

AhsanBLAH at the FinEval 2026: Grounded Two-Stage Financial QA with LLM-Judge Verification

Muhammad Ahsan Atiq¹ and Mohib Ahmad Butt²

¹Department of Computer Science, COMSATS University Islamabad

²Department of Robotics & Artificial Intelligence, SZABIST

¹sp24-bds-028@isbstudent.comsats.edu.pk ²mohib.ahmad.butt@gmail.com

Abstract

Financial question answering demands rigorous synthesis of standardized SEC filings alongside unstructured, cross-document text. This complexity is significantly amplified in multilingual environments, where terminology and reporting conventions vary across languages. We present a two-stage pipeline for the FinNLP 2026 PolyFiQA shared task, with all inference routed through the OpenRouter API under the 70B parameter limit. Stage 1 uses Qwen3.8-27B with category-calibrated few-shot prompts to extract facts and synthesize news across five languages (English, Chinese, Japanese, Spanish, Greek). Stage 2 is an LLM-as-a-Judge formatting guard that checks format compliance, strips filler, and verifies citations, without rewriting content. On our 76-instance training/development set (used for both calibration and evaluation, no held-out split), the calibrated pipeline reaches 48.19% Macro ROUGE-1 F1, an 88.3% relative gain over a zero-shot 70B baseline and a 6.32-point margin over Llama-3-70B’s own few-shot result. Gains concentrate in categories needing narrative synthesis (Revenue Trends, Cash Flow Dynamics) and are smaller for terse, formulaic ones (R&D Investments). On the official blind test set, our system scores 37.02% macro ROUGE-1 F1 on the private leaderboard (52 instances) and 31.79% on the public slice (24 instances).

1 Introduction

Financial analysts routinely reconcile forward-looking commentary from regional press coverage with statutory disclosures in corporate filings. Automating this reconciliation process using large language models introduces two central challenges: discursive meta-commentary that dilutes analytical precision, and numeric hallucination that is inadequately grounded in source evidence.

The FinNLP 2026 PolyFiQA shared task evaluates cross-document reasoning across SEC filings

(10-Q, 10-K) and financial news in five languages (English, Chinese, Japanese, Spanish, and Greek). Questions span four core financial categories: Revenue Trends & Focuses, Balance Sheet Health, Cash Flow Dynamics, and Research & Development (R&D) investments. Submissions require an exact two-part format: a concise factual answer followed by verbatim filing evidence citations.

Our contribution is an API-driven, two-stage framework:

- **Category-Aware Few-Shot Conditioning:** We calibrate prompt structures and exemplar formulas specifically tailored for revenue growth, balance sheet ratios, cash flow dynamics, and R&D expense metrics, operating strictly within the $\leq 70B$ parameter budget via OpenRouter (<https://openrouter.ai/api/v1>), the sole inference route used throughout this work, in accordance with the shared task’s resource constraints.
- **LLM-as-a-Judge Formatting Guard:** Stage 1 drafts the grounded answer; Stage 2 acts as a verification guard that audits mandatory two-section format compliance, strips conversational preambles, and verifies verbatim citation grounding against source text.
- **Transparent Evaluation:** We report detailed internal ablations comparing Qwen3.8-27B and Llama-3-70B, under both zero-shot and few-shot conditioning, across all question categories and languages, alongside official results on the public and private competition leaderboards.

2 Related Work

Benchmark datasets such as FinQA (Chen et al., 2021) and TAT-QA (Zhu et al., 2021) established numerical reasoning standards over tabular English financial documents. MultiFinBen (Peng et al.,

2026) and PolyFiQA extend this paradigm to cross-document, multilingual contexts, where terminology and disclosure conventions vary substantially across languages. Prior approaches to such generalization have largely relied on resource-intensive parameter fine-tuning or larger proprietary models, both of which are impractical under strict compute or API budgets. In contrast, our approach demonstrates that category-calibrated prompt conditioning paired with an LLM-judge guard stage on frontier open-weight models provides competitive precision under strict compute bounds, without any fine-tuning.

3 Task Formulation & Data Splits

3.1 Dataset Structure & Document Grouping

PolyFiQA pairs multi-page SEC filings from multinational corporations with multilingual press coverage. The benchmark is partitioned into two distinct splits:

1. Training Split. Comprises seventy-six annotated instances with reference answers evenly distributed across four financial reasoning categories (nineteen instances per category). This split is utilized for prompt calibration, few-shot exemplar curation, and internal comparative ablations.

2. Official Blind Test Split. The hidden-answer test benchmark contains seventy-six instances structured into nineteen document groups, ensuring all four question types associated with a single corporation and period remain grouped together during blind evaluation:

- **Public Leaderboard Slice:** Encompasses twenty-four instances across six document groups, utilized during active competition for initial progress tracking.
- **Final Private Leaderboard:** Encompasses the remaining fifty-two instances across thirteen document groups, utilized for final blind competition ranking based on macro-averaged ROUGE-1 F1.

3.2 Required Submission Schema

Outputs must strictly adhere to the following schema:

Answer: A concise synthesis specifying company name, period, and exact numerical figures (\$B, \$M, %).

Financial Statement Evidence: Verbatim quote from the multilingual financial news, or “None.”

Note that the training-stage exemplars in Appendices A.2–A.3 use the label `News Evidence:`, an earlier iteration of this field; the schema above, using `Financial Statement Evidence:`, is the one enforced in the test-set pipeline (Appendix A.4) and validated by the Stage 2 guard (Appendix A.5).

4 System Architecture

4.1 Backbone Model Selection

We deploy Qwen3.8-27B (Qwen Team, 2026) via the OpenRouter API as our primary engine, while benchmarking against Llama-3-70B (Grattafiori et al., 2024), also routed through OpenRouter. With 27B parameters, Qwen3.8 operates well below the shared task’s 70B ceiling, offering strong multilingual performance, high token-generation speed, and low latency. No local model hosting or third-party routing services outside OpenRouter were used at Stage 1. See Figure 1 in Appendix A.1 for a visual overview of the pipeline.

4.2 Stage 1: Grounded Synthesis

The Stage 1 generation prompt incorporates the query, filing context, and multilingual news excerpts alongside category-calibrated few-shot exemplars drawn from the training distribution. The model extracts relevant figures, identifies cross-lingual news alignment across European and Asian languages, computes standardized ratios, and defaults to None when evidence or statements are absent.

4.3 Stage 2: LLM-as-a-Judge Formatting Guard

Rather than attempting generative rewriting or self-correction, Stage 2 operates as a post-generation LLM-as-a-Judge formatting and citation-parsing guard, implemented with Qwen3.8-27B (temperature 0.1) served via the OpenRouter API. It checks the candidate draft against the following criteria:

- 1. Format Compliance:** Enforces the mandatory two-section header structure (Answer: and Financial Statement Evidence:).
- 2. Anti-CoT & Filler Pruning:** Strips chain-of-thought tokens (`<think>`), code fences, and conversational preambles (e.g., “Based on the provided text...”).

3. **Evidence Grounding:** Validates that cited line items correspond verbatim to source excerpts.

Stage 2 is deliberately conservative: it corrects violations of these criteria while preserving the grounded numerical content produced in Stage 1. Its role is structural rather than score-driven: since Stage 1 already produces largely well-formed output on the current 76-instance benchmark, the guard is intended as a fixed safeguard for robustness on future, uncalibrated inputs (new companies, filings, or news sources) rather than a component expected to lift in-sample ROUGE. We did not run a controlled Stage-1-only vs. Stage-1+Stage-2 ROUGE ablation; its effect is reported qualitatively, not as a measured score gain.

5 Experimental Results

5.1 Training Benchmark Ablations

We report in-sample development metrics evaluated against gold reference answers across all seventy-six instances in the training/development set; the same instances were used for few-shot exemplar selection and prompt calibration, so this is not a held-out validation score.

Zero-Shot vs. Few-Shot, Across Models. Table 1 compares zero-shot and category-calibrated few-shot prompting across both Llama-3-70B and Qwen3.8-27B, all runs executed via OpenRouter.

Few-shot conditioning improves both models substantially. For Qwen3.8-27B, precision rises sharply from 15.99% (zero-shot) to 47.98% (few-shot), while recall stays comparatively stable (55.23% to 56.63%), indicating that category-calibrated exemplars primarily sharpen answer specificity and format adherence rather than expanding raw coverage. Llama-3-70B shows a similar overall gain, with F1 rising from 25.59% to 41.87%.

Performance by Question Category. Table 2 details the breakdown of our Qwen3.8-27B few-shot pipeline across all four financial reasoning categories on the training dataset.

Baseline vs. Final System, by Category. Table 3 contrasts the Llama-3-70B zero-shot baseline with the final calibrated Qwen3.8-27B system on a per-category basis, isolating where category-aware calibration contributes most. Note this comparison also changes the backbone model, so the deltas

reflect the joint effect of model switch and calibration, not calibration alone. Holding the model fixed (Table 1), Llama-3-70B alone improves from 25.59% to 41.87% zero-shot to few-shot, indicating calibration contributes independently of the model switch.

Gains concentrate in Revenue Trends, Cash Flow Dynamics, and Balance Sheet Health, where the baseline’s ungrounded, low-recall zero-shot answers benefited most from exemplars that anchor exact figures and formatting. R&D Investments is the exception: the baseline already scores highly (53.01%) and the final system scores slightly lower (47.17%, -5.84), likely because R&D disclosures are short, formulaic line items that zero-shot extraction already captures well, leaving little headroom and some risk of exemplar-induced lexical drift from terse gold phrasing. We report this regression directly, since it is an informative finding about where calibration helps.

Language-by-Language Multilingual Performance. Table 4 reports the multilingual performance of Qwen3.8-27B across the five evaluation languages in the dataset. Citation counts exceed the seventy-six instance total because a single answer may ground claims in multiple language sources.

The model exhibits strong cross-lingual consistency across European and East Asian scripts, with Japanese and Spanish achieving above 50% ROUGE-1 F1.

Post-Guard Model Comparison. Table 5 summarizes the overall post-guard, few-shot metrics across both evaluated models on the internal training benchmark.

5.2 Official Test & Leaderboard Results

Table 6 reports official evaluation results on the hidden test set, scored under the competition’s macro-averaged ROUGE-1 F1 metric.

On the shared public leaderboard slice, our Qwen3.8-27B pipeline scored 31.79% macro ROUGE-1 F1, compared to 30.17% for a Llama-3-70B reference run on the same slice. On the full private leaderboard comprising fifty-two instances across thirteen document groups, our system achieved 37.02% macro ROUGE-1 F1. Public and private splits differ in size and composition and are reported as standalone results rather than a controlled comparison.

Model / Configuration	ROUGE-1 F1	ROUGE-1 Precision	ROUGE-1 Recall	ROUGE-2 F1	ROUGE-L F1
Llama-3-70B (Zero-Shot)	25.59%	63.68%	22.25%	18.58%	24.34%
Qwen3.8-27B (Zero-Shot)	22.25%	15.99%	55.23%	—	—
Llama-3-70B (Few-Shot)	41.87%	41.23%	51.48%	23.03%	34.31%
Qwen3.8-27B (Few-Shot)	48.19%	47.98%	56.63%	27.94%	40.65%

Table 1: Benchmark evaluation across model and prompting configurations on the training/development dataset (seventy-six instances), all runs routed via the OpenRouter API. ROUGE-2/ROUGE-L were not computed for the Qwen3.8-27B zero-shot run.

Category	Count	R-1 F1	R-1 P	R-1 R	R-2 F1
Balance Sheet Health	19	53.14%	51.03%	58.19%	28.50%
Revenue Trends	19	52.46%	53.59%	55.86%	33.31%
R&D Investments	19	47.17%	50.62%	63.80%	32.52%
Cash Flow Dynamics	19	39.98%	36.67%	48.67%	17.41%
Macro Average	76	48.19%	47.98%	56.63%	27.94%

Table 2: Qwen3.8-27B category-level performance breakdown across the four financial evaluation tasks.

Category	Baseline R-1 F1	Final R-1 F1	Δ
Revenue Trends	10.51%	52.46%	+41.95
Cash Flow Dynamics	17.97%	39.98%	+22.01
Balance Sheet Health	20.90%	53.14%	+32.24
R&D Investments	53.01%	47.17%	−5.84
Macro Average	25.59%	48.19%	+22.60

Table 3: Per-category ROUGE-1 F1 for the Llama-3-70B zero-shot baseline versus the final calibrated Qwen3.8-27B few-shot system.

Target Language	Citations	R-1 F1	Precision	Recall	R-2 F1
Japanese (JA)	9	55.58%	56.62%	59.31%	36.83%
Spanish (ES)	16	50.52%	53.41%	50.82%	30.73%
Greek (EL)	8	48.21%	48.62%	51.60%	30.50%
Chinese (ZH)	16	47.21%	49.65%	48.88%	28.22%
English (EN)	61	47.14%	46.41%	56.87%	26.20%

Table 4: Language-by-language performance and citation grounding fidelity on the benchmark.

Metric	Llama-3-70B	Qwen3.8-27B
ROUGE-1 F1 (%)	41.87	48.19
Precision (%)	41.23	47.98
Recall (%)	51.48	56.63
ROUGE-2 F1 (%)	23.03	27.94
ROUGE-L F1 (%)	34.31	40.65
Mean Word Count	142	128

Table 5: Comprehensive evaluation metrics on the training benchmark comparing both models after few-shot conditioning and the Stage 2 guard. These figures match the few-shot rows in Table 1.

Model	Split	Instances	Macro R-1 F1
Llama-3-70B	Public	24 queries (6 groups)	0.3017 (30.17%)
Qwen3.8-27B	Public	24 queries (6 groups)	0.3179 (31.79%)
Qwen3.8-27B	Private	52 queries (13 groups)	0.3702 (37.02%)

Table 6: Official leaderboard results on the blind test dataset, all inference via the OpenRouter API.

6 Discussion & Analysis

Stage 2 as a Formatting and Grounding Guard. Qwen3.8-27B produces largely well-formed, factually grounded drafts at Stage 1 given strong few-shot conditioning. The post-guard metrics in Table 5 reflect this stability. Stage 2 serves as a structural safeguard rather than a score-improving component: it enforces the mandatory two-section header format, strips conversational filler and chain-of-thought tokens, and validates citation grounding against source text. We view it as insurance against distribution shift as the pipeline meets new filings or companies beyond the current calibration set, rather than a lever expected to move in-sample ROUGE.

Cross-Lingual Observations. Table 4 shows citation retrieval holds up consistently across non-English sources, reflecting Qwen’s multilingual pretraining.

Where Calibration Helps. Table 3 shows the largest gains where zero-shot baselines struggle most with recall and grounding (Revenue Trends, Cash Flow Dynamics, Balance Sheet Health), lifting macro F1 by 22.60 points (+88.3% relative) — reflecting the joint effect of the backbone switch and few-shot calibration together. R&D Investments is the exception, suggesting calibration effort is best directed at categories with diffuse, narrative-style disclosures rather than terse, formulaic ones.

Evaluation Metric Nuances. ROUGE rewards exact lexical overlap. Numerically correct answers expressed under differing unit conventions (e.g.,

\$35,000M vs. \$35.0B) can incur minor surface penalties despite conceptual accuracy.

7 Limitations

No held-out development split was used: the same 76 instances informed exemplar selection, prompt calibration, and the reported 48.19% score. Our set is constrained to a small set of statutory filings, and category-level results carry natural statistical variance given nineteen instances per category. We did not run a controlled Stage-1-only vs. Stage-1+Stage-2 ROUGE ablation; the guard’s benefit is reported qualitatively. ROUGE metrics reward lexical overlap rather than direct symbolic algebraic verification. Public and private leaderboard splits differ in size and document composition, so cross-split comparisons are treated as indicative rather than direct controlled trials. ROUGE-2/ROUGE-L were not computed for the Qwen3.8-27B zero-shot configuration.

8 Conclusion

We presented a two-stage financial QA architecture combining category-calibrated few-shot extraction with an LLM-judge formatting guard, with all inference executed via the OpenRouter API. On the FinNLP 2026 PolyFiQA benchmark, our system achieves 48.19% Macro ROUGE-1 F1 in-sample, 31.79% on the official public leaderboard slice, and 37.02% on the official private leaderboard, within the 70B parameter ceiling.

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*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Aaron Grattafiori et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- 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. *Preprint*, arXiv:2506.14028. To

appear in Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026).

Qwen Team. 2026. Qwen3.8-27b: Open-weights frontier and multimodal benchmarks.

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 and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

A Appendix

A.1 Pipeline Overview

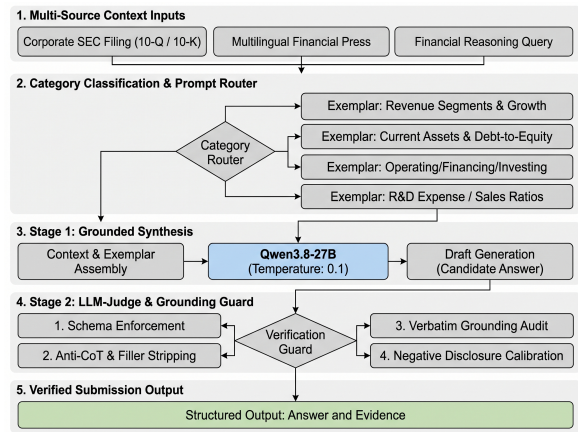


Figure 1: Overview of the two-stage pipeline: Stage 1 category-calibrated few-shot generation followed by the Stage 2 LLM-as-a-Judge formatting guard.

A.2 Zero-Shot Prompt

Zero-Shot Prompt.

SYSTEM PROMPT:
You are an expert financial analyst answering cross-document questions using corporate SEC filings (10-Q/10-K) and multilingual news (English, Chinese, Japanese, Spanish, Greek).

MANDATORY INSTRUCTIONS:

1. Start your response IMMEDIATELY with 'Answer:' followed by your concise factual finding.
2. End your response IMMEDIATELY with 'News Evidence:' (or 'Financial Statements Evidence:') quoting the verbatim supporting line items or write 'None.' if absent.
3. If the requested financial statement or specific metric is not present in the context, write:
Answer: None.
News Evidence: None.
4. Do not include any reasoning, preambles, introductory filler, or conversational meta-talk.

USER PROMPT:
QUESTION:

{question}
CONTEXT:
{query}

A.3 Few-Shot Prompt

Few-Shot Prompt (Category-Calibrated Exemplars).

SYSTEM PROMPT:

You are an expert financial analyst answering questions from corporate SEC financial filings and multilingual news (English, Chinese, Japanese, Spanish, Greek).

MANDATORY INSTRUCTIONS:

1. Start IMMEDIATELY with 'Answer:' followed by your financial findings.
2. If the requested financial statement or specific line item (e.g. Cash Flow Statement or R&D expense) is completely absent in the context, write:
Answer: None.
News Evidence: None.
3. Structure your response based on query category:
 - R&D Ratio: Write "[Company]'s R&D ratio for [Period] was [X]% (R&D expense of [A] divided by revenue/sales of [B]). This represents [an increase / stable / decrease] from [Y]% in [Prior Period], indicating consistent investment in innovation relative to revenue." If R&D is not disclosed, write "Answer: None. There is no R&D expense listed in the financial statements."
 - Balance Sheet: State Total Current Assets for current and prior periods, Total Liabilities, Total Stockholders' Equity, the calculated Total-Liability-to-Equity ratio (Total Liabilities / Total Equity) for current vs prior periods, and conclude with financial health/leverage.
 - Cash Flow: State Operating cash flow (current vs prior), Financing cash flow (current vs prior), and Investing cash flow (current vs prior), and conclude with "No major irregularities were observed."
 - Revenue: State Total Revenue for current and prior period, YoY % growth, main segment drivers, followed by exact multilingual news quotes.
4. End IMMEDIATELY with 'News Evidence:' quoting multilingual news or 'None.'

CATEGORY-CALIBRATED FEW-SHOT EXEMPLARS:

[CATEGORY: R&D RATIO]

User:

QUESTION:

What is the company's R&D ratio (R&D divided by revenue)? Please provide only the final result(s); do not include any calculation steps. Please quote the piece(s) of news aligned with the finding(s), if any.

CONTEXT:

[Financial Statement for J&J: Sales to customers \$22,321M in Q1 2021 vs \$20,691M in Q1 2020. Research and development expense \$3,178M in Q1 2021 vs \$2,581M in Q1 2020.]

A:

Answer: Johnson & Johnson's R&D ratio for Q1 2021 was 14.2% (R&D expense of \$3,178 million divided by sales of \$22,321 million). This represents an increase from 12.5% in Q1 2020, indicating the company is investing more in research and development relative to revenue.
News Evidence: None.

[CATEGORY: CASH FLOW / CAPITAL ALLOCATION]
User:

QUESTION:

Are there any significant changes or irregularities in the amount of operating, investing and financing cash flow? Please quote the piece(s) of news aligned with the finding(s), if any.

CONTEXT:

[Financial Statement for Microsoft: Net cash from operations \$17,504M in Q3 FY2020 vs \$13,520M in Q3 FY2019. Net cash used in financing \$(14,645)M vs \$(7,601)M. Net cash from investing \$51M vs \$(1,363)M.]

A:

Answer: In Q3 FY2020, Microsoft reported \$17.5B inflow in operating cash flow up from \$13.5B in Q3 FY2019; Microsoft used \$14.6B in financing cash outflow, up from \$7.6B outflow in Q3 FY2019; and Microsoft had minimal investing activity at \$0.05B net inflow, vs. \$1.4B outflow in Q3 FY2019. No major irregularities were observed.
News Evidence: None.

User:

QUESTION:

Are there any significant changes or irregularities in the amount of operating, investing and financing cash flow? Please quote the piece(s) of news aligned with the finding(s), if any.

CONTEXT:

[Financial Statement: Only Balance Sheet and Statement of Earnings are provided. No Statement of Cash Flows is in the filing.]

A:

Answer: None.

News Evidence: None.

[CATEGORY: BALANCE SHEET HEALTH]

User:

QUESTION:

How does the company's balance sheet reflect its financial health in terms of total amount of current assets and the ratio of total liability divided by total equity? Please provide only the final result(s); do not include any calculation steps. Please quote the piece(s) of news aligned with the finding(s), if any.

CONTEXT:

[Financial Statement for Microsoft: Total current assets \$170,505M at March 31, 2020 vs \$175,552M at June 30, 2019. Total liabilities \$170,948M vs \$184,226M. Total stockholders' equity \$114,501M vs \$102,330M.]

A:

Answer: Microsoft's current assets totaled \$170.5B in Q3 FY2020, down from \$175.6B in Q3 FY2019. Microsoft's total-liability-to-equity ratio reduced to 1.49 in Q3 FY2020 from 1.80 in Q3 FY2019, indicating an improvement in financial health.
News Evidence: None.

[CATEGORY: REVENUE TRENDS & MULTILINGUAL NEWS]

User:

QUESTION:

What trends can be observed in the company's revenue amount over the past few years? Please quote the piece(s) of news aligned with the finding(s), if any.

CONTEXT:

[Financial Statement for Microsoft: Revenue \$35,021M in Q3 FY2020 vs \$30,571M in Q3 FY2019] [Japanese news, Spanish news, Chinese news excerpts reporting the same \$35.0B revenue figure and ~15% YoY growth.]

A:

Answer: Microsoft's total revenue increased to \$35.0B in Q3 FY2020 (Q1 2020 calendar year), a

14.6% YoY growth from \$30.6B in Q3 FY2019, driven by both product and service revenue growth. News Evidence: Japanese news, Spanish news, and Chinese news excerpts corroborating the \$35.0B revenue figure and approximate 15% YoY growth (see Figure 2 for verbatim quotations).

Margin, Operating Income, and Net Income for current vs prior periods.

4. Capital Expenditures: In Answer, synthesize capital expenditure amounts and strategic projects from news; if news does not discuss strategic capex, write 'Answer: None.'. In Financial Statements Evidence, quote additions to property and equipment / capital expenditures for reported periods.



Figure 2: Verbatim multilingual news quotations (Japanese, Spanish, Chinese, Greek) for the Revenue Trends few-shot exemplar, reproduced here as an image due to CJK glyph rendering constraints under pdflatex.

A.5 Stage 2 Judge Prompt

Stage 2 Formatting & Citation-Parsing Guard Prompt.

You are an expert financial answer rectifier.

MANDATORY RECTIFICATION RULES:

1. Output EXACTLY two sections with NO introduction, commentary, or conclusion:
Answer: [Your concise answer]
Financial Statement Evidence: [Verbatim line quotes with exact dollar amounts]
2. Strip all meta-talk (e.g. "According to the news...", "While details are not provided...").
3. Ensure all specific phrasing, dollar amounts (\$B, \$M) and percentages (%) from the draft are accurately preserved. Do not over-summarize if it deletes valid context.

A.4 Test Pipeline System Prompt

Test Pipeline System Prompt (Test-Set Variant).

TEST PIPELINE SYSTEM PROMPT (B.2 - Test Set Variant):
You are an expert financial analyst answering cross-document questions using corporate SEC filings (10-Q/10-K) and multilingual news (English, Chinese, Japanese, Spanish, Greek).

MANDATORY RESPONSE FORMAT:
Every output MUST contain EXACTLY two labeled sections:
Answer: [Your concise synthesized findings based on the question and multilingual news. If the specific aspect asked is not discussed in the news, write 'Answer: None.']
Financial Statements Evidence: [Relevant line items and numerical facts quoted from the financial statements (Income Statement, Balance Sheet, Statement of Cash Flows) supporting the findings.]

QUESTION GUIDELINES:

1. Top Focuses on Revenue: In Answer, list the 3 main business segments / growth drivers highlighted in the news (e.g. 1. [Segment] - [driver/growth], 2. [Segment] - [driver/growth], 3. [Segment] - [driver/growth]). In Financial Statements Evidence, quote Total Revenue, Service/Product revenue, and segment revenues for current and prior periods.
2. Capital Allocation: In Answer, synthesize how the company allocates capital (investments, share repurchases, dividends, capex) from the news; if news does not discuss capital allocation plans, write 'Answer: None.'. In Financial Statements Evidence, quote Share repurchases, Dividends paid, Capital expenditures, and Net cash from operations.
3. Maintaining Profit Margins: In Answer, synthesize the company's strategies to navigate/maintain profit margins from news (e.g. high-margin cloud mix, digital transformation, operational cost control). In Financial Statements Evidence, quote Gross