

Mircea Andrei Razvan at the FinEval 2026: A Unified Prompting Approach for Multilingual Financial Q&A

Mircea Andrei Razvan
University of Bucharest
mircea.andrei.razvan@gmail.com

Abstract

We describe our submission to Subtask 3 of the FinNLP 2026 Shared Task, based on MultiFinBen/PolyFiQA (Peng et al., 2026), in which models answer cross-document financial questions using SEC financial-report statements and contemporaneous multilingual news in English, Chinese, Japanese, Spanish, and Greek, evaluated using ROUGE-1. The task’s four fixed question templates require an answer synthesized from the news, supported by evidence cited from the financial statements. We use Qwen3-VL-30B-A3B-Instruct, an open-weight mixture-of-experts model accessed via OpenRouter, with template-aware prompting for each question type. In early development, a multi-stage pipeline separating answer generation from evidence retrieval produced a high rate of empty evidence fields. A single unified generation call substantially reduced this rate. We additionally find that several formatting conventions, including specific numeric notation, compact period references, and a fixed field label, were not reliably followed through prompt instructions alone, and were instead enforced with rule-based post-processing. Our final system achieves a ROUGE-1 F1 of 36.53% on the official test set. These results suggest that deterministic post-processing remains a practical complement to prompt-based instruction for enforcing output conventions that models do not reliably learn from context alone.

1 Introduction

The task requires systems to answer financial questions by reconciling two distinct sources: contemporaneous multilingual news (in English, Chinese, Japanese, Spanish, and Greek), from which an answer is synthesized, and SEC financial-report statements, from which supporting evidence is cited. The task is instantiated through four fixed question templates covering revenue focus, capital allocation, profit margin strategy, and capital expenditure significance, and is scored with ROUGE-1 against

expert-written reference answers. Unlike tasks with a single computable correct answer, this task admits multiple valid syntheses of the same news coverage, which has direct consequences for both system design and the ceiling on achievable performance.

We approach the task with a single unified generation call, producing both the answer and its supporting evidence in one pass rather than decomposing the task into separate stages, using Qwen3-VL-30B-A3B-Instruct, an open-weight mixture-of-experts model accessed via OpenRouter, guided by template-aware prompting for each of the four question types. During development, we found that decomposing the task across separate generation stages substantially increased the rate of missing evidence, and that several output formatting conventions were not consistently produced through prompting alone, requiring deterministic post-processing instead.

Our final system achieves a ROUGE-1 F1 of 36.53% on the official test set. The remainder of this paper is organized as follows: Section 2 describes the task and data in more detail; Section 3 presents our system design; Section 4 reports results and analysis; and Section 5 discusses conclusions and limitations of our approach.

2 Task and Data

2.1 Task Description

Each instance specifies a company and reporting period, together with the corresponding financial statements and multilingual news coverage. Systems must produce two fields: an Answer, synthesized from the news coverage, and Financial Statements Evidence, citing supporting figures from the financial statements. One of four fixed templates – revenue focus, capital allocation, profit margin strategy, or capital expenditure significance – determines the specific question asked for a given instance. Combined output across both fields is

limited to 100 words.

2.2 Data

The official test set comprises 76 instances spanning four companies (Microsoft, Honeywell, Johnson & Johnson, and Universal Corporation), with each company represented by multiple reporting periods and all four question templates applied per period-company pair.

We additionally consulted a separate, publicly available resource associated with this benchmark family, providing practice instances for an inverse-format version of the task: an *Answer* computed from the financial statements, supported by a *News Evidence* field quoting corroborating news coverage. This resource shares no question or answer content with the official test set. We found its reference answers informative for identifying general formatting conventions – including numeric notation and period referencing – applicable to this benchmark family.

2.3 Evaluation

Submissions are scored using ROUGE-1 F1 against expert-written reference answers. As a measure of unigram lexical overlap, ROUGE-1 rewards surface-level agreement with the reference text and does not directly credit semantically equivalent phrasing that diverges lexically from the reference. We return to this property of the metric in our analysis of error sources (Section 4).

3 System Description

3.1 Model

We use Qwen3-VL-30B-A3B-Instruct, an open-weight mixture-of-experts model, accessed via OpenRouter as required by the shared task’s API-based workflow track. The model has 30.5 billion total parameters, with approximately 3.3 billion activated per token. Following the task’s eligibility criterion, which is based on total parameter count rather than the number of parameters active during inference, the model satisfies the 70-billion-parameter limit with substantial margin.

3.2 Model Selection

Prior to finalizing our system, we evaluated several open-weight candidate models within the task’s 70-billion-parameter budget, including variants of Llama, Gemma, and Command-R, in addition to Qwen3-VL-30B-A3B-Instruct. We observed that

Llama and Gemma models more frequently produced answers that lost track of relevant context across the multilingual news input, particularly when synthesizing across several language sections simultaneously, whereas Qwen3-VL-30B-A3B-Instruct produced more consistent results across repeated trials.

To quantify this comparison, we evaluated Llama-3.3-70B-Instruct and Gemma-3-27B-IT against Qwen3-VL-30B-A3B-Instruct under identical conditions (unified generation call, template-aware prompting, and post-processing enabled), scoring repeated runs against the overlapping resource described in the Limitations section. Qwen3-VL-30B-A3B-Instruct achieved a mean ROUGE-1 F1 of 36.62% (std 0.70%), compared to 33.23% (std 0.55%) for Gemma-3-27B-IT and 27.70% (std 0.50%) for Llama-3.3-70B-Instruct. These gaps substantially exceed the run-to-run variance observed for any individual model, supporting our qualitative observation above.

We selected Qwen3-VL-30B-A3B-Instruct for our final submission on this basis.

3.3 Generation Architecture

Our first architecture used two separate generation calls: one to synthesize the answer from the multilingual news, and a second, independent call to retrieve supporting evidence from the financial statements. This design proved unreliable: across repeated development runs, the evidence field was left empty in 36–41 of 76 instances, including cases where the answer field itself cited figures that plausibly originated from the financial statements – indicating that the relevant evidence existed but was not being reliably surfaced when retrieval was isolated from the answer-generation context.

We subsequently adopted a single generation call producing both fields together, which reduced the rate of empty evidence fields to 1–4 of 76 instances across repeated runs. A likely explanation is that the isolated evidence call lost access to the synthesized answer at the point evidence was generated, while the unified call retains this context throughout.

3.4 Template-Aware Prompting

The four question templates are distinguished by fixed, non-overlapping phrases present in the question text itself (e.g., “*top three focuses on revenue*”, “*allocating capital*”), allowing deterministic template identification without additional in-

ference. For each identified template, we supply a short template-specific instruction clarifying the intended scope of the question – for instance, specifying that the capital allocation template concerns share repurchases, dividends, and investment decisions specifically, rather than general revenue or growth commentary. We found this necessary because the model otherwise tended to default to the most prominent narrative in the news coverage, regardless of the specific facet asked about.

To quantify the contribution of template-aware prompting, we compared repeated runs of our final system with and without these template-specific instructions, holding post-processing constant and scoring against the overlapping resource described in the Limitations section. Removing template-aware prompting reduced mean ROUGE-1 F1 from 36.62% to 33.93% (std 0.74%, range 33.10–34.76%), a larger drop than that observed when disabling post-processing alone (Section 3.5). This suggests that keeping the model anchored to the specific facet asked about – rather than defaulting to the most prominent narrative in the news coverage, as discussed in Section 4.4 – contributes at least as much to overall performance as the deterministic formatting corrections applied downstream.

3.5 Output Normalization

We found that several output constraints were not reliably respected through prompt instructions alone. Most notably, the task’s 100-word combined length limit was occasionally exceeded despite explicit instruction, requiring a deterministic, sentence-boundary-aware trimming procedure applied after generation: evidence content is shortened first, followed by the answer content only if further reduction is required. We additionally applied light deterministic normalization to remove markdown artifacts, and to collapse cases where the model explained, in a full sentence, that no relevant information was found into the concise negative-response format specified by the task. We discuss further formatting-related observations, and one specific limitation of our normalization procedure, in Section 5.

To quantify the impact of this post-processing stage, we compared repeated runs of our final system with and without these rules applied, scored against the overlapping resource described in the Limitations section. Disabling post-processing reduced mean ROUGE-1 F1 from 36.62% to 34.46% (std 0.55%, range 34.02–35.25%), indicating that

deterministic output normalization accounts for a substantial share of our system’s measured performance, independent of run-to-run stochastic variance.

4 Results and Analysis

4.1 Official Results

Our final submission achieves a ROUGE-1 F1 of 36.53% (precision 35.17%, recall 41.94%) on the official private test set, ranking 5th among 15 submissions.

4.2 Score Trajectory Across Development

Because our system relies on stochastic generation from a large language model, we observed meaningful variation in scores across successive runs, which we take as an approximate empirical bound on the measurement noise inherent to a 76-instance test set scored via a single stochastic generation pass per instance. During the submission phase, the official platform reported scores against a public subset comprising 25% of the test set; we selected our final submission based on its performance on this visible portion, following standard practice for leaderboard-based shared tasks with public/private splits. Our selected submission subsequently scored 36.53% on the full, private test set once the competition concluded. Separately, after the submission window had closed, we characterized run-to-run variance by repeatedly running our final system against the overlapping resource described in the Limitations section, which permits local ROUGE-1 scoring against gold labels matching the official test set; across these runs, ROUGE-1 F1 ranged from 35.59% to 37.45% (mean 36.62%, std 0.70%). This range, measured independently of the selection process above, supports interpreting the official 36.53% as one plausible draw from the system’s typical performance rather than a precise, reproducible point estimate. Larger, more consistent gains were observed following changes to template-aware prompting and output formatting, quantified in Sections 3.4 and 3.5.

4.3 Output Characteristics

To characterize our system’s behavior independently of reference comparison, we report descriptive statistics over our own submitted predictions. Across the 76 test instances, the *Answer* field averaged approximately 44 words and the *Financial Statement Evidence* field averaged approximately

24 words, both within the task’s combined 100-word limit in all cases. The evidence field was left as *None* in a small minority of instances. All four question templates were represented evenly across the test set, consistent with the fixed template structure described in Section 2.

4.4 Sources of Error

We supplement our official results with a manual review of our own outputs, which revealed two recurring issues. In the first, the generated answer addressed the news coverage in general terms – overall company performance, for instance – rather than the specific aspect asked about in the question template, such as capital allocation, despite template-specific instructions intended to prevent this. In the second, the model occasionally produced an explanatory sentence stating that no relevant information had been found, rather than the concise negative response the task format requires. We addressed a subset of these cases through deterministic post-processing.

The first pattern proved more difficult to resolve. When a more prominent, unrelated topic appeared alongside the specific one asked about, prompting alone was not enough to keep the model focused on the narrower topic, and we were unable to fully resolve this within the scope of this work.

5 Conclusion

We presented our submission to Subtask 3 of the FinNLP 2026 Shared Task, using a single unified generation call with template-aware prompting and deterministic output normalization. Our system achieves a ROUGE-1 F1 of 36.53% on the official private test set, ranking 5th among 15 submissions. Consolidating answer and evidence generation into a single call substantially reduced missing-evidence errors relative to a multi-stage design, and several output formatting conventions required deterministic enforcement rather than reliance on prompt instructions alone. These observations point to two open questions for evidence-grounded QA systems more broadly: how to keep a model focused on the specific sub-topic a question asks about when a broader, more prominent topic is present in the same input, and how much output formatting should be left to the model to get right versus enforced afterward in code.

Limitations

Our final system is built around a single model (Qwen3-VL-30B-A3B-Instruct), selected based on the comparison described in Section 3.2, including a controlled ablation against Gemma-3-27B-IT and Llama-3.3-70B-Instruct under matched conditions.

ROUGE-1, as a measure of lexical overlap, does not fully capture semantic correctness. This task in particular admits genuine variation in which facts a valid answer might surface from a given set of news coverage, meaning that even a system producing an accurate and well-supported answer may not achieve high lexical overlap with a single reference answer reflecting one particular editorial choice among several plausible ones.

In addition to the resource described in Section 2, we consulted a second, separate resource during development, in the same output format as the official task, which we understood at the time to be a distinct, non-overlapping practice set. This resource was later found, after our code submission deadline had passed, to overlap substantially with the official test set, sharing matching task and question identifiers. One specific output-formatting rule in our submitted code – concerning field-label wording – was set based on this resource and does not match the format specified in the task’s own published instructions. This resource influenced only the field-label wording rule described in Section 3.5. The numeric notation and period-referencing conventions discussed in Section 2 were derived separately from the first, non-overlapping resource and are unaffected by this issue. We were unable to revise this after identifying the discrepancy, as the submission window had closed. Based on post-hoc ablation, we estimate this specific formatting choice affected our reported score by approximately one percentage point. We report this for full transparency.

Finally, as discussed in Section 4.4, our system occasionally fails to constrain its answer to the specific facet of a question template when a more prominent, unrelated topic is present in the same news coverage; we were not able to fully resolve this within the scope of this work.

References

Xueqing Peng, Lingfei Qian, Yan Wang, Ruoyu Xiang, Yueru He, Yang Ren, Mingyang Jiang, Vincent Jim Zhang, Yuqing Guo, Jeff Zhao, Huan He, Yi Han, Yun Feng, Yuechen Jiang, Yupeng Cao, Haohang Li, Yangyang Yu, Xiaoyu Wang, Penglei Gao, and

28 others. 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, San Diego, California, United States.