

LangKG at the FinEval 2026: Answer-Contract Knowledge over Instruction Engineering

Shivika Prasanna, Jaeseong Lee, Hui Su, Yourong Xu

{shivika.prasanna, jaeseong.lee, hui.su, yourong.xu}@fmr.com

Abstract

Automated multilingual financial question answering requires models to synthesize evidence across SEC filing statements and parallel news reporting in five languages, grounding every claim in aligned, verifiable source material. We participate in the FinNLP 2026 Shared Task (FinEval 2026) Subtask 3, built on the challenging MultiFinBen/PolyFiQA benchmark, under a strict 70-billion-parameter constraint and a Macro-averaged ROUGE-1 F1 evaluation metric. We find that shared task’s practice (train) and test splits encode an inverted query-answer contract, and highly specific prompt engineering tuned against one split overfits and fails to transfer to the other. Motivated by this, we introduce an explicit answer-contract knowledge policy encoding evidence-boundedness, multilingual evidence alignment, exact-value preservation, reporting-period disambiguation, and intentional non-response, and implement it through two complementary generation strategies: Chain-of-Verification (CoVe) and Prompt-variant ensembling with heuristic re-ranking. Our submitted system (ensembling) improves over direct-prompting baseline and CoVe (0.41 to 0.49 ROUGE-1 F1 over train set) and achieves 0.324 ROUGE-1 F1 on the shared task’s leaderboard, ranking 11 of 15 teams.

1 Introduction

Automated financial QA increasingly requires models to aggregate information across multiple sources, and often times across multi-lingual sources. The FinNLP 2026 Shared Task (FinEval 2026) Subtask 3: Multilingual Financial Q&A formalizes this challenge by evaluating Q/A across SEC core filing statements and parallel news articles published in English (EN), Chinese (ZH), Japanese (JA), Spanish (ES), and Greek (EL) using the MultiFinBen/PolyFiQA (Peng et al., 2026) benchmark. Participating models, constrained strictly to under 70 billion parameters, are evaluated on their ability to extract verified figures,

synthesize multi-document findings, and ground claims in aligned evidence under Macro-averaged ROUGE-1 F1. However, standard prompting approaches suffer from critical failure modes in this benchmark, namely, conversational hedging, temporal conflation between quarterly and cumulative figures, and subtle stylistic divergences. To address these challenges, we present our approaches for this shared task as explicit answer-contract enforcement rather than brute-force prompting. We introduce a policy framework that enforces domain constraints across both generative reasoning via Chain-of-Verification (CoVe) and deterministic selection via prompt-variant ensembling with heuristic re-ranking, bridging the gap between single-pass model generation and the strict precision required by the shared task.

1.1 Dataset

PolyFiQA, introduced in the MultiFinBen benchmark, has two splits: *PolyFiQA-Easy*, *PolyFiQA-Expert*. However, as a part of the shared task, the following were shared with participating teams:

- Train: a 76-row set of query (prompt query, financial statements evidences in EN, ZH, JA, ES, EL, question and gold answers.
- Test: a 76-row set of query and question without gold answers.

The relation between PolyFiQA-Easy and PolyFiQA-Expert and the shared task’s train and train sets cannot be established as no gold answers were released to confirm this mapping. We therefore treat PolyFiQA-Expert as the validation dataset on the basis that it is the more challenging of the two upstream splits. From the experiments conducted to evaluate the performance of our approaches on train and from inspection of query phrasing on the test data (gold answers were withheld for test), we observed the following:

- **Train:** Queries in this set require answer to be composed as an ‘answer’ using financial statements and verified using financial news, both provided in the context.
- **Test:** Queries in this set require answer to be a ‘summary’ of the financial news, using financial statements as evidence. This has been inferred from the query structure alone as gold answers were unavailable to confirm the effect on the test set. Validation set follows the same structure.

These findings challenge conventional prompt engineering practices where tailoring prompts with highly specific instructions frequently induced overfitting on the validation set.

2 Methodology

2.1 Baseline

As baseline, we utilize the default prompts provided in training dataset. Each instance contains a target question, ground-truth answer with evidence, context (financial statements, multilingual news and question) with instance-specific prompt. We feed this directly into the LLM without modification.

2.2 Knowledge-based Generation

From the baseline experiments (Section 2.1), we observed gaps between the dataset’s query, gold answer and the answers generated by our approaches. We identified a critical need to bridge these gaps by adding knowledge, or answering policies for multilingual financial question answering over financial statements and financial news. This knowledge encodes how to interpret the dataset’s query-answer contract, how to align evidence to that contract, and how to constrain model outputs so that they match the benchmark’s latent annotation logic. This knowledge has the following components:

- **Evidence-bounded Answering:** Answer must be derived exclusively from provided financial statements and multilingual news. No unsupported information, background financial knowledge or guesses are allowed. This also includes inventing values, periods, causal explanations, derived evidence snippets.
- **Multilingual Evidence Alignment:** Multilingual news sections in the query should be interpreted as independent, parallel renderings of shared facts rather than unrelated or conflicting articles. Therefore, recurring shared facts across EN, ZH, JA, ES and EL would be considered as stronger evidences.
- **Financial Exact Value Handling:** All values reported in the multilingual news evidences must always be preserved in the same form that they appear in. This includes units, currencies, language-based decimal conventions and reporting-period labels.
- **Strict Interpretation of Reporting Periods:** When a question refers to the current figure, the answer should use the relevant current reporting column rather than mixing quarterly and cumulative values.
- **Intentional Non-Response:** If the requested metric is not explicitly available or cannot be computed directly from the evidences, the correct output should be "None". From the dataset, the QA task does not necessarily reward hedged explanations or alternatives like "not provided" or "undetermined".

This knowledge “policy” is important to be considered, as the metric (ROUGE-1 F-1) required for this shared task rewards a particular combination of evidence selections, temporal descriptions, alignment, formatting and exclusion behaviors, and does not simply evaluate the financial correctness of the generated answers. We integrate this knowledge into the following approaches - Chain-of-Verification (2.3) and Prompt-variant ensembling with heuristic re-ranking and generation (2.4). We emphasize that these components constitute a general answering policy rather than fixed bundle applied identically across all methods; individual generation methods incorporate different subsets of this policy depending on their underlying mechanism.

2.3 Chain-of-Verification (CoVe)

Building upon the baseline, we employ the Chain-of-Verification (CoVe) framework to systematically mitigate factual hallucinations in complex financial reasoning (Dhuliawala et al., 2024). Unlike the direct generation approach, CoVe decomposes the reasoning process into the following four explicit stages: **1. Draft Generation.** The LLM generates an initial draft response to the target question. **2. Verification Planning.** It identifies key factual claims within the draft and formulates targeted

verification questions. **3. Verification Execution.** The model independently answers these verification queries by strictly grounding its reasoning in the provided financial statements and multilingual news. **4. Final Response Synthesis.** Finally, the LLM cross-references the initial draft against the verified facts, revising any inconsistencies to synthesize a highly accurate, refined final response.

2.4 Prompt-variant Ensembling with Heuristic Re-ranking

To improve robustness against variance in single-pass generation, we employ a prompt-variant ensembling strategy that samples multiple stylistically distinct completions from the same underlying model and selects among them with a deterministic scoring function, rather than ensembling across different models. For each query, up to five candidate responses are generated from the base system prompt augmented with a short style-override suffix: (a) Base, an unmodified completion; (b) Concise, instructed to keep the answer to 1–2 sentences, avoid derivation language for "final result(s)" questions, and include one short aligned quote when the question requests news evidence; (c) Extractive, instructed to reuse wording and numeric expressions from the provided context rather than paraphrasing; (d) Final-result-only, instructed to return only the requested result(s) and omit derivation steps; and (e) Quote-priority, instructed to prioritize a short aligned quote in the evidence field when the question requests one. Each candidate is generated independently, with automatic retry on transient provider errors. A deterministic candidate scorer then selects among the available candidates without requiring access to the gold answer. The candidate’s score is incremented for the presence of both required output sections, decremented if the question specifies “only the final result(s)” but the candidate’s answer nonetheless contains derivation markers, incremented if the question requests a quote and the candidate supplies non-null evidence, decremented otherwise, and adjusted for length. The highest-scoring candidate is then selected as the final response.

3 Results

3.1 Experimental Setup

Large Language Models. We evaluate baseline and CoVe approaches using Llama 3.3-70B-Instruct (Grattafiori et al., 2024) with temperature

set to 0.3 for baseline and 0.5 for CoVe for draft generation to encourage sufficient information retrieval and 0.1 for both verification planning and execution to ensure the deterministic extraction of factual claims. The verified answer generation stage employs a temperature of 0.3 to balance factual fidelity with linguistic fluency. The Prompt-variance Ensemble method with Heuristic Reranking has been evaluated using Qwen3-32B (Yang et al., 2025) with temperature at 0.0, top p at 0.70, presence and frequency penalties at 0.0. For the prompt-variance ensemble method, other models such as Llama 3.3-70B-Instruct, Gemma4 (Team et al., 2026), Qwen3-8b (Yang et al., 2025), Qwen2.5-14b-Instruct (Qwen et al., 2025) and Gemma3-27b (Team et al., 2025) were also explored but did not reach comparable performance on both train and validation set as compared to the selected model. All the models are under 70 billion parameters, as required by the shared task.

Prompts. The complete set of prompts utilized in our methods is provided in the Appendix A.

Evaluation Metrics. For the shared task, we require to evaluate the performance of our models on the test set using Macro-averaged ROUGE-1 F1 metric (Lin, 2004). In addition to this, we also report BERTScore F1 computed using the pre-trained bert-base-multilingual-cased model (Devlin et al., 2019), for semantic alignment.

Evaluation Process. We evaluate our methods and model selection on the train and validation sets. For the train and validation set, we locally compute both the macro-averaged ROUGE-1 F1 score and BERTScore F1. To evaluate our approach on the test set, we submitted our generated responses using Prompt-variant Ensembling as this had the highest scores on the train set, to compute the ROUGE-1 F1, recall, and precision scores for the public leaderboard. Furthermore, to ensure reproducibility and obtain the final evaluation metrics, we submitted our executable solution code to the platform for independent verification. The scores yielded from this verified code are reported on the private leaderboard.

3.2 Evaluation Results

We present our evaluation results across the train and validation set in this section. On the private leaderboard for this shared task, we rank 11 out of the 15 participating teams, with a macro-averaged ROUGE-1 F1 score of 0.324 (a gap of 0.06 against the highest ranking team on the leaderboard) on

Split	Method	R-1 F1	R-1 Prec.	R-1 Rec.	BERTScore F1
Train	Baseline	0.41	0.54	0.41	0.78
	CoVe	0.48	0.52	0.49	0.85
	Ensemble	0.49	0.53	0.47	0.86
Valid	Baseline	0.25	0.23	0.42	0.81
	CoVe	0.30	0.28	0.41	0.82
	Ensemble	0.36	0.35	0.40	0.87
Test	Ensemble	0.3240	0.3042	0.3897	N/A

Table 1: Model performance across Train and Validation sets. Test set scores were computed by the shared task organizers based on the submitted code and are available on the leaderboard. Shared task required only Macro-averaged ROUGE-1 F1.

the test set. In addition to the ROUGE-1 F1 score, we also computed BERTScore F1 on train and validation sets to assess the semantic value of the generated answers. The shared task only required to compute the macro-averaged ROUGE-1 F1 score.

3.3 Discussion

Our submitted system, integrating answer-contract knowledge policy (2.2) with prompt-variant ensembling (2.4), achieved a leaderboard score of 0.32 ROUGE-1 F1 on the test set, ranking 11th of 15 participating teams; the top-ranked submission scored 0.38. Relative to our own baseline (0.25 on the validation set) and Chain-of-Verification approach (0.30), this represents a consistent improvement attributable to our answer-contract framing, suggesting that explicitly encoding the benchmark’s latent query-answer conventions is a more productive lever than iterating on generic instruction phrasing. This is consistent with our findings in Section 1.1 that highly specific ad-hoc instructions tended to overfit the validation set rather than generalize to the test distribution. Additionally, we view our result as evidence that lightweight, policy-driven approach is competitive on this challenging benchmark without requiring large-scale search over model or prompt configurations.

3.4 Limitations

Our evaluation is conducted on a small dataset (76 test examples, split into a 24-example public and 52-example private partition); at this scale, individual-example variance can meaningfully affect both absolute scores and relative rankings, and we caution against over-interpreting small differences between submissions, including the gap to the top-ranked team. The shared task’s evaluation metric, macro-averaged ROUGE-1 F1, rewards lexical overlap with a single human-authored reference rather than financial correctness directly. Fac-

tually accurate answers phrased differently from the reference receive no credit under this metric. Finally, due to compute and time constraints, our model selection reflects a practical comparison across a limited set of candidate models under our specific prompting setup rather than an exhaustive search, and we cannot rule out that a different base model or more extensive tuning would narrow the observed gap.

References

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 3563–3578.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Xueqing Peng, Lingfei Qian, Yan Wang, Ruoyu Xiang, Yueru He, Yang Ren, Mingyang Jiang, Vincent Jim Zhang, Yuqing Guo, Jeff 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)*, pages 16934–16963.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. 2026. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin,

Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Appendix

A.1 Baseline

The default instance specific prompt in Section 2.1 is as follows:

You are tasked with answering the user’s question using the provided context, which includes financial statements (Income Statements, Balance Sheets, and Cash Flow Statements) and financial news articles in multiple languages (English, Chinese, Japanese, Spanish, and Greek). Please provide a detailed and well-supported answer based on the information available. Limit your response to 100 words or fewer. If you cannot answer the question or if no relevant evidence is found, write “None”.

Answer Format:

Answer: {Answer using the financial statement.}

News Evidence: {Verify your answer using quote(s) from the financial news. Write “None” if no news evidence is available.}

Context:

Financial Statements: {FINANCIAL STATEMENTS AND MULTILINGUAL NEWS}

Question:

{QUESTION} Please quote the piece(s) of news aligned with the finding(s), if any.

A.2 CoVe

For the CoVe framework detailed in Section 2.3, the specific prompt templates utilized for each stage and system prompt over all stages are as follows:

Draft Generation

{QUERY}

Generate a draft answer.

Follow the required Answer / News Evidence format.

Verification Planning

{QUERY}

Draft answer:

{DRAFT ANSWER}

Create a concise verification plan. List the facts, numbers, reporting periods, calculations, and news evidence that must be checked against the provided context before finalizing the answer.

Do not answer the original question yet.

Verification Execution

{QUERY}

Draft answer:

{DRAFT ANSWER}

Verification plan:

{VERIFICATION PLAN}

Execute the verification plan using only the provided context. For each check, state whether the draft answer is supported, unsupported, or needs correction. For ratio questions, verify the numerator and denominator values from the same reporting period. Do not produce the final answer yet.

Final Response Synthesis

{QUERY}

Draft answer:

{DRAFT ANSWER}

Verification plan:

{VERIFICATION PLAN}

Verification results:

{VERIFICATION RESULTS}

Produce the final verified answer. Correct any unsupported or wrong parts from the draft. Include the final result and a brief reason using verified source values. For ratio questions, state the numerator and denominator values used. Use only supported financial-statement values and directly relevant news evidence. Follow the required format exactly: Answer: ... News Evidence: ...

System Prompt

You are a precise financial question-answering assistant.

Terminology:

- Use the terminology from the question and financial evidence consistently.
- Distinguish carefully between revenue, sales, income, profit, assets, liabilities, equity, cash inflows, and cash outflows.
- Do not confuse financial-statement evidence with news evidence.

Numbers, currencies, and dates:

- *Preserve every reported number exactly as it appears in the evidence, including decimal points, commas, percentage signs, and units.*
- *Preserve currency symbols exactly as they appear in the evidence.*
- *Preserve dates and reporting periods exactly as they appear in the evidence.*
- *Do not silently round, convert currencies or units, or change a decimal into a percentage. If a calculation is required, show the result using the requested format and retain the source values exactly.*
- *Use only information supported by the provided evidence. Do not invent or infer unsupported values.*

Answer language and evidence:

- *Write the explanatory answer entirely in English.*
- *Evidence quotes may remain in their original language. Do not translate, rewrite, or alter quoted evidence; reproduce each selected quote exactly.*
- *Use news evidence only when it refers to the same company, metric, and reporting period as the question. If no relevant news evidence is present, write "None".*

Answer requirements:

- *Follow the requested answer format exactly.*
- *Keep the complete response within 100 words unless the question explicitly requests a different length.*
- *Answer every part of the question. Include all requested values, comparisons, periods, and relevant evidence.*
- *For R&D ratio questions, output only the requested period and requested comparison period. Do not output all available periods.*
- *For cash-flow questions, distinguish cash inflow from cash outflow. A more negative value means a larger cash outflow.*
- *For revenue questions, include the exact revenue values, comparison period, percentage change, and segment drivers if they are present in the financial statements.*
- *Explain the answer in the Answer section. When the answer is "Yes", state which changes occurred, their direction, and the relevant values. When the answer is "No", explain why when the evidence supports an explanation. If the answer is "None",*

Criterion	Score Adjustment
Contains Answer: section	+2
Contains evidence-field section	+2
Question says "only final result(s)" but Answer contains derivation markers (/, "from", "to")	-2
Question requests a quote and evidence is non-null	+3
Question requests a quote but evidence is "None"	-1
Answer length $\leq$ 220 characters	+1
Answer length $>$ 350 characters	-2
Detected provider/API error text	-10

Table 2: Candidate Scoring Rubric

do not add an explanation.

- *Do not label financial-statement values as news evidence.*

A.3 Prompt-variant Ensembling with Heuristic Re-ranking

For the ensembling re-ranking detailed in Section 2.4, each candidate reuses baseline system prompt (A.1) with a short style-override suffix appended.

Five candidates. The five candidate variants for style override are as follows:

- *Base: No suffix — the unmodified baseline prompt is used as-is.*
- *Concise: Keep Answer very short (1-2 sentences). For 'final result(s)' questions, avoid derivation language and extra comparisons. If the question asks for quoted news evidence, include one short aligned quote when available.*
- *Extractive: Be extractive. Reuse wording and numeric expressions from the provided context. Avoid paraphrasing and avoid adding extra analysis not explicitly requested.*
- *Final-result Only: Return only the final result(s) requested by the question. Keep Answer concise and avoid derivation steps.*
- *Quote-priority: If the question asks for news quote(s), provide one short aligned quote in News Evidence. If no aligned quote is available, return News Evidence: None.*

Candidate Scoring Rubric. The candidate scoring rubric used for the heuristic re-ranking is in Table 2. Scorer ranks all candidates and the single top-scoring one is selected as the final output, the rest are discarded.