pjmathematician @ FinMMEval 2026: Systems for Tasks 1, 2 and 3

FinMMEval Lab at CLEF 2026

Poojan Vachharajani[1]

[1]*Amazon*

Abstract

Financial natural language processing has emerged as a critical research frontier, demanding systems capable of multilingual reasoning, domain-specific knowledge retrieval, and real-time market decision-making. This paper presents our participation in FinMMEval at CLEF 2026, where we submitted systems for all three tasks: Task 1 (Financial Exam Q&A), Task 2 (Multilingual Financial Q&A), and Task 3 (Financial Decision Making).

For Task 1, we developed a retrieval-augmented multiple-choice answering system leveraging Google Gemini with real-time Google Search grounding. By combining the model's parametric financial knowledge with live web retrieval, our system answered exam-style questions spanning valuation, accounting, ethics, corporate finance, and regulatory knowledge across multiple languages. Our approach achieved competitive results across all evaluated languages, ranking 1st in English (97.50%, 195/200), 1st in Chinese (96.50%, 193/200), 1st in Arabic (97.50%, 195/200), and 1st-2nd tied in Hindi (92.00%, 184/200), demonstrating strong cross-lingual generalization.

For Task 2, we designed a structured multilingual financial reasoning pipeline using Gemini to generate concise, evidence-grounded answers from SEC filings and multilingual news articles. Our pipeline incorporated iterative validation steps enforcing structural formatting and a strict 100-word answer length constraint, ensuring output quality and consistency. Our system achieved a ROUGE-1 F1 score of 29.72%, ranking 5th on the final leaderboard among all complete submissions.

For Task 3, we developed a multi-signal ensemble trading agent combining technical analysis (RSI, moving averages, z-scores), fundamental analysis from SEC 10-K/10-Q filings, news sentiment, and real-time Google Search grounding. A weighted ensemble strategy aggregated signals from all components, with a final LLM-based arbitration step designed to suppress decision hysteria and ensure position stability.

Our results demonstrate that grounding large language models with real-time retrieval and structured multi-signal ensembling is a highly effective strategy for financial NLP tasks across diverse languages and reasoning modalities.

Keywords

financial question answering, multilingual reasoning, retrieval-augmented generation, trading decision making

1. Introduction

The integration of artificial intelligence within the financial sector demands high levels of reasoning capability, multilingual comprehension, and structured decision-making [1]. Traditional Large Language Models (LLMs) often lack up-to-date market information and are susceptible to hallucinations, limiting their reliability on domains requiring absolute accuracy, such as regulatory compliance, professional examinations, and automated trading.

The FinMMEval Lab 2026 offers a comprehensive platform to test the capabilities of current state-of-the-art AI systems across three core financial domains [1]:

1. **Task 1: Financial Exam Q&A**, which measures accuracy on professional financial examinations such as the CFA, CPA, and EFPA across multiple languages [2].
2. **Task 2: Multilingual Financial Q&A**, which requires extracting evidence-grounded, concise answers from a combination of SEC filings and multilingual news articles [3].
3. **Task 3: Financial Decision Making**, which demands a daily news-driven trading agent that executes positions over real-time market assets [4].

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

pjmathematician@gmail.com (P. Vachharajani)

To tackle these tasks, we present the solutions developed by the **pjmathematician** team. For Task 1, we present a Retrieval-Augmented Generation (RAG) system utilizing Google Gemini grounded with live Google Search queries. For Task 2, we implement an iterative validation pipeline that aims to systematically enforce strict formatting syntax and word count constraints. Finally, we build a multi-signal ensemble trading agent that mathematically synthesizes technical indicators with fundamental sentiment and search-backed data, governed by a Chief Investment Officer (CIO) LLM arbitration layer.

2. Related Work

2.1. Financial Question Answering and RAG

Evaluating LLMs on finance-specific datasets has progressed rapidly. Early research focused on numerical reasoning datasets such as FinQA [5] and document-grounded question answering via FinanceBench [6]. Traditional retrieval-augmented generation (RAG) mechanisms often struggle with complex domain reasoning. To address this, current benchmarks focus on evaluating implicit premises and missing information, as demonstrated by RealFin [7], as well as regional and regulatory compliance contexts, such as the SAHM Arabic financial benchmark [8].

2.2. Multilingual Financial NLP

Most initial financial LLM evaluations were heavily biased toward English. Benchmarks such as MultiFinBen [9] address this limitation by evaluating model capabilities in cross-lingual environments where the source data and query languages diverge. These benchmarks emphasize that multilingual financial question answering requires high-stakes semantic synthesis and strict formatting alignment to capture real-world operational standards.

2.3. Agentic Financial Trading Systems

LLM-based quantitative systems represent a paradigm shift from traditional reinforcement learning approaches like FinRL [10]. Recent frameworks such as FinMem [11] implement memory and character designs to preserve long-term context. More advanced architectures organize systems as cooperative teams or multi-agent ensembles, mitigating individual model biases and ensembling quantitative signals (such as technical indicators) with qualitative textual features (such as news or regulatory filings) to generate stable trade executions on live testbeds like Agent Market Arena [12].

3. Task 1: Financial Exam Q&A

3.1. Task Definition and Challenges

Task 1 tests an agent's ability to answer multiple-choice questions from professional finance exams covering core areas like accounting, valuation, corporate finance, ethics, and regulatory laws [2]. The dataset represents a diverse linguistic footprint, incorporating source benchmarks such as RealFin (English and Chinese) [7], BhashaBench-Finance (Hindi) [13], and SAHM (Arabic) [8]. Because financial regulations change and complex calculations are often needed, parametric knowledge alone is frequently insufficient.

3.2. System Architecture

Our solution utilizes a RAG pipeline powered by the `gemini-3-flash-preview` model [14]. To overcome knowledge gaps regarding local financial laws and recent accounting changes, we integrated the Google Search grounding tool directly into the generation layer using the official Gemini API SDK.

The workflow proceeds as follows:

1. **Prompt Generation**: For each question Q and candidate options $\{A, B, C, D\}$, we format a unified query string instructing the model to perform a detailed search, analyze the findings, and generate its final decision.
2. **Grounding and Execution**: The Gemini client invokes Google Search queries internally using the `google_search` tool, compiling the returned context.
3. **Strict Parsing**: The output format is explicitly constrained to terminate with `Correct Answer: <Option Letter>` to enable automated extraction.
4. **Post-processing**: In the event of minor extraction anomalies (e.g., non-alphabetic characters yielded due to regional script translation errors in Hindi/Arabic), regex patterns clean the responses to default options.

The final submissions were produced in a single, direct inference run; no programmatic retry loops were permitted. The post-processing regex mapping affected only 1.5% of overall outputs (specifically, 3 questions out of 200 within the Hindi subset experienced minor translation artifacts which were successfully parsed, while 0% of English, Chinese, and Arabic answers required fallback adjustments).

The prompt utilized for Task 1 is presented in Appendix A.

4. Task 2: Multilingual Financial Q&A

4.1. Task Definition and Challenges

Task 2 evaluates cross-lingual reasoning across heterogeneous financial documents [3]. Given a query, English SEC 10-K/10-Q filing segments, and related news articles written in Spanish, Chinese, Japanese, or Greek, systems must formulate a highly structured response [9]. Crucially, the final output must strictly follow two rigid constraints:

- It must preserve separate sections labeled `Answer:` and `Financial Statements Evidence:`.
- The text within the `Answer:` subsection must remain strictly concise and contain 100 words or fewer.

4.2. System Architecture

Instead of relying on single-turn inference, which often fails to adhere to rigid length constraints, we designed an **Iterative Multimodal Validation Pipeline**. The sequential execution flow is illustrated below:

1. **Initial Inference**: We prompt `gemini-3-flash-preview` to read the composite inputs and generate the answer and supporting evidence.
2. **Structural Validation**: A rule-based parser checks the text for the exact presence of the two required header patterns. If either is absent, a specialized formatting correction prompt is issued to regenerate the response structure.
3. **Length Validation**: The text bounded by the `Answer:` header is parsed to calculate the precise word count. If the word count exceeds 100 words, a subsequent compression prompt is executed, demanding a reduction in length while preserving the informational integrity of the response.

The implementation details of this sequential self-correction flow ensure robust compliance with the metrics without degrading factual ROUGE scores.

5. Task 3: Financial Decision Making

5.1. Task Definition and Challenges

Task 3 simulates a daily news-driven trading workflow. Every day at 00:00 UTC, the system receives a payload containing past price history, current asset price, daily news headlines, momentum metrics, and optional SEC filings. The agent must yield an action: BUY (long), HOLD (flat), or SELL (short) [4, 12].

The primary challenges lie in preventing trading hysteria (excessive swapping of positions due to noisy micro-signals) and managing the diverse signals across technical, fundamental, and sentiment perspectives.

5.2. System Architecture

Our system consists of a multi-signal ensemble agent governed by a hierarchical architecture. Rather than relying on a single model to digest all inputs, we split the analysis into specialized modules:

- **Technical Indicator Analyzer**: A deterministic Python parser computes standard quantitative metrics over the asset's historical price list, including the Relative Strength Index (RSI), a 5-day Simple Moving Average (SMA), and standard Z-scores. It generates a numeric raw direction along with a confidence metric representing signal convergence.
- **Fundamental Analyzer**: A specialized LLM block analyzes SEC 10-K and 10-Q inputs to determine if long-term fundamentals are bullish, bearish, or flat.
- **Real-Time Sentiment Analyzer**: Extracts sentiment signals from the daily news inputs using targeted sentiment prompts.
- **Google Search Grounding module**: Calls Google Search to query breaking developments regarding the asset on the request date, aligning temporal context.

5.3. Ensemble Optimization and CIO Arbitration

To consolidate these disparate indicators, we define a mathematically weighted raw score:

$$S_{raw} = w_{news}S_{news} + w_{past}S_{past} + w_{search}S_{search} + w_{tech}S_{tech} + w_{mom}S_{mom} + w_{fund}S_{fund} \quad (1)$$

where:

- $S_{raw} \in [-1.0, 1.0]$ represents the final aggregated, raw consolidated direction score.
- $S_i \in \{-1.0, 0.0, 1.0\}$ represents the individual directional signal from module i, mapping to SELL (-1.0), HOLD (0.0), and BUY (1.0).
- $w_i \in [0, 1]$ represents the assigned normalized weight of module i (such that $\sum w_i = 1.0$).

The specific input signal variables are defined as:

- S_{news}: Sentiment score from the daily news sentiment analyzer module.
- S_{past}: The past trading action or position (serving as a trend-following stabilizer to prevent panic-trading).
- S_{search}: Score derived from real-time web verification grounding.
- S_{tech}: Technical signal yielded by the deterministic indicator logic.
- S_{mom}: Baseline momentum indicator provided directly in the raw data payload (where bullish mapped to 1.0, bearish to -1.0, and neutral/flat to 0.0).
- S_{fund}: Long-term fundamental health score evaluated by parsing SEC filings.

The empirically optimized weights reflect our design hypothesis prioritising news, past state preservation, and search grounding over raw filing statistics: $w_{news} = 0.25$, $w_{past} = 0.25$, $w_{search} = 0.15$, $w_{tech} = 0.15$, $w_{mom} = 0.10$, and $w_{fund} = 0.10$.

Finally, the compiled inputs, the calculated raw score, and the previous position are passed to the **Chief Investment Officer (CIO) Arbitration Agent** (powered by `gemini-3.5-flash`). The CIO enforces trading stability rules: if the raw score is near zero, it overrides signals to produce a HOLD, avoiding unnecessary market friction and trading hysteria.

5.4. Implementation Details and Hyperparameters

Our pipeline uses different backbone configurations depending on the computational reasoning profile required by each task:

- **Model Selection**: For Tasks 1 and 2, we utilized `gemini-3-flash-preview` due to its low latency, high parallel throughput, and cost-efficiency, which are ideal for parallel execution over long documents. For Task 3, we deployed `gemini-3.5-flash` as our CIO arbiter. This model was selected because decision arbitration involves reasoning over contradictory signals and adhering to strict multi-agent execution rules (e.g., dynamic "hysteria" filtering), which benefit from the model's superior instruction-following capabilities.
- **Execution Hyperparameters**: Our configurations are described in Table 1. To maximize determinism for Task 1's multiple-choice exam answers, the temperature was set to $T = 0.0$.

Table 1
Pipeline Execution Configurations and Hyperparameters

Module	Backbone Model	Temperature (T)	Top-p / Top-k	Grounding Search Calls
Task 1 Exam Pipeline	`gemini-3-flash-preview`	0.0	1.0 / 1	1 call per question
Task 2 QA Pipeline	`gemini-3-flash-preview`	0.2	1.0 / –	None
Task 3 Sentiment / Search	`gemini-3-flash-preview`	0.1	1.0 / –	1 call per asset per day
Task 3 CIO Arbiter	`gemini-3.5-flash`	0.15	1.0 / –	None

6. Results

6.1. Task 1: Financial Exam Q&A

Our system achieved competitive performance across all four evaluated languages on the final hidden test sets. Our results are presented in Table 2.

Table 2
Official Task 1 Final-Test Accuracy Leaderboard Results

Language	Benchmark Dataset Source	Correct Answers	Accuracy (%)	Official Rank
English	CFA [7]	195 / 200	97.50%	**1st**
Chinese	CPA [7]	193 / 200	96.50%	**1st**
Arabic	SAHM [8]	195 / 200	97.50%	**1st**
Hindi	BhashaBench [13]	184 / 200	92.00%	**1st-2nd (Tied)**

6.1.1. Task 1 Ablation Study

To quantify the impact of real-time search grounding, we performed an ablation study on the English (CFA) and Arabic (SAHM) test configurations, comparing our search-grounded approach against a purely parametric, zero-shot configuration. The results are reported in Table 3. This ablation reveals that static model knowledge is frequently out-of-date regarding regional regulations, showing that real-time grounding is a major driver of our high accuracy.

6.2. Task 2: Multilingual Financial Q&A

The iterative validation pipeline successfully optimized answer formats under strict constraints. Our submission achieved an official **ROUGE-1 F1 score of 29.72%**, positioning our system **5th overall** on the final completed leaderboard.

Table 3
Task 1 Ablation Study: Grounded RAG vs. Static Zero-Shot Accuracy

Configuration	English CFA Accuracy	Arabic SAHM Accuracy
Zero-Shot (Static Parametric)	83.50%	79.00%
Our System (Google Search Grounded)	**97.50%**	**97.50%**

6.2.1. Task 2 Constraint Satisfaction and Ablation

We evaluated the necessity of our iterative feedback loop by measuring constraint compliance with and without our validation module. As shown in Table 4, omitting the feedback step led to a significant failure rate for both the structural syntax (required headers) and the maximum length (100-word) limit. The validation pipeline successfully achieved 100% constraint satisfaction across all processed test samples.

Table 4
Task 2 Constraint Satisfaction Comparison

Configuration	Structural Header Compliance (%)	Word Count Compliance ($\leq$ 100 words) (%)
Without Validator Loop	93.8%	85.5%
With Validator Loop	**100.0%**	**100.0%**

6.2.2. Qualitative Error Analysis

Although our validator loop ensured 100% compliance with the strict format and length parameters, our ROUGE-1 F1 score (29.72%) reflects the inherent limitations of unigram-matching evaluation for open-ended generation tasks. A qualitative review of our errors revealed two main sources of score degradation:

1. **Linguistic Synonymy**: The system generated semantically exact financial facts using valid synonyms (e.g., generating *"revenue expansion of $12M"* or *"operational costs increased"*), whereas the gold-standard references used direct unigrams (e.g., *"sales growth of $12 million"* or *"administrative expenses rose"*). This severely penalized ROUGE-1 precision despite equivalent meaning.
2. **Structural Style Mismatches**: The reference answers frequently consisted of highly condensed bullet points, while our generator synthesized complete continuous prose sentences to preserve logical consistency.

Table 5 displays a typical mismatch example.

Table 5
Example of Semantic Equivalence with Mismatched ROUGE unigram overlap (Task 2)

Query	Gold Expert Reference	System Output (Under 100-wo
What drove capital expenditures in Q1 2026?	Capex driven by Mojo and Optimus AI infrastructure expansion.	The primary driver of capital ex extensive capital allocation to t computer project and Optimus opment, as highlighted in the filing.

6.3. Task 3: Financial Decision Making

Our technical indicator modules and weighted CIO model successfully generated stable trading actions. Due to the ongoing evaluation structure over the designated market period (running through mid-2026),

quantitative trading results (Cumulative Return, Sharpe Ratio, Maximum Drawdown) remain subject to live market execution by the organizers. However, backtesting simulations indicated that our CIO arbitration filter successfully reduced portfolio turnover by 34% compared to unconstrained sentiment signals.

7. Discussion

The results across the FinMMEval tasks highlight key lessons for applying LLMs to financial AI:

- **Search Grounding is Essential for Compliance**: Across Task 1, static knowledge bases struggle with shifting legal and regulatory examination environments. Dynamically enabling search engines at the LLM reasoning layer provides a robust solution.
- **Structural Constraints Require Multi-Turn Validation**: Single-turn prompting often fails on strict syntax constraints such as exact token lengths or target headers. Programmatic validation loops with targeting adjustment prompts significantly reduce error rates.
- **Decoupled Ensembles Stabilize Multi-Agent Systems**: Combining specialized algorithms (such as mathematical technical indicators) with qualitative models (sentiment/fundamentals) reduces the risk of extreme individual signal failures.

8. Conclusion

In this paper, we described our participation as the **pjmathematician** team in the CLEF 2026 FinMMEval Lab. By leveraging Google Gemini models grounded in real-time search, implementing robust iterative validation loops, and deploying mathematical ensemble systems, we demonstrated consistent performance across Q&A, multilingual reasoning, and financial decision-making tasks. Future research will explore deeper integration of multimodal signals for complex financial risk analysis.

Acknowledgments

We thank the FinMMEval 2026 lab organizers for preparing the datasets and establishing the robust evaluation framework.

Declaration on Generative AI

During the preparation of this work, the author(s) used Google Gemini models (including gemini-3-flash-preview and gemini-3.5-flash) in order to execute the automated pipelines for Tasks 1, 2, and 3. Additionally, assistive LLMs were employed to perform spelling and grammar verification on the drafted sections of this manuscript. After using these services, the authors reviewed, refined, and edited all text and take full responsibility for the content of this publication.

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, X. 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, Y. Dai, R. Elbadry, V. Jani, G. Georgiev, D. Dimitrov, F. Zhang, X. Peng, L. Qian, J. Huang, J. Geng, Y. Chen, Y. Yuan, H. Wu, Y. Wang, I. Koychev, V. Stoyanov, M. Song, Y. Chen, X. Liu, P. Nakov, Overview of the FinMMEval 2026 task 1: Multilingual financial multiple-choice question answering, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[3] 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, X. 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.

[4] Z. Xie, L. Qian, G. Georgiev, D. Dimitrov, R. Elbadry, F. Zhang, X. Peng, J. Huang, V. Jani, Y. Dai, J. Geng, Y. Chen, Y. Yuan, H. Wu, Y. Wang, I. Koychev, V. Stoyanov, M. Song, Y. Chen, X. Liu, P. Nakov, Overview of the FinMMEval 2026 task 3: Financial decision making, in: CLEF 2026 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Jena, Germany, 2026.

[5] Z. Chen, W. Chen, C. Smiley, S. Shah, G. Borradaile, 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, 2021, pp. 3697–3711.

[6] A. Ismail, et al., FinanceBench: A new benchmark for financial question answering, arXiv preprint arXiv:2311.11944 (2023).

[7] Y. Dai, Y. Lin, Z. Xie, Y. Wang, RealFin: How well do LLMs reason about finance when users leave things unsaid?, in: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2026.

[8] R. Elbadry, S. Ahmad, A. Heakl, D. Bouch, M. Ahsan, M. AlMahri, M. E. Khalil, Y. Wang, S. Lahlou, S. Ananiadou, V. Stoyanov, J. Huang, X. Peng, P. Nakov, Z. Xie, SAHM: A benchmark for Arabic financial and shari
ah-compliant reasoning, in: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, San Diego, California, United States, 2026, pp. 34509–34536. URL: https://aclanthology.org/2026.acl-long.1593/. doi:10.48550/arXiv.2604.19098.

[9] 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.

[10] X.-Y. Liu, H. Yang, J. Chen, R. Li, D. Zhou, X. Wang, J. W. Suchow, M. Shou, K. Khashanah, FinRL: A deep reinforcement learning library for automated stock trading in quantitative finance, arXiv preprint arXiv:2011.09607 (2020).

[11] Y. Yu, H. Li, Z. Chen, Y. Jiang, Y. Li, D. Zhang, R. Liu, J. W. Suchow, K. Khashanah, FinMem: A performance-enhanced LLM trading agent with layered memory and character design, arXiv preprint arXiv:2309.03736 (2023).

[12] L. Qian, X. Peng, Y. Wang, V. J. Zhang, H. He, H. Smith, Y. Han, Y. He, H. Li, Y. Cao, Y. Yu, A. Lopez-Lira, P. Lu, J.-Y. Nie, G. Xiong, J. Huang, S. Ananiadou, When agents trade: Live multi-market trading benchmark for LLM agents, 2025. URL: https://arxiv.org/abs/2510.11695. doi:10.48550/arXiv.2510.11695. arXiv:2510.11695.

[13] V. Devane, M. Nauman, B. Patel, A. M. Wakchoure, Y. Sant, S. Pawar, V. Thakur, A. Godse, S. Patra, N. Maurya, S. Racha, N. K. Singh, A. Nagpal, P. Sawarkar, K. V. Pundalik, R. Saluja, G. Ramakrishnan, BhashaBench V1: A comprehensive benchmark for the quadrant of indic domains, 2025. URL: https://arxiv.org/abs/2510.25409. arXiv:2510.25409.

[14] Google LLC, Gemini 2.0: Multimodal Reasoning and Agentic Capabilities at Scale, Technical Report, Google Tech Report, 2025.

A. Task 1: System Prompts

The prompt format utilized to query Gemini grounded with real-time web search for Task 1 is defined as follows:

```
MAIN_PROMPT = """
You are given a finance question, and related options.
You have to think, and google search and finally give the answer in the following format
    :
Correct Answer: <Option Letter>

Example:
Correct Answer: A

Question: {}
Options: {}
"""
```

B. Task 3: System Prompts and Code

B.1. Fundamental Analysis Prompt

SEC filings are parsed with the following structural instruction set:

```
FUNDAMENTAL_PROMPT = """
You are a Senior Equity Analyst. Analyze the provided SEC Filing Summaries (10-K/10-Q)
    for {symbol}.
Focus on: Revenue growth, Debt levels, Risk factors, and Management guidance.

Filings Data:
- 10-K (Annual): {data_10k}
- 10-Q (Quarterly): {data_10q}

Evaluate if the company's fundamentals are improving, stable, or deteriorating.
Assign a momentum label: 'bullish' (improving), 'bearish' (deteriorating), or 'flat' (no
     major change).

Return strictly as JSON:
{
  "recommended_action": "BUY/HOLD/SELL",
  "confidence": float,
  "fundamental_momentum": "bullish/bearish/flat",
  "rationale": "Max 50 words focusing on financial health"
}
"""
```

B.2. Real-Time Google Search Grounding Prompt

Live verification queries utilize this grounding context:

```
SEARCH_ANALYSIS_PROMPT = """
You are a financial research agent. Your goal is to verify the current market sentiment
    for {symbol} using Google Search to complement the provided news.

Date of Request: {date}
Symbol: {symbol}
Provided News Summary: {news}
```

```
Instructions:
1. Search for any breaking financial news or price action for {symbol} on or around {
    date}.
2. Compare the search results with the provided news.
3. Determine if the momentum is Bullish, Bearish, or Flat.

Return the result STRICTLY in this JSON format:
{
  "recommended_action": "BUY/HOLD/SELL",
  "confidence": float,
  "momentum_label": "bullish/bearish/flat",
  "rationale": "A 50-word summary of how search data influenced this decision."
}
"""
```

B.3. Chief Investment Officer Ensemble Prompt

The consolidated arbiter utilizes the following strict system instruction context to output finalized positions:

```
ENSEMBLE_PROMPT = """
You are the Chief Investment Officer (CIO). Your job is to aggregate inputs from
    multiple specialized financial agents and output a final trade decision.

AGENT INPUTS:
- Technical Signal: {tech}
- Fundamental Signal: {fundamental}
- News Sentiment: {news}
- Google Search Research: {search}
- Past Position Context: {past}
- Raw Weighted Score: {score} (-1.0 for SELL, 0 for HOLD, 1.0 for BUY)

ENFORCED STRATEGY:
- If current Raw Score is near 0, prefer HOLD to avoid unnecessary turnover (hysteria).
- If the Raw Score contradicts the Past Position, provide a clear reason for the pivot.

Final Output MUST be strictly JSON:
{
  "recommended_action": "BUY/HOLD/SELL",
  "confidence": float,
  "momentum_label": "bullish/bearish/flat",
  "rationale": "Max 50 words explaining the ensemble consensus."
}
"""
```