

[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** 시스템입니다. 단순 검색(Retrieval)을 넘어, **상승론(Bull)과 하락론(Bear)의 논리적 경합(Debate)**을 통해 시장의 괴리(Divergence)를 분석하고, 그 결과를 바탕으로 퀀트 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 조정.
- **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 엡지 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 괴리 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:** Uvicorn, 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 -ngl 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 상황에 한국 옵션 시장에 미치는 영향은?"
}
```

Execute

Clear

Responses

Curl

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

Request URL

```
http://127.0.0.1:8088/analyze
```

Server response

Code Details

200

Code Details

200

Response body

```
{
  "status": "success",
  "evaluation": {
    "final_score": 0,
    "total_attempts": 1,
    "critique": "생성된 결과는 원본 질문에 대한 답변을 제공하지만, 몇 가지 개선점이 필요합니다. 1) 직관적인 판단 답변 부분에 부족합니다. 원본 질문은 개별 위시가 지명에 한국 옵션 시장에 대한 영향을 미치는지에 대한 직접적인 답변을 요구합니다. 2) 최신 자료 반영 부분에 부족합니다. 생성된 결과는 2026년 2월 3일의 뉴스를 기반으로 하지만, 한국 옵션 시장의 최근 동향이나 분석을 더 포함할 수 있습니다. 3) 논리적 모순성은 현재는 없습니다만, Bear의 의견을 포함한 혼조적인 영향에 대한 설명이 명확해야 합니다.",
    "manager_view": " [AT Debate Market View]\n- News Sentiment: Positive\n- Price Action: 상승\n\n:: [Judge Consensus]\n개별 위시가 연준 의장으로 지명된 상황은 한국 옵션 시장에 복합적인 영향을 미칠 수 있습니다. 그러나, 상승론자의 시각에서 분석하면, 위시 후보의 지명에 한국 옵션 시장에 긍정적인 영향을 미칠 가능성이 높습니다. 그러나, Bear의 의견을 고려할 때, 혼조적인 영향에 있을 수 있으며, 이는 금리 정책, 원금 변동성, 불확실성 증가 등 다양한 요인에 따라 다르게 반영될 수 있습니다. 따라서, 한국 옵션 시장에서는 이러한 혼조적인 영향을 고려하여 전략을 조정하거나, 다양한 시나리오를 고려하여 대응할 필요가 있습니다.\n\n[Divergence]\n- 현재 가격은 상승 중입니다. 그러나, Bear의 의견을 고려할 때, 혼조적인 영향에 있을 수 있으며, 이는 금리 정책, 원금 변동성, 불확실성 증가 등 다양한 요인에 따라 다르게 반영될 수 있습니다. 따라서, 한국 옵션 시장에서는 이러한 혼조적인 영향을 고려하여 전략을 조정하거나, 다양한 시나리오를 고려하여 대응할 필요가 있습니다.\n\n[RAG Anchor: Goldilocks]\n- 설명: 저평가-적정 성장 속 예상적인 우상향. 낮은 변동성 및 골 옵션 수익성 개선.\n- 기준 mu: [0.1676, -0.0683, -0.2374, 0.0894, 0.1393]\n- 기준 vol: [0.1356, 0.1328, 0.1849, 0.1773, 0.1207]\n",
    "quant_params": {
      "expected_returns": [
        0.01,
        0.01,
        0.01,
        0.01,
        0.01,
        0.01
      ],
      "risk_score": 3
    },
    "portfolio": {
      "weights": [

```

Response headers

```
content-length: 2462
content-type: application/json
date: Tue, 03 Feb 2026 13:38:09 GMT
server: uvicorn
```

Responses

Code Description

200 Successful Response

Media type

application/json

Controls Accept header.

Example Value Schema

```
"string"
```

422 Validation Error

Media type

application/json

Example Value Schema

```
{
  "detail": [
    {
      "loc": [
        "string",
        0

```

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

5

금융 뉴스 (websocket 수신 via LS 증권 Open API)

The screenshot displays a database management interface with a SQL query editor and a results grid. The query is as follows:

```

SELECT * FROM news_data LIMIT 20 ;

SELECT date, time, category, title, body , realkey
FROM news_data
WHERE category LIKE '저시경제%'
ORDER BY date DESC, time DESC LIMIT 120 ;

SELECT date, time, category, title, body , realkey

```

The results grid shows the following data:

date	time	category	title	body	realkey
2026-02-03	20:41:03	저시경제_KEY	"워시, 2번 아닌 4~5번 글리 내릴 것"	워시 연준 의장 로보가 글리 인화를 적극적으로 추진할 수 있다는 관측이 나오고 ...	20260203204103358FQ37L38
2026-02-03	20:06:58	저시경제_LLM	트레이더들이 영한은행 결정을 주시하는 가운데 소탈형은 ...	소탈형 확률은 영한은행의 글리 결정을 앞두고 안정세를 보이며, 올해 들어 1.4...	202602032006582800001513
2026-02-03	16:35:21	저시경제_KEY	[한글마당] 한·일 금반기, 워시 최종형+주가 폭등+호주 금...	한글마당은 1445.4원으로 마감했으며, 금 연휴 전 1430원~1460원 사이에서 ...	20260203163521120300394
2026-02-03	16:34:41	저시경제_KEY	첫연두, 일본은 지난 1945 부안기동 통행 저점...10억곳 구...	첫연두는 일본은 시장가치를 위한 JAR 불투명한 하이퍼어를 공급하며, 이는 물신...	202602031634411144401042
2026-02-03	16:29:12	저시경제_KEY	대이원합리니, 독자 전환 성공...수익성 중심 제철 전환 가시화	대이원합리니는 2024년 사상 최대 영업이익 46억 원을 기록하며 독자 전환에 성...	20260203162912120300390
2026-02-03	16:27:49	저시경제_KEY	동부건설, 지난해 영업이익 606억원... 독자 전환	동부건설은 지난해 606억원의 영업이익을 기록하며 독자 전환에 성공했으며, 신...	202602031627492320303159
2026-02-03	16:27:11	저시경제_KEY	형지IC는 작년 연간 연결 영업실적 70억... 독자 전환	형지IC는 작년 연간 연결기준으로 70억원의 영업실적을 기록, 전년 대비 독자 지...	202602031627113300014116
2026-02-03	16:27:00	저시경제_KEY	포탈신라, 지난해 영업이익 135억원... 독자 전환 성공	포탈신라는 지난해 연간 매출 4조683억원으로 전년 대비 3.1% 증가, 영업이익 1...	202602031627002300067752
2026-02-03	16:25:47	저시경제_KEY	"강제장 면 7500도 문제 없어"...포토전 코스피 목표지 '향' 올...	포토전 코스피 목표지를 6000으로 올렸으며, 강제장에서는 7500까지 상승할 ...	202602031625472511951557
2026-02-03	16:17:26	저시경제_KEY	RE 100 일 불렀다...저항한 구역전기 써드'재정여지' 지구...	장부가 산입단지 입주 기업들이 저항한 전기용 공급받더라도 재정여지 PPA를 ...	20260203161726260000889
2026-02-03	16:16:25	저시경제_KEY	SK하이닉스, 지난해 영업실적 1235억원... 독자 지속	SK하이닉스는 지난해 매출이 6314억원으로 전년 대비 143.3% 증가했으며...	20260203161625220303109
2026-02-03	16:12:33	저시경제_KEY	아스코 '리라'를 송파구, 삼성대문로 독자 전환'기' 점다	삼성전자는 대리점 구상으로 인해 2주 동안의 수익을 얻었지만, 대리점의 독...	202602031612331423301011
2026-02-03	16:11:43	저시경제_KEY	포탈신라, 작년 영업이익 135억 독자 전환... '내실경영 주력'	포탈신라는 지난해 연결 기준으로 135억원의 영업이익을 기록해 독자 전환했으...	202602031611432710668820
2026-02-03	16:10:45	저시경제_KEY	포탈신라, 작년 영업이익 135억원으로 독자... '내실경영 주력'	포탈신라는 작년 연결 기준으로 135억원의 영업이익을 기록, 전년 대비 독자 전...	202602031610451111299515
2026-02-03	16:09:04	저시경제_KEY	솔트웨이, 작년 연간 연결 영업이익 5억... 독자 전환	솔트웨이는 작년 연간 연결기준으로 5억66616만원의 영업이익을 기록해 독자전...	202602031609043300013987
2026-02-03	16:08:33	저시경제_KEY	"잠실까지 20분대"...구리역 하이네리 리버파크 이달 분양	DL이앤씨, GS건설, SK에프를 통해 컨소시엄이 경기 구리시 수택동에 3022가구...	20260203160833200383544
2026-02-03	16:07:16	저시경제_KEY	유화·노도구 고기류 영안지 포탈스... '인더스트리'로 고기류	포탈스의 '인더스트리'로 고기류 영안지 유화·노도구 고기류 영안지 2024년 ...	2026020316071611110000

The bottom section shows the execution output of the queries:

#	Time	Action	Message	Duration / Fetch
1	21:33:26	SELECT * FROM news_data WHERE category LIKE '저시경제%' ORDER BY date DESC, time DESC ...	200 row(s) returned	0.219 sec / 0.000 sec
2	00:31:47	SELECT * FROM news_data WHERE category LIKE '저시경제%' 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

admin - Warning - not support...

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,
36 strategy TEXT NOT NULL,
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,
47 mu JSON NOT NULL,
48 vol JSON NOT NULL,
49 corr JSON NOT NULL,
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

id	scenario_name	market_description	mu	vol	corr
1	OPTQ_EXTREME_BEAR_0000	Deleveraging: 파산론 및 부채 축소로 인한 강제 매도 장세	[-0.3385, 0.1038, 0.5128, -0.1338, -0.2469]	[0.3599, 0.34, 0.4803, 0.4954, 0.4017]	[[1.0, 0.771854844224368, -0
2	OPTQ_BULLISH_0001	Goldilocks: 적절가치형 성장 속 이상적인 주상장	[0.1676, -0.0683, -0.2374, 0.0894, 0.1393]	[0.1356, 0.1328, 0.1848, 0.1731, 0.1207]	[[1.0, 0.7992983004977631, -0
3	OPTQ_BEARISH_0002	Emerging Market Crisis: 신흥국 자금 유출로 인한 급락 하락	[-0.3558, 0.1106, 0.483, -0.1448, -0.2394]	[0.3465, 0.3464, 0.5189, 0.4907, 0.4208]	[[1.0000000000000002, 0.8066
4	OPTQ_VOLATILE_0003	Flash Crash: 알고리즘 매도 폭주에 의한 단기 급락	[0.0543, -0.0338, 0.0519, -0.0579, 0.0094]	[0.4385, 0.4605, 0.456, 0.4416, 0.3444]	[0.9999999999999996, 0.8741
5	OPTQ_VOLATILE_0004	CPI Volatility: 물가 지표 발표 전후의 극심한 방향성 함석	[0.0352, -0.0488, 0.0557, -0.0465, 0.0052]	[0.458, 0.4464, 0.4533, 0.4473, 0.3516]	[[1.0, 0.9239239553292844, -0
6	OPTQ_VOLATILE_0005	Gamma Squeeze: 옵션 밀려 파장으로 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, 0.8624417197085427, -0
7	OPTQ_STRONG_BULL_0006	AI Super Cycle: 빅테크 주도 실적 기반 강력한 상승세	[0.1728, -0.0389, -0.2352, 0.0958, 0.1014]	[0.161, 0.1581, 0.1602, 0.1714, 0.1078]	[[0.9999999999999998, 0.7938
8	OPTQ_VOLATILE_0007	Dead Cat Bounce: 하락 추세 중 일시적 기술적 반등	[0.0684, -0.0369, 0.0569, -0.069, -0.015]	[0.447, 0.4577, 0.455, 0.4305, 0.3462]	[[1.0000000000000002, 0.8671
9	OPTQ_BLACK_SWAN_0008	Systemic Risk: 금융 시스템 붕괴 우려 및 전 자산 추락	[-0.3425, 0.1112, 0.5008, -0.1522, -0.2608]	[0.3575, 0.3509, 0.4979, 0.494, 0.4029]	[[1.0, 0.7773371646976963, -0
10	OPTQ_BULLISH_0009	V-shaped Recovery: 급락 후 펀더멘탈 개선에 따른 빠른 회복	[0.1874, -0.0307, -0.2355, 0.0882, 0.1245]	[0.1462, 0.1453, 0.1926, 0.1964, 0.1017]	[[1.0, 0.7928497293148665, -0
11	OPTQ_VOLATILE_0010	Gamma Squeeze: 옵션 밀려 파장으로 인한 변동성 증폭	[0.0408, -0.0492, 0.0671, -0.0303, 0.0167]	[0.4513, 0.4547, 0.456, 0.4612, 0.345]	[0.9999999999999998, 0.8096
12	OPTQ_EXTREME_BULL_0011	Melt-up: 과열된 투심으로 인한 비이성적 폭등	[0.1896, -0.0359, -0.2543, 0.0752, 0.1296]	[0.1454, 0.1334, 0.1604, 0.1891, 0.1183]	[[1.0, 0.7801812632198319, -0
13	OPTQ_NEUTRAL_0012	Liquidity Trap: 돈을 풀어도 시중이 돌지 않는 정체성 정보	[-0.0005, 0.0348, 0.0194, 0.043, 0.0278]	[0.1015, 0.1308, 0.1018, 0.1282, 0.0753]	[[1.0, 0.9433121089482541, -0
14	OPTQ_NEUTRAL_0013	Liquidity Trap: 돈을 풀어도 시중이 돌지 않는 정체성 정보	[-0.0029, 0.0684, -0.0069, 0.0441, 0.0003]	[0.1266, 0.1199, 0.1023, 0.1198, 0.0627]	[[1.0, 0.9082758996772807, 0
15	OPTQ_BULLISH_0014	V-shaped Recovery: 급락 후 펀더멘탈 개선에 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, 0.7998742711774398, -0
16	OPTQ_BEARISH_0015	Safe Haven Flight: 안전자산 선호로 인한 지수 정체 및 차권 강세	[-0.3392, 0.1019, 0.5085, -0.145, -0.2438]	[0.3411, 0.3512, 0.498, 0.5178, 0.3835]	[0.9999999999999998, 0.7681
17	OPTQ_VOLATILE_0016	Flash Crash: 알고리즘 매도 폭주에 의한 단기 급락	[0.0376, -0.0507, 0.0303, -0.0587, 0.0118]	[0.4536, 0.4579, 0.453, 0.4467, 0.3634]	[[1.0000000000000002, 0.8046

macro_scenarios 2 x

Output

Action Output

#	Time	Action	Message	Duration / Fetch
2	00:31:47	SELECT * FROM news_data WHERE category LIKE '가시경제%' 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