-ROUGE-1 winner. The winner then undergoes a ROUGE-refinement rewrite pass to further improve lexical overlap. As a by-product, every (winner, loser) pair is logged as a DPO preference pair [5], yielding up to 688 preference instances per tier at zero additional annotation cost.

Best-of-N achieves R1=0.413/0.411 (Easy/Expert) — a 19%/20% relative improvement over the P7 single-strategy baseline — suggesting that candidate diversity and selection add substantial value. However, this comes at a cost: format compliance falls to 64–67% across Easy and Expert pools (vs 100% for all single-strategy runs), and the language-coverage reward dimension averages only 0.12/3.00 (Expert), revealing a fundamental *quality-ROUGE tension*: candidates that score highest on ROUGE-1 tend to be more verbose and English-centric, sacrificing multilingual grounding and output-schema compliance. Because candidate selection relies on ROUGE-1 against gold labels, Best-of-N is strictly an oracle method on development data and cannot be applied to the blind test set. This disqualifies it from our official portal submission and reinforces P7 as the best *deployable* strategy.

Tables 10, 11, and 12 report per-tier ROUGE scores and the breakdown of mean reward dimensions.

Table 10
Exp. 6: Reward-Guided Best-of-N (N=5 candidates, ROUGE-1 selection + refine pass) vs single-strategy runs (Easy). Reward max=26 (5-dim LLM judge/14 + ROUGE-1 bonus/12).

System	R1	R2	RL	Reward/26
P0 Baseline	0.2939	0.1146	0.2265	—
P5 Combined	0.3041	0.1235	0.2349	—
P7 ROUGE-Opt	0.3461	0.1525	0.2679	—
Structured IR	0.3018	0.1309	0.2399	—
Best-of-N (+Refine)	**0.4133**	**0.2008**	**0.3073**	9.97

7. Blind-Test Submission and Official Results

7.1. Submission Configuration

Following the protocol described in Section 8.4 below, we generated answers for all 128 Easy and 128 Expert blind-test questions (256 total) using P7 with deterministic decoding (temperature ≈ 0) and enforced the required `Answer:`/`Evidence:` schema in post-processing. The submission file

Table 11
Exp. 6: Reward-Guided Best-of-N (N=5 candidates, ROUGE-1 selection + refine pass) vs single-strategy runs (Expert). Reward max=26. Note: format compliance drops to 64–67% across tiers in Best-of-N vs 100% in single-strategy runs.

System	R1	R2	RL	Reward/26
P0 Baseline	0.3058	0.0995	0.1878	—
P5 Combined	0.2864	0.0856	0.1740	—
P7 ROUGE-Opt	0.3411	0.1330	0.2201	—
Structured IR	0.3253	0.1068	0.1971	—
Best-of-N (+Refine)	**0.4111**	**0.1733**	**0.2559**	9.46

Table 12
Exp. 6: Mean reward dimensions from the LLM reward model (Expert tier, $n = 172$). Low language coverage and format compliance scores reveal a quality-ROUGE tension.

Reward Dimension	Max	Mean
Factual accuracy	4	2.24
Groundedness	3	0.70
Language coverage	3	0.12
Format compliance	2	0.00
Conciseness	2	1.58
ROUGE-1 bonus	12	4.81
Total reward	26	9.46

`FinMMeval_T2_FTest_submission.jsonl` was exported and uploaded to the external evaluation portal.

7.2. Official Portal Evaluation

Table 13 shows the per-tier development scores for reference; Table 14 reports the official blind-test scores returned by the evaluation portal. The system achieved **Rank #1** on the Task 2 leaderboard with ROUGE-1 F1 = 0.3118 and full 100% coverage across all 256 blind-test instances (128 Easy + 128 Expert), which differ from the 344-instance development set (172 Easy + 172 Expert) used for all ablations above. The test aggregate (R1 F1 = 0.3118) is slightly below the development average (P7 avg R1 $\approx$ 0.344), consistent with the expected development-to-test distribution shift on unseen financial documents. The deterministic, inference-only nature of the P7 pipeline — no test-set tuning, no gold-label dependency — suggests that the generalisation holds without any blind-test-specific adaptation.

Table 13
Development-set ROUGE scores (P7, from Exp. 2) by tier. The portal reports aggregate scores only; tier breakdown is from internal evaluation.

Tier	ROUGE-1	ROUGE-2	ROUGE-L
Easy	0.346	0.153	0.268
Expert	0.341	0.133	0.220
Avg	0.344	0.143	0.244

Table 14
Official blind-test results from the evaluation portal (P7 submission, 256 instances: 128 Easy + 128 Expert). **Rank #1** on the Task 2 leaderboard as of 2026-05-14. Note: the blind-test set (256) differs from the development set used for all ablations (344: 172 Easy + 172 Expert).

System	ROUGE-1 F1	Precision	Recall	Coverage
Calibrated Signals (P7)	**0.3118**	0.3025	0.3764	100.0%

8. Evaluation and Improvement Opportunities

8.1. Key Findings

- **P7 beats P5 as the best single strategy.** Despite P5 being our designed primary submission, P7 (ROUGE-Opt) achieves R1=0.346/0.341 Easy/Expert versus P5's 0.304/0.286. Critically, P5 *underperforms P0 zero-shot* on Expert (-0.019 ΔR1), suggesting the combined role+CoT+tier-routing prompt adds noise for analytical questions.
- **LLM-as-Judge is the most effective inference-time technique.** Judge filtering raises Easy R1 from 0.2904 to 0.4096 (+0.119) — the single largest gain in any experiment.
- **Greek (EL) is the most informative non-English language on Expert.** Removing EL drops Expert R1 by 0.010, the largest single-language drop; removing English reduces Expert R1 by 0.045, indicating multilingual context is genuinely useful.
- **Best-of-N is the empirical ceiling.** Exp. 6 achieves R1=0.413/0.411 but at a cost: format compliance falls to 64–67% (vs 100% for all single-strategy runs), and language coverage reward averages only 0.12/3.00, revealing a quality-ROUGE tension.
- **Hallucination is the dominant failure mode** (60% of low-ROUGE cases), followed by English bias (30%) — both addressable by fine-tuning on the DPO pairs generated by Exp. 6.

8.2. Known Issues and Improvements

Table 15 summarises the issues identified and proposed fixes.

Table 15
Issues identified and proposed fixes.

Issue	Proposed Fix
P5 (Combined) underperforms P0 on Expert (R1=0.286 vs 0.306)	Decouple the three sub-rules; run ablation to identify which sub-rule hurts Expert
LLM-as-Judge scores are low (3.7/10 Easy, 3.1/10 Expert) despite high ROUGE	Calibrate judge threshold; judge may be penalising format compliance rather than factual quality
Best-of-N format compliance drops to 64–67%	Filter candidates by format compliance before ROUGE selection; add format penalty to reward
Language coverage reward near zero (0.02 Easy, 0.12 Expert)	Add a hard constraint to include at least one non-English citation in the reward model prompt
ROUGE refinement prompt may hallucinate	Constrain rewrite to swap vocabulary only; prohibit changing numeric values
No deduplication of DPO pairs when best strategy changes	Hash (question, chosen, rejected) tuples before appending

8.3. Remaining Limitations and Improvement Roadmap

Three concrete directions for future work follow directly from our results:

1. **Fine-tuning with DPO preference data.** Exp. 6 generated 172×4 = 688 (winner, loser) preference pairs per tier as a by-product at zero extra cost. Fine-tuning an open-weight multilingual model (e.g. Qwen2.5 [11] or mT5 [12]) on these pairs with DPO [5] or GRPO [13] should directly reduce hallucination and English bias — the two dominant failure modes.
2. **Multilingual retrieval augmentation.** The consistency check (Exp. 4) shows 41.3% of responses ignore non-English sources. Prepending a per-language retrieved sentence (dense retrieval over the news corpus) before each news section would encourage the model to anchor its answer in non-English text without relying solely on prompt instructions.
3. **Format-ROUGE joint optimisation.** Best-of-N (Exp. 6) raises R1 to 0.41 but format compliance drops to 64–67%. Adding a binary format-compliance filter before ROUGE selection (reject any candidate lacking both `Answer:` and `Evidence:` sections) would trade a small ROUGE reduction for fully compliant outputs, which is the requirement for official submission scoring.

8.4. Final Blind-Test Submission Protocol

To align with the external portal setup (questions only, hidden labels), we use the following fixed inference-only protocol:

1. Use **P7** as the system prompt (best deployable strategy in Exp. 2).
2. Generate one response per item with deterministic settings (temperature near 0).
3. Enforce required schema in post-processing: `Answer:` and `Evidence:`, with answer length $\leq$100 words.
4. If schema is violated, run one constrained retry with the same prompt.
5. Export outputs as task-compliant JSONL for portal submission.

This protocol is reproducible, low-cost, and free of any gold-answer dependency.

9. Conclusion

We presented an inference-only system for cross-lingual financial QA, evaluated across six experiments designed to progressively identify the strongest deployable strategy for the blind test portal.

Experiment progression and P7 selection. Exp. 1 established P0 zero-shot as the baseline (R1=0.294/0.308 Easy/Expert). Exp. 2 (prompting ablation over P0–P7) showed that adding role, CoT, and tier-routing instructions does not consistently help and that P5 (Combined) actually underperforms P0 on Expert (-0.019 ΔR1), while **P7 (ROUGE-Opt)** — which instructs the model to reuse question key-terms and exact numeric values — achieves the best *average* score across both tiers (R1=0.346/0.341 Easy/Expert; avg 0.344). P6 (Few-Shot) achieves a higher Easy R1 (0.350) but drops on Expert (0.297); its cross-tier average of 0.324 is below P7's 0.344, making P7 a robust single strategy. Exp. 3 suggests that P7's advantage holds across all per-language ablations and that non-English sources — Greek (EL) in particular — provide genuine signal on Expert queries.

Exps. 4–6 tested techniques that could potentially surpass P7 but each introduced a deployment obstacle. Exp. 4 showed that LLM-as-Judge filtering raises Easy R1 to 0.410 (+0.119 over P0) — the largest single-technique gain — but this relies on retry passes with gold-label-calibrated thresholds that add latency and cost. Exp. 5 (error analysis) identified hallucination (60%) and English bias (30%) as the two dominant failure modes, motivating the reward dimensions in Exp. 6. Exp. 6 (Reward-Guided Best-of-N oracle) reached R1=0.413/0.411 but requires reference ROUGE for candidate selection, which is unavailable on the blind test set, and incurs a 64–67% format-compliance drop — disqualifying it from submission.

Submission decision. Because all techniques in Exps. 4–6 that outperform P7 either depend on gold-answer signals or degrade schema compliance, **P7 is the strongest deployable configuration in our experiments**. We froze P7 with deterministic decoding and strict post-processing as our official portal submission and report the resulting blind-test scores in Section 7 (Table 14).

Limitations and future work. A key finding is the quality-ROUGE tension: Best-of-N maximises ROUGE but format compliance drops to 64–67%, and language coverage rewards average only 0.12/3.00 on Expert, indicating the agent still defaults to English-centric generation despite explicit multilingual prompting. The 688 DPO preference pairs generated by Exp. 6 (per tier) are a concrete asset for future fine-tuning with GRPO/PPO [13, 14] on open-weight multilingual LLMs such as Qwen2.5 [11] or mT5 [12], which should directly reduce hallucination and English bias — the two dominant failure modes. Real-time multilingual retrieval augmentation and a binary format-compliance filter in the reward-guided candidate selection loop are the two highest-priority engineering improvements.

Acknowledgements

We thank the research community for releasing open multilingual financial QA benchmarks and evaluation resources.

Declaration on Generative AI

During the preparation of this work, the authors used **Claude Sonnet 4.6** (Anthropic) and **ChatGPT/GPT-5.5** (OpenAI) in order to improve the grammar, writing style, and sentence flow of portions of the paper text (usage categories: *sentence polishing* and *paraphrasing*). After using this tool, the authors reviewed and edited all AI-assisted content as needed and take full responsibility for the publication's content.

References

[1] Z. Xie, Y. Dai, R. Elbadry, V. Jani, X. Peng, L. Qian, G. Georgiev, D. Dimitrov, F. Zhang, J. Huang, J. Geng, Y. Chen, Y. Yuan, H. Wu, Y. Wang, I. Koychev, V. Stoyanov, M. Song, Y. Chen, S. Liu, P. 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, 2026.

[2] Z. Xie, X. Peng, G. Georgiev, D. Dimitrov, R. Elbadry, F. Zhang, L. Qian, J. Huang, V. Jani, Y. Dai, J. Geng, Y. Chen, Y. Yuan, H. Wu, Y. Wang, I. Koychev, V. Stoyanov, M. Song, Y. Chen, S. Liu, P. Nakov, Overview of the FinMMEval 2026 task 2: Financial question answering and summarization, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[3] Z. Chen, W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.-H. Huang, B. R. Routledge, W. Y. Wang, FinQA: A dataset of numerical reasoning over financial data, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 3697–3711. doi:10.18653/v1/2021.emnlp-main.300.

[4] F. Zhu, W. Lei, Y. Huang, C. Wang, S. Zhang, J. Lv, F. Feng, T.-S. Chua, 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 (ACL-IJCNLP), Association for Computational Linguistics, Online, 2021, pp. 3277–3287. doi:10.18653/v1/2021.acl-long.254.

[5] R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, C. Finn, Direct preference optimization: Your language model is secretly a reward model, in: Advances in Neural Information Processing Systems (NeurIPS), volume 36, Curran Associates, Inc., 2023.

[6] X. Peng, L. Qian, Y. Wang, R. Xiang, Y. He, Y. Ren, M. Jiang, V. J. Zhang, Y. Guo, J. Zhao, H. He, Y. Han, Y. Feng, Y. Jiang, Y. Cao, H. Li, Y. Yu, X. Wang, P. Gao, S. Lin, K. Wang, S. Yang, Y. Zhao, Z. Liu, P. Lu, J. Huang, S. Wang, T. Papadopoulos, P. Giannouris, E. Soufleri, N. Chen, Z. Deng, H. Fu, Y. Zhao, M. Lin, M. Qiu, K. E. Smith, A. Cohan, X.-Y. Liu, J. Huang, G. Xiong, A. Lopez-Lira, X. Chen, J. Tsujii, J.-Y. Nie, S. Ananiadou, Q. Xie, MultiFinBen: Benchmarking large language models for multilingual and multimodal financial application, 2025. URL: https://arxiv.org/abs/2506.14028. doi:10.48550/arXiv.2506.14028. arXiv:2506.14028.

[7] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Advances in Neural Information Processing Systems (NeurIPS), volume 35, Curran Associates, Inc., 2022, pp. 24824–24837.

[8] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Advances in Neural Information Processing Systems (NeurIPS), volume 33, Curran Associates, Inc., 2020, pp. 1877–1901.

[9] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, P. Clark, Self-Refine: Iterative refinement with self-feedback, in: Advances in Neural Information Processing Systems (NeurIPS), volume 36, Curran Associates, Inc., 2023.

[10] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging LLM-as-a-judge with MT-Bench and chatbot arena, in: Advances in Neural Information Processing Systems (NeurIPS), volume 36, Curran Associates, Inc., 2023.

[11] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Guo, Q. Team, Qwen2 technical report, arXiv preprint arXiv:2407.10671 (2024).

[12] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A massively multilingual pre-trained text-to-text transformer, arXiv preprint arXiv:2010.11934 (2021). doi:10.18653/v1/2021.naacl-main.41, published at NAACL 2021.

[13] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu, D. Guo, DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, arXiv preprint arXiv:2402.03300 (2024).

[14] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, in: arXiv preprint arXiv:1707.06347, 2017.

A. P7 System Prompt (ROUGE-Optimised)

The complete system prompt used for strategy P7 (ROUGE-Optimised), referenced in Table 2 and selected as the best deployable strategy in Exp. 2.

Listing 1: P7 (ROUGE-Optimised) system prompt.

```
You are a financial analyst. Your answers will be scored against
reference answers using ROUGE-1 unigram overlap.
```

```
RULES to maximise score:
1. Re-use KEY TERMS and EXACT PHRASES from the question.
2. Include SPECIFIC NUMBERS with exact units
   ($35.0 billion, 15 percent, Q3 FY2020).
3. State COMPANY NAME, METRIC, and TIME PERIOD explicitly.
4. Keep Answer section to 60 words or fewer -- high precision
   beats recall.
5. Do NOT use filler phrases. Start the answer directly.
6. Consult ALL five language sources (EN, ZH, JA, ES, EL).
7. Follow the exact Answer/Evidence format in the query.
```

B. LLM-as-Judge Prompts (Exp. 4)

The following prompts are used for the Judge Filtered technique in Exp. 4 (Table 3). A retry is triggered when the returned `total` field is below the threshold of 6/10.

System prompt

```
You are an expert evaluator of financial question answering
systems. You will evaluate answers on four dimensions.
```

User prompt template (`{question}`, `{reference}`, `{generated}` are filled at runtime)

Listing 2: LLM-as-Judge user prompt template (Exp. 4).

```
Evaluate the following financial QA answer.

QUESTION: {question}
REFERENCE ANSWER: {reference}
GENERATED ANSWER: {generated}

Score on these four dimensions:

1. GROUNDEDNESS (0-3):
   0 = Contains hallucinated facts not in context
   1 = Mostly grounded, minor unsupported claims
   2 = Fully grounded in provided documents
   3 = Fully grounded with explicit source attribution

2. LANGUAGE COVERAGE (0-3):
   0 = Only English sources used
   1 = Mostly English, minimal non-English
   2 = At least 2 non-English sources consulted
   3 = All 5 language sources clearly consulted

3. FORMAT COMPLIANCE (0-2):
   0 = Missing required Answer/Evidence sections
   1 = Partially correct format
   2 = Perfect format compliance

4. CONCISENESS (0-2):
   0 = Over 100 words in answer section
   1 = Under 100 words but verbose
   2 = Concise and complete under 100 words
```

```
Respond in this exact JSON format:
{
  "groundedness": <0-3>,
  "language_coverage": <0-3>,
  "format_compliance": <0-2>,
  "conciseness": <0-2>,
  "total": <0-10>,
  "reasoning": "<one sentence explanation>"
}
```