

[기술 보고서] LLM Multi-Agent 기반 금융 RAG 및 포트폴리오 최적화 시스템

1. 시스템 개요 (Executive Summary)

본 시스템은 뉴스 데이터와 매크로 시나리오를 결합하여 최적의 자산 배분 가이드를 제공하는 **Self-Improving Financial RAG 시스템**입니다. 단순 검색(Retrieval)을 넘어, **상승론(Bull)과 하락론(Bear)의 논리적 경합(Debate)**을 통해 시장의 괴리(Divergence)를 분석하고, 그 결과를 바탕으로 퀸트 모델의 파라미터를 미세 조정(Fine-tuning)한 뒤 수학적 최적화(SLSQP)를 수행하는 End-to-End 파이프라인을 구축하였습니다.

2. 핵심 아키텍처 및 알고리즘 설명

2.1 Multi-Agent 추론 워크플로우 (LangGraph 활용)

단일 LLM의 편향성을 제거하기 위해 **LangGraph**를 이용한 상태 중심(Stateful) 에이전트 구조를 설계하였습니다.

- Debate Node:** 뉴스 컨텍스트를 바탕으로 '상승론자'와 '하락론자' 페르소나가 각각 독립적인 분석 수행.
- Judge Node (CIO):** 두 의견을 종합하여 최종 합의문(Consensus)을 도출하고 시장 트렌드 및 리스크 점수 산출.
- Evaluator Node (Feedback Loop):** 생성된 결론이 질문의 요지 및 원본 데이터와 일치하는지 검증. 점수 미달 시(8점 미만) 다시 토론 단계로 회귀하여 논리 보완(Self-Refinement).

2.2 하이브리드 RAG 전략

- Vector DB:** FAISS를 활용하여 **Deleveraging**, **Goldilocks** 등 과거 핵심 경제 시나리오를 인덱싱.
- Dynamic Search:** 단순 사용자 질문이 아닌, 에이전트가 도출한 '최종 합의문'을 검색 쿼리로 사용하여 현재 상황과 가장 유사한 과거 매크로 앵커(Anchor) 데이터를 추출.

2.3 Quant Engine & Optimization

- **Parameter Tuning:** RAG로 추출된 과거 통계값(Expected Return, Volatility)을 LLM이 현재 시장 상황에 맞춰 보정.
- **SLSQP Optimizer:** `scipy.optimize` 를 사용하여 각 자산의 비중 제한(0~45%) 및 총합 (1.0) 제약 조건을 만족하는 샤프 지수 기반 최적 포트폴리오 산출.

3. 주요 요구사항 구현 상세

요구사항	구현 내용
1. 데이터 전처리	<code>SCENARIO_KB</code> 를 통해 구조화된 금융 시나리오 데이터 구축 및 임베딩 처리
2. 벡터 DB 색인	<code>langchain_community.vectorstores.FAISS</code> 를 활용한 고속 벡터 검색 구현
3. 관련 문서 검색	<code>similarity_search</code> 를 통해 현재 시장 상황과 매칭되는 시나리오 메타데이터 추출
4. LLM 답변 생성	<code>local-llama</code> 모델을 활용, Structured Output(Pydantic)을 통한 정형 데이터 생성
5. 정확성 평가/개선	<code>Evaluator</code> 노드를 통한 반복 추론(Iterative Refinement) 로직 및 조건부 엣지 구현
6. REST API 구축	<code>FastAPI</code> 기반 <code>/analyze</code> (분석), <code>/metrics</code> (성능 측정) 엔드포인트 제공
7. 성능 측정	<code>PerformanceTracker</code> 클래스를 통한 Latency, Eval Score, Retry Rate 실시간 집계

4. 시스템 동작 알고리즘 (Flow Chart)

1. **Input:** 사용자의 금융 질문 (예: "현재 금리 인하 기대감이 시장에 미치는 영향은?")
2. **Debate:** 뉴스 기반 Bull/Bear 의견 대립 생성.
3. **Judgement:** CIO 에이전트가 시장 심리(Sentiment) 및 가격 곤충 분석.
4. **Evaluation:** 비판적 검토 (Pass 시 다음 단계, Fail 시 Debate 재수행).
5. **Retrieval:** 합의된 내용을 키워드로 FAISS에서 유사 시나리오 로드.
6. **Estimation:** LLM이 시나리오 통계치와 현재 지표를 결합해 μ (기대수익률), Σ (공분산) 추정.
7. **Optimization:** SLSQP 알고리즘으로 최적 자산 비중 w^* 산출.
8. **Output:** 최종 투자 뷰(Manager View) 및 포트폴리오 가이드 반환.

5. 기술 스택 (Tech Stack)

- **Language:** Python 3.10+
- **Frameworks:** LangChain, LangGraph, FastAPI
- **LLM:** Local Llama (via OpenAI-compatible API)
- **Database:** FAISS (Vector DB)
- **Math/Stats:** NumPy, SciPy (SLSQP), Pydantic
- **DevOps:** Unicorn, Python-dotenv

6. 성능 측정 및 개선 결과

시스템 내부의 `PerformanceTracker` 를 통해 다음과 같은 지표를 관리합니다:

- **평균 응답 시간(Latency):** Multi-agent 토론 과정으로 인해 단일 호출보다 시간은 소요되나, 논리적 완성도 확보.
- **재시도율(Retry Rate):** 평가 노드를 통해 초기 논리 모순을 발견하고 수정함으로써 최종 답변의 신뢰도(Reliability) 향상.
- **정확도(Eval Score):** 금융 전문가 페르소나를 통한 자가 채점 시스템으로 지속적인 프롬프트 개선 가능.

[부록] 실행 방법

1. `.env` 파일에 필요한 환경 변수 설정.
2. 로컬 LLM 서버(Port: 8090) 실행 확인.
3. `python main.py` 실행 후 `http://localhost:8088/docs` 접속하여 API 테스트 수행.

[2] e.g.

```
.\llama-server -m "V:\PythonProject\hf_cache_gguf\Llama-3.2-3B-Instruct-Q4_K_M.gguf" --port 8090 --host 0.0.0.0 -nlg 99 -c 8196 -fa auto --embedding --pooling mean
```

[1] e.g.

```
# .env 파일 내용
LS_ACCESS_TOKEN=...
DUMMY_TOKEN=...
HF_LOGIN_TOKEN=...
# LS 증권 API 정보
APP_KEY = "..." # 발급받은 APP Key
APP_SECRET = "..." # 발급받은 APP Secret
OPEN_AI_KEY=sk-proj-...
DB_USER=admin
DB_PASSWORD=...
DB_NAME=LLM
MARKET_IV=32.50
TOTAL_CAPITAL=60_000_000
MINI_FUTURE_FOCODE=A0562000
MINI_FUTURE_INIT_MARGIN=4000000

# [Email 설정 (Gmail 예시)]
# 구글 계정 설정 -> 보안 -> 앱 비밀번호 생성 필요
EMAIL_SENDER=...
EMAIL_PASSWORD="..."
EMAIL_RECEIVER_A=...
EMAIL_RECEIVER_B=...
```

결과:

default

POST /analyze Analyze Market

Parameters

No parameters

Request body required

application/json

Edit Value | Schema

```
{
  "question": "현재 개별 위시가 연준 의장으로 지명된 상황이 한국 음선 시장에 미치는 영향은?"
}
```

Execute Clear

Responses

Curl

```
curl -X 'POST' \
  'http://127.0.0.1:8088/analyze' \
  -H 'Content-Type: application/json' \
  -H 'Accept: application/json' \
  -d '{
    "question": "현재 개별 위시가 연준 의장으로 지명된 상황이 한국 음선 시장에 미치는 영향은?"
  }'
```

Request URL

<http://127.0.0.1:8088/analyze>

Server response

Code Details

200

Code	Details
200	<p>Response body</p> <pre>{ "status": "success", "evaluation": { "final_score": 0, "total_attempts": 2, "critique": "결론은 일본 질문에 대한 답변을 제공하지만, 몇 가지 계산점이 필요합니다. 1) 적절적인 질문 답변 부분이 부족합니다. 일본 질문은 개별 위시의 지명이 한국 음선 시장에 어떤 영향을 미치는지에 대한 적절한 답변을 요구합니다. 2) 최신 시장 분분야 부족입니다. 결론은 2026년 2월 3일의 뉴스를 기반으로 하지만, 한국 음선 시장의 최근 동향이나 분석을 더 포함하면 더욱 정확해질 수 있습니다. 3) 논리적 모순은 현재는 없습니다만, Bear의 의견을 포함한 혼조적인 영향에 대한 설명이 명확해야 합니다." }, "anager_view": "➡️ [AI Debate Market View]➡️ News Sentiment: Positive➡️ Price Action: 상승➡️: [Judge Consensus]➡️: 위시가 연준 의장으로 지명된 상황은 한국 음선 시장에 복합적인 영향을 대칠 수 있습니다. 그러나, 경승론자와 시각에서 분석하면, 위시 후보의 지명이 한국 음선 시장에 긍정적인 영향을 미칠 가능성이 높습니다. 그러나, Bear의 의견을 고려할 때, 혼조적인 영향이 있을 수 있으며, 이는 금리 정책, 원통 분동장, 불확실성 증가 등 다양한 요인으로부터 유발될 수 있습니다. 따라서, 한국 음선 시장에서는 이러한 혼조적인 영향을 고려하여 전략을 조정하거나, 다양한 시나리오를 고려하는 대응방법을 고려할 것입니다. 그러나, Bear의 의견을 고려할 때, 혼조적인 영향이 있을 수 있으므로, 이는 금리 정책, 원통 분동장, 불확실성 증가 등 다양한 요인으로부터 유발될 수 있습니다. 따라서, 한국 음선 시장에서는 이러한 혼조적인 영향을 고려하여 전략을 조정하거나, 다양한 시나리오를 고려하여 대응할 필요가 있습니다.➡️[RAG Context]➡️[RAG Anchor: Goldilocks]➡️ 설명: 제품과 액정 강당 속 미강적인 우승팀. 낮은 분동장과 끝 음선 수익성 개선➡️ 가중: mu: [0.1674, -0.0683, 0.1393]➡️ 기준 vol: [0.1356, 0.1328,</pre> <p>Download</p> <p>Response headers</p> <pre>content-length: 2462 content-type: application/json date: Tue, 03 Feb 2026 13:30:09 GMT server: unicorn</pre> <p>Responses</p> <p>Code Description Links</p> <p>200 Successful Response No links</p> <p>Media type application/json</p> <p>Controls Accept header.</p> <p>Example Value Schema</p> <pre>"string"</pre> <p>422 Validation Error</p> <p>Media type application/json</p> <p>Example Value Schema</p> <pre>{ "detail": [{ "loc": ["string", 0] }] }</pre> <p>No links</p>

금융 뉴스 (websocket 수신 via LS 증권 Open API)

```

admin - Warning - not support... ×
File Edit View Query Database Server Tools Scripting Help
MySQL Workbench 8.0.28
Navigator: ex1 ex2 ex3
MANAGEMENT
  ● Server Status
  ● Client Connections
  ● Users and Privileges
  ● Status and System Variables
  ● Data Export
  ● Data Import/Restore
INSTANCE
  ● Startup / Shutdown
  ● Server Logs
  ● Options File
PERFORMANCE
  ● Dashboard
  ● Performance Reports
  ● Performance Schema Setup
  ● Administration
  ● Schemas
Information: No object selected
Result Grid | Filter Rows: | Edit | Export/Import: | Wrap Cell Contents: | Fetch rows: | Result Grid | Form Editor | Field Types | Query Stats | Execution Plan |
news_data 2 ×
Output
Action Output
# Time Action Message Duration / Fetch
1 21:33:26 SELECT * FROM news_data WHERE category LIKE '%가시경제%' ORDER BY date DESC, time DESC ... 200 rows(s) returned 0.219 sec / 0.000 sec
2 00:31:47 SELECT * FROM news_data WHERE category LIKE '%가시경제%' ORDER BY date DESC, time DESC ... 200 rows(s) returned 0.016 sec / 0.000 sec
3 00:32:05 SELECT * FROM macro_scenarios ORDER BY created_at DESC LIMIT 0, 1000 100 row(s) returned 0.000 sec / 0.000 sec
  
```

The screenshot shows the MySQL Workbench interface. The left sidebar contains navigation links for Management, Instance, Performance, Administration, Schemas, and Information. The main area has three tabs: 'ex1', 'ex2', and 'ex3'. Tab 'ex2' is active and displays a SQL query:

```

49
50 이는 미국과 이런 간접 완화로 인해 발생한 것으로, 투합포 대중들의 대화 기록 언급이 주요 원인이다.
51
52 이로 인해 평화 국제 유가의 변동성이 주목된다.
53
54 */
55
56
57 ● SELECT * FROM news_data LIMIT 20 ;
58
59
60 ● SELECT date, time, category, title, body , realkey
61   FROM news_data
62   WHERE category LIKE '%가시경제%'
63   ORDER BY date DESC, time DESC LIMIT 120 ;
64
65 /*
66
67 */
68
69 ● SELECT date, time, category, title, body , realkey
  
```

Below the query, the 'Result Grid' tab is selected, showing the results of the executed queries. The results for the first query show 200 rows returned in 0.219 seconds. The second query also shows 200 rows returned in 0.016 seconds. The third query shows 100 rows returned in 0.000 seconds.

admin - Warning - not support... x

File Edit View Query Database Server Tools Scripting Help

Navigator ex1 ex2 ex

MANAGEMENT

- Server Status
- Client Connections
- Users and Privileges
- Status and System Variables
- Data Export
- Data Import/Restore

INSTANCE

- Startup / Shutdown
- Server Logs
- Options File

PERFORMANCE

- Dashboard
- Performance Reports
- Performance Schema Setup

```

35     key_factors TEXT NOT NULL,          -- 한단 균거
36     strategy TEXT NOT NULL,           -- 주문 투자 전략
37     created_at DATETIME DEFAULT CURRENT_TIMESTAMP
38   );
39
40
41
42
43 • CREATE TABLE IF NOT EXISTS macro_scenarios (
44   id INT AUTO_INCREMENT PRIMARY KEY,
45   scenario_name VARCHAR(255) NOT NULL,
46   market_description TEXT NOT NULL,    -- LLM이 검색할 때 사용할 상세 설명
47   mu JSON NOT NULL,                  -- [C_L, C_S, P_L, P_S, FUT]
48   vol JSON NOT NULL,                -- [v1, v2, v3, v4, v5]
49   corr JSON NOT NULL,               -- 5x5 행렬
50   tags VARCHAR(255),
51   created_at TIMESTAMP DEFAULT CURRENT_TIMESTAMP
52 );
53
54 • SELECT * FROM macro_scenarios
55 ORDER BY created_at DESC ;

```

Result Grid | Filter Rows: | Edit: | Export/Import: | Wrap Cell Contents: |

	scenario_name	market_description	mu	vol	corr
1	OPTQ_EXTREME_BEAR_0000	Dileveraging: 마진콜 및 부채 축소로 인한 길게 매도 장세	[{"-0.3385, 0.1038, 0.5128, -0.1338, -0.2460"}]	[{"0.3599, 0.34, 0.4803, 0.4954, 0.4017"}]	[{"1.0, 0.7718548442224368, -0.1676, 0.0683, -0.2374, 0.0894, 0.1393"}]
2	OPTQ_BULLISH_0001	Goldilocks: 기울기-액정 설정 속 이산화적인 우상향	[{"0.1356, 0.1328, 0.1848, 0.1731, 0.1207"}]	[{"0.3465, 0.3494, 0.5189, 0.4907, 0.4083"}]	[{"1.0, 0.7992883004977631, -0.1358, 0.1106, 0.483, -0.1448, 0.2394"}]
3	OPTQ_BEARISH_0002	Emerging Market Crisis: 신종コ 채권 유동화로 인한 언센스 하락	[{"0.0543, -0.0338, 0.0519, -0.0579, 0.0094"}]	[{"0.4585, 0.4605, 0.456, 0.4476, 0.3444"}]	[{"1.0, 0.9999999999999998, 0.8741, 0.0352, -0.0498, 0.0557, -0.0465, 0.0052"}]
4	OPTQ_VOLATILE_0003	Flash Crash: 알고리즘 매도 폴리주로 인한 단기 급락	[{"0.0543, -0.0338, 0.0519, -0.0579, 0.0094"}]	[{"0.4585, 0.4605, 0.456, 0.4476, 0.3444"}]	[{"1.0, 0.9999999999999998, 0.8741, 0.0352, -0.0498, 0.0557, -0.0465, 0.0052"}]
5	OPTQ_VOLATILE_0004	CPI Volatility: 물가 지표 발표 전후의 국소화된 밸류션 탐색	[{"0.0543, -0.0338, 0.0519, -0.0579, 0.0094"}]	[{"0.458, 0.4464, 0.4533, 0.4473, 0.3516"}]	[{"1.0, 0.9239239552692844, 0.458, 0.4464, 0.4533, 0.4473, 0.3516"}]
6	OPTQ_VOLATILE_0005	Gamma Squeeze: 옵션 페더 차질으로 인한 변동성 증폭	[{"0.0382, -0.041, 0.0383, -0.0507, 0.0107"}]	[{"0.4494, 0.4615, 0.4448, 0.4443, 0.3645"}]	[{"1.0, 0.8624417197054527, 0.0382, -0.041, 0.0383, -0.0507, 0.0107"}]
7	OPTQ_STRENG_BULL_0006	At Super Cycle: 빅마크 주도 성장으로 인한 급락한 상승세	[{"0.1728, -0.0389, -0.2352, 0.0958, 0.1014"}]	[{"0.161, 0.1581, 0.1602, 0.1714, 0.1078"}]	[{"1.0, 0.9999999999999998, 0.7938, 0.1728, -0.0389, -0.2352, 0.0958, 0.1014"}]
8	OPTQ_VOLATILE_0007	Dead Cat Bounce: 하락 추세 중 절시적 기술적 반등	[{"0.0684, -0.0369, 0.0569, -0.069, 0.015"}]	[{"0.447, 0.4577, 0.455, 0.4305, 0.3462"}]	[{"1.0, 0.0000000000000002, 0.8675, 0.0684, -0.0369, 0.0569, -0.069, 0.015"}]
9	OPTQ_BLACK_SWAN_0008	Systemic Risk: 금융 시스템 붕괴 위험 및 전자신 투매	[{"0.3425, 0.1112, 0.5008, -0.1522, -0.2608"}]	[{"0.3575, 0.3509, 0.4979, 0.494, 0.4293"}]	[{"1.0, 0.777371646976953, 0.1874, -0.0307, -0.2355, 0.0882, 0.1249"}]
10	OPTQ_BULLISH_0009	V-shaped Recovery: 금락 후 원모멘텀 기선에 따른 빠른 회복	[{"0.1874, -0.0307, -0.2355, 0.0882, 0.1249"}]	[{"0.1462, 0.1453, 0.1926, 0.1964, 0.1017"}]	[{"1.0, 0.75249723148665, 0.1874, -0.0307, -0.2355, 0.0882, 0.1249"}]
11	OPTQ_VOLATILE_0010	Gamma Squeeze: 옵션 페더 차질으로 인한 변동성 증폭	[{"0.0408, -0.0492, 0.0671, -0.0303, 0.0167"}]	[{"0.4513, 0.4547, 0.456, 0.4612, 0.345"}]	[{"1.0, 0.9999999999999998, 0.888, 0.0408, -0.0492, 0.0671, -0.0303, 0.0167"}]
12	OPTQ_EXTREME_BULL_0011	Melt-up: 과정된 투심으로 인한 비이상적 폭등	[{"0.1896, -0.0353, -0.2545, 0.0752, 0.1296"}]	[{"0.1454, 0.1334, 0.1604, 0.1891, 0.1181"}]	[{"1.0, 0.7801812632198319, 0.1896, -0.0353, -0.2545, 0.0752, 0.1296"}]
13	OPTQ_NEUTRAL_0012	Liquidity Trap: 물가도 시장에 들지 않는 침체상 환보	[{"-0.0005, 0.0348, 0.0194, 0.043, 0.0278"}]	[{"0.1015, 0.1308, 0.1018, 0.1282, 0.0755"}]	[{"1.0, 0.9433121089492541, 0.1015, 0.0348, 0.0194, 0.043, 0.0278"}]
14	OPTQ_NEUTRAL_0013	Liquidity Trap: 물가도 시장에 들지 않는 침체상 환보	[{"-0.0029, 0.0684, -0.0069, 0.0441, 0.0003"}]	[{"0.1266, 0.1199, 0.1023, 0.1198, 0.0627"}]	[{"1.0, 0.908278896722807, 0.1266, 0.1199, 0.1023, 0.1198, 0.0627"}]
15	OPTQ_BULLISH_0014	V-shaped Recovery: 금락 후 원모멘텀 기선에 따른 빠른 회복	[{"0.1857, -0.0316, -0.2451, 0.0606, 0.1351"}]	[{"0.1572, 0.1693, 0.1932, 0.1641, 0.1062"}]	[{"1.0, 0.799874271174398, 0.1857, -0.0316, -0.2451, 0.0606, 0.1351"}]
16	OPTQ_BEARISH_0015	Safe Haven Flight: 안전자산 선호로 인한 지수 정체 및 재원 강세	[{"-0.3392, 0.1019, 0.5085, -0.145, -0.2438"}]	[{"0.3411, 0.3512, 0.498, 0.5178, 0.3835"}]	[{"1.0, 0.9999999999999998, 0.7681, -0.3392, 0.1019, 0.5085, -0.145, -0.2438"}]
17	OPTQ_VOLATILE_0016	Flash Crash: 알고리즘 매도 폴리주로 인한 대규모	[{"0.0225, 0.0507, 0.0303, 0.0482, 0.0015"}]	[{"0.4576, 0.4576, 0.4576, 0.4576, 0.4576"}]	[{"1.0, 0.9999999999999998, 0.7681, 0.0225, 0.0507, 0.0303, 0.0482, 0.0015"}]

macro_scenarios 2 x

Output

Action Output

#	Time	Action	Message	Duration / Fetch
2	00:31:47	SELECT * FROM news_data WHERE category LIKE '%거시경제%' ORDER BY date DESC, time DESC ..	200 row(s) returned	0.016 sec / 0.000 sec
3	00:32:05	SELECT * FROM macro_scenarios ORDER BY created_at DESC LIMIT 0, 1000	100 row(s) returned	0.000 sec / 0.000 sec
4	00:32:16	SELECT date, time, category, .. realkey, title, body FROM news_data WHERE LENGTH(id) > 10 ...	54 row(s) returned	0.094 sec / 0.000 sec