RiskRadar: Multi-Agent LLM System for Early Warning Detection

Group 9:
Adeyinka Abdulrahman
Jessica A. Abreu
Alex Hamilton
Muhammad Latif
Rajen Lavingia
Arkadiusz Tomczak

The Bank of England Challenge

The Regulatory Burden

Current State:

- 200 UK banks × 4 quarterly reviews = 800 annual assessments
- 300-500 page financial documents per bank
- 40-60 hours manual analysis per document
- £2.4M-£3.6M annual analyst cost

The Problem:

- Weeks-long lag between report publication and response
- Manual analysis may miss subtle warning signals
- Inconsistent assessment across analysts (15-25% variability)
- Cannot scale to handle increasing regulatory complexity

Business Need:

✓ Faster, more consistent, fully auditable risk assessments

What Makes Manual Analysis So Difficult?

SCALE & SPEED
📚 442-page documents
⏱️ 40-60 hours per report
🔁 800 reviews annually

QUALITY & CONSISTENCY
🔍 Subtle risks across sections
⚖️ Analyst variability 15-25%
📊 Sampling vs full coverage

REGULATORY REQUIREMENT
✓ CAMELS framework
✓ Complete audit trail
✓ Cited evidence
✓ Reproducibility

RISK CONSEQUENCES
⚠️ False negatives → missed crises
⚠️ False positives → wasted resources
⚠️ Delayed action → materialised risk

🎯 We needed a system that could read like an expert, think like a regulator, and document like an auditor.

🎯 RiskRadar

AI-Powered Risk Intelligence

What It Is:

An automated financial risk assessment system powered by 16 specialized AI modules that analyse regulatory documents to produce comprehensive CAMELS risk ratings with full citation tracking.

Core Innovation:

Instead of training 16 separate models, we use prompt engineering to transform a single LLM into 16 domain experts through carefully crafted instructions.

Key Capabilities:

- Complete document coverage (100% of pages analyzed)
- 16-dimensional risk assessment (CAMELS + linguistic analysis)
- Every finding cited to specific page/section
- Structured JSON output for regulatory systems
- 40-60 minutes automated analysis (vs 40-60 hours manual)

Output: Traffic-light risk signals 🟢 🟡 🔴 with actionable recommendations

🎯 RiskRadar

Key Results:

METRIC	MANUAL ANALYSIS	RISKRADAR	IMPROVEMENT
Time per Document	40-60 hours	50 minutes	**48-72x faster**
Cost per Document	£1,500	£3	**99.8% reduction**
Document Coverage	Selective sampling	100% (all pages)	**Complete coverage**
Annual Time Savings			

RiskRadar Implementation

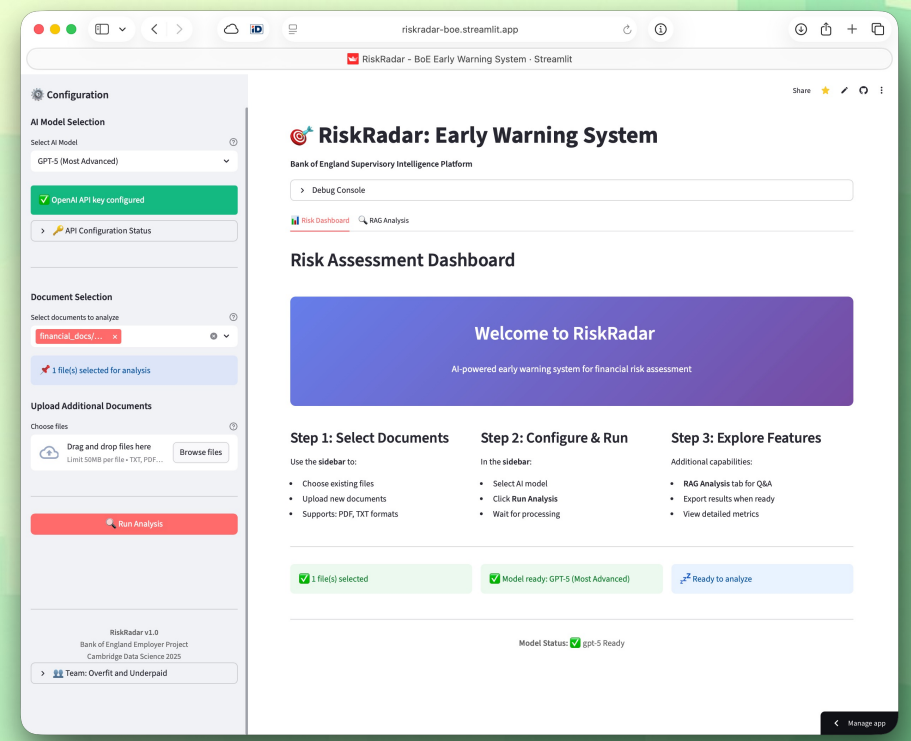
Multi-Analysis

+

RAG

+

Multi-Analysis + RAG



Specialized AI Modules for Comprehensive Analysis

Tier 1: Linguistic & Behavioral (4 modules)

- 😊 Sentiment Tracker → Defensive language, tone shifts
- 📋 Topic Analyzer → Narrative changes, omissions
- 🎯 Confidence Evaluator → Management evasiveness, credibility
- ❓ Analyst Concern Detector → Implicit worries from Q&A

Tier 2: Quantitative Risk Metrics (9 modules)

- 💰 Capital Buffers
- 🏦 Credit Quality
- 💧 Liquidity & Funding
- 📈 Earnings Quality
- 📊 Market & Interest Risk
- ⚖️ Governance & Controls
- ⚖️ Legal & Regulatory
- 🏢 Business Model
- 📄 Off-Balance Sheet

Tier 3: Pattern Recognition (2 modules)

- 🚩 Red Flag Detector → Critical warnings (going concern, breaches)
- 🔍 Discrepancy Auditor → Cross-validation across agents

Tier 4: Risk Synthesis (1 module)

- 🎯 CAMELS Fuser → Final comprehensive rating (0-10 scale)

Why 16 Agents?

Cognitive Decomposition Strategy

Tier 1: Linguistic Analysis (4 agents)

Purpose: Capture subjective risk signals humans detect

- **sentiment_tracker** → Detects defensive language, hedging
- **topic_analyzer** → Identifies narrative shifts, omissions
- **confidence_evaluator** → Assesses management evasiveness
- **analyst_concern** → Extracts implicit worries from Q&A

Why separate? Different linguistic dimensions require different prompts

Tier 2: Quantitative Metrics (9 agents)

Purpose: Extract structured financial data

- **capital_buffers** → CET1, Tier 1, leverage ratios
- **liquidity_funding** → LCR, NSFR, funding mix
- **market_irrbb** → Interest rate risk, unrealized losses
- **credit_quality** → NPL, Stage 2/3, ECL coverage
- **earnings_quality** → ROE, ROA, NIM, one-off items
- **governance_controls** → Material weaknesses, audit opinions
- **legal_reg** → Enforcement actions, litigation
- **business_model** → Revenue concentration, growth anomalies
- **off_balance_sheet** → Commitments, derivatives exposure

Why separate? Each agent optimized for specific regulatory domain

Tier 3: Pattern Recognition (2 agents)

- **red_flags** → Detects critical warning phrases
- **discrepancy_auditor** → Cross-validates agent consistency

Why separate? Different pattern matching vs validation logic

Tier 4: Synthesis (1 agent)

- **camels_fuser** → Aggregates all evidence into final CAMELS

Why separate? Final synthesis requires holistic view

5-Step Analysis Pipeline

Step 1: Document Ingestion

PDF → Text extraction → Intelligent chunking
Credit Suisse 2019: 442 pages, 1.87M characters → 3 chunks

Step 2: Parallel Analysis

14 agents analyze each chunk simultaneously
Runtime per chunk: ~15 minutes
7x faster than sequential execution

Step 3: Cross-Chunk Aggregation

Results merged using strategy-specific algorithms:

- Linguistic agents: Average scores + merge findings
- Quantitative agents: Coalesce values + maximum risk score
- Pattern agents: Merge all + deduplicate

Step 4: Meta-Analysis

Discrepancy Auditor → Validates consistency
CAMELS Fuser → Synthesizes final assessment

Step 5: Regulatory Output

Structured report with risk scores, traffic lights, citations, and actionable recommendations

Single Model, Multiple Experts

Traditional Approach

- ❌ Train 16 separate models
- ❌ Months of model development
- ❌ Requires labeled training data
- ❌ Hard to update for new rules

RiskRadar Approach

- ✅ One LLM + 16 specialized prompts
- ✅ Days of prompt engineering
- ✅ No training data needed
- ✅ Update prompt, instant deployment

Executing 14 Agents Simultaneously

Sequential Approach

14 agents × 10 minutes
= 140 minutes per chunk

Parallel Approach

14 agents / 2 workers
= ~15 minutes per chunk

9x speedup!

How Parallel Execution Works:

1. ThreadPoolExecutor with 2 workers
2. Each worker handles one agent at a time
3. Agents are I/O-bound (waiting for API responses)
4. Python GIL released during network calls
5. Rate limiter coordinates across threads

Credit Suisse Annual Report 2019 Analysis:

- 3 chunks × 14 agents = 42 agent executions
- Chunk 1: 14 minutes | Chunk 2: 12 minutes | Chunk 3: 18 minutes
- Meta-agents: 6 minutes
- Total: 50 minutes

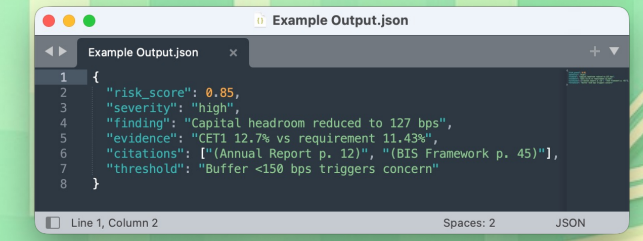
vs Manual Analysis: 40-60 hours (48x faster)

Regulatory AI Requires Complete Transparency

The Citation Requirement:

Every finding MUST reference its source:

- ✓ "CET1 ratio 12.7% (Annual Report p. 12)"
- ✓ "Material uncertainty regarding going concern (Auditor Report p. 260)"
- ✓ "Management discussed IBCM restructuring (Q&A p. 95)"



Why This Matters:

Auditability	→ Regulators can verify every claim
Accountability	→ AI decisions are traceable
Trust	→ Humans can validate AI reasoning
Reproducibility	→ Same input = same output with same evidence

"If you can't cite it, you can't trust it."

RiskRadar vs Manual Analysis

Speed:

- Manual: 40-60 hours per document
- RiskRadar: 50 minutes per document
- Speedup: 48-72x faster

Cost:

- Manual: ~£1,500 per document (£25/hour × 60 hours)
- RiskRadar: ~£3 per document (API costs)
- Cost reduction: 99.8%

Coverage:

- Manual: Selective sampling (time constraints)
- RiskRadar: 100% document coverage (all 442 pages)
- Coverage: Complete vs partial

Scalability:

- Manual: Limited by analyst availability
- RiskRadar: Linear scaling with API capacity
- Can process entire UK banking sector quarterly

Testing on Known Risk:

Test Example: Credit Suisse Annual Report 2019

Ground Truth (What We Know Now):

- 2019 report released before March 2023 collapse
- Known regulatory issues:
  - Federal Reserve CCAR conditional non-objection
  - Capital planning weaknesses required remediation
  - Material legal exposures (Mozambique, RMBS)
  - Earnings reliant on one-time gains
  - COVID-19 exposure flagged

Testing on Known Risk:

Credit Suisse Annual Report 2019

RiskRadar Assessment (December 2019 analysis):

Overall Risk: 0.680 🟡 AMBER (Medium-High Risk)

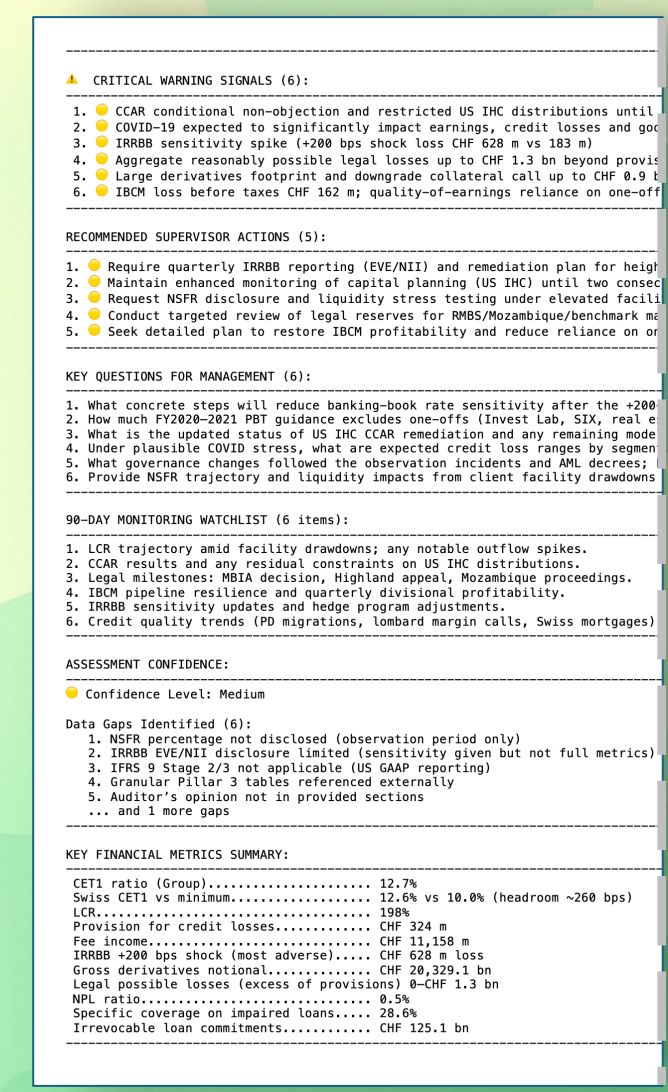
CAMELS Breakdown:

- 🟢 Capital: GREEN CET1 12.7%, adequate headroom
- 🟡 Assets: AMBER NPL 0.5% but thin coverage (28.6%)
- 🟡 Management: AMBER Fed CCAR restrictions, AML remediation
- 🟡 Earnings: AMBER ROE 7.7%, one-off reliance, IBCM losses
- 🟢 Liquidity: GREEN LCR 198%, strong buffer
- 🟡 Sensitivity: AMBER IRRBB spiked (CHF 628M vs CHF 183M)

Key Finding Correctly Identified:

"Material uncertainty regarding going concern"
detected by red flags agent (Annual Report p. 260) - 2 years before actual collapse

✅ **System correctly signaled elevated risk profile**



```
⚠️  CRITICAL WARNING SIGNALS (6):
1. 🟡 CCAR conditional non-objection and restricted US IHC distributions until
2. 🟡 COVID-19 expected to significantly impact earnings, credit losses and goo
3. 🟡 IRRBB sensitivity spike (+200 bps shock loss CHF 628 m vs 183 m)
4. 🟡 Aggregate reasonably possible legal losses up to CHF 1.3 bn beyond provis
5. 🟡 Large derivatives footprint and downgrade collateral call up to CHF 0.9 b
6. 🟡 IBCM loss before taxes CHF 162 m; quality-of-earnings reliance on one-off

RECOMMENDED SUPERVISOR ACTIONS (5):
1. 🟡 Require quarterly IRRBB reporting (EVE/NII) and remediation plan for heigh
2. 🟡 Maintain enhanced monitoring of capital planning (US IHC) until two consec
3. 🟡 Request NSFR disclosure and liquidity stress testing under elevated facili
4. 🟡 Conduct targeted review of legal reserves for RMBS/Mozambique/benchmark ma
5. 🟡 Seek detailed plan to restore IBCM profitability and reduce reliance on or

KEY QUESTIONS FOR MANAGEMENT (6):
1. What concrete steps will reduce banking-book rate sensitivity after the +200
2. How much FY2020–2021 PBT guidance excludes one-offs (Invest Lab, SIX, real e
3. What is the updated status of US IHC CCAR remediation and any remaining mode
4. Under plausible COVID stress, what are expected credit loss ranges by segmen
5. What governance changes followed the observation incidents and AML decrees;
6. Provide NSFR trajectory and liquidity impacts from client facility drawdowns

90-DAY MONITORING WATCHLIST (6 items):
1. LCR trajectory amid facility drawdowns; any notable outflow spikes.
2. CCAR results and any residual constraints on US IHC distributions.
3. Legal milestones: MBIA decision, Highland appeal, Mozambique proceedings.
4. IBCM pipeline resilience and quarterly divisional profitability.
5. IRRBB sensitivity updates and hedge program adjustments.
6. Credit quality trends (PD migrations, lombard margin calls, Swiss mortgages)

ASSESSMENT CONFIDENCE:
🟡 Confidence Level: Medium

Data Gaps Identified (6):
   1. NSFR percentage not disclosed (observation period only)
   2. IRRBB EVE/NII disclosure limited (sensitivity given but not full metrics)
   3. IFRS 9 Stage 2/3 not applicable (US GAAP reporting)
   4. Granular Pillar 3 tables referenced externally
   5. Auditor's opinion not in provided sections
   ... and 1 more gaps

KEY FINANCIAL METRICS SUMMARY:
CET1 ratio (Group)..................... 12.7%
Swiss CET1 vs minimum.................. 12.6% vs 10.0% (headroom ~260 bps)
LCR.................................... 198%
Provision for credit losses............ CHF 324 m
Fee income............................. CHF 11,158 m
IRRBB +200 bps shock (most adverse)..... CHF 628 m loss
Gross derivatives notional............. CHF 20,329.1 bn
Legal possible losses (excess of provisions) 0–CHF 1.3 bn
NPL ratio.............................. 0.5%
Specific coverage on impaired loans..... 28.6%
Irrevocable loan commitments........... CHF 125.1 bn
```

Quantitative Risk Assessment Summary

- Aggregate Risk Scores
- Individual Scores
- Overall Risk Classification
- Threshold Definitions

```
================================================================================
RISK ASSESSMENT SUMMARY
================================================================================

Individual Agent Risk Scores (0.0=low, 1.0=high):
================================================================================
🔴 earnings_quality.................. 1.000 (HIGH)
🔴 legal_reg......................... 1.000 (HIGH)
🔴 red_flags......................... 1.000 (HIGH)
🔴 discrepancy_auditor............... 1.000 (HIGH)
🔴 analyst_concern................... 0.913 (HIGH)
🔴 capital_buffers................... 0.850 (HIGH)
🔴 credit_quality.................... 0.850 (HIGH)
🟡 camels_fuser...................... 0.680 (MEDIUM)
🟡 topic_analyzer.................... 0.673 (MEDIUM)
🟡 governance_controls............... 0.600 (MEDIUM)
🟡 liquidity_funding................. 0.500 (MEDIUM)
🟡 business_model.................... 0.500 (MEDIUM)
🟡 off_balance_sheet................. 0.500 (MEDIUM)
🟡 sentiment_tracker................. 0.477 (MEDIUM)
🟢 confidence_evaluator.............. 0.390 (LOW)
🟢 market_irrbb...................... 0.200 (LOW)
--------------------------------------------------------------------------------

🟡 AVERAGE RISK SCORE: 0.696 (MEDIUM RISK)
   Total agents analysed: 16
   Highest risk score: 1.000
   Lowest risk score: 0.200
   Score variance (σ): 0.247

Risk Distribution:
   🟢 Low Risk (0.0-0.4):     2 agents (12.5%)
   🟡 Medium Risk (0.4-0.7):  7 agents (43.8%)
   🔴 High Risk (0.7-1.0):    7 agents (43.8%)

Interpretation:
   ⚠️  7 agent(s) detected high risk
   → Focus on high-risk areas identified
   → Review CAMELS assessment for details

Risk by Agent Category:
   🟡 Linguistic/Behavioural: 0.613 (4 agents)
   🟡 Quantitative Metrics:   0.667 (9 agents)
   🔴 Pattern Detection:      1.000 (1 agent)
   🔴 Meta-Analysis:          0.840 (2 agents)

================================================================================
```

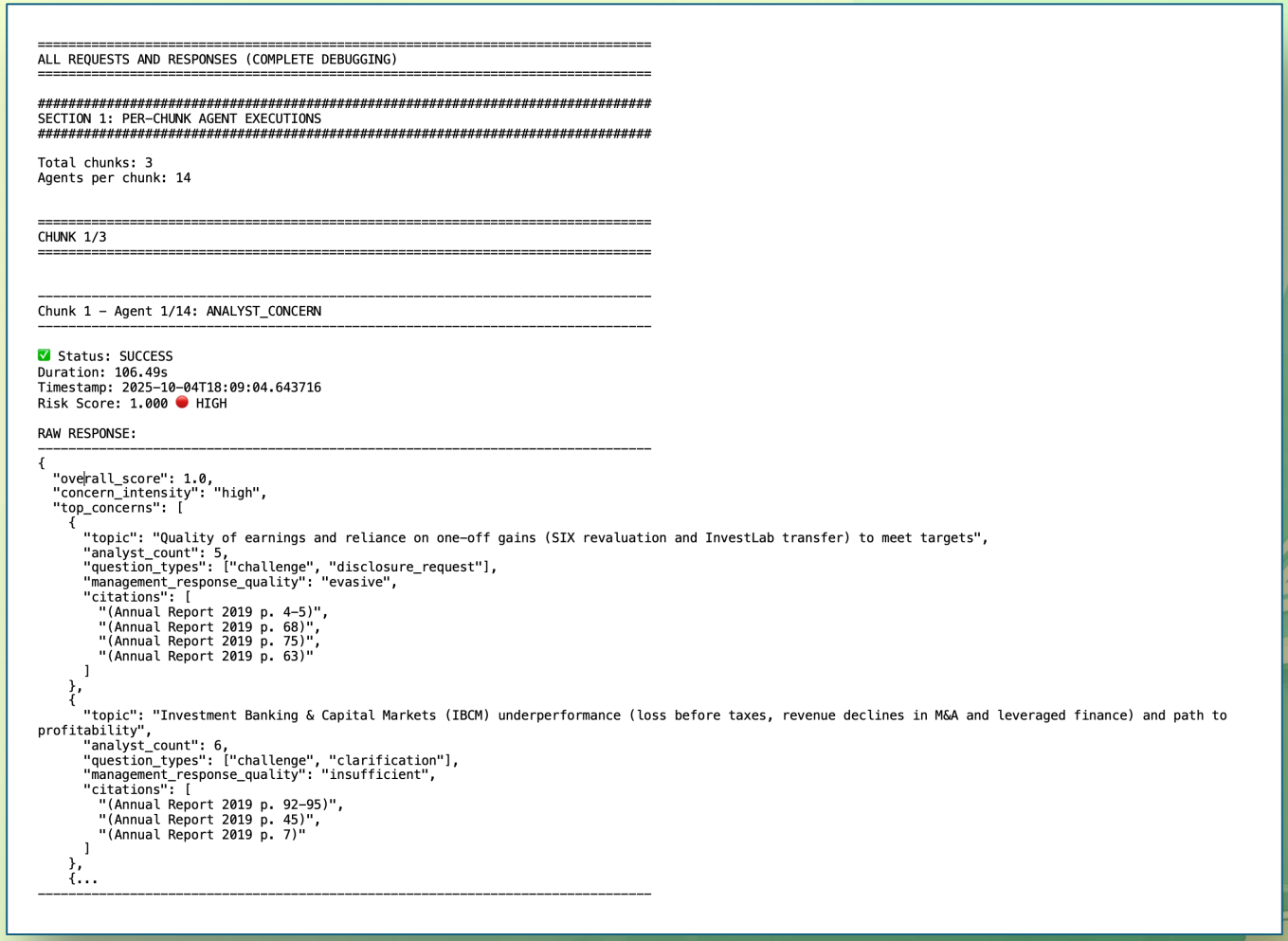
System Transparency & Audit Trail

Complete Request-Response Log Key Features:

- Every agent prompt captured verbatim
- Full LLM responses logged with timestamps
- Model metadata tracked (tokens, duration, model version)
- Auto-export to timestamped debug files

Compliance Benefits:

- Full audit trail for regulatory review
- Source traceability for every finding
- Reproducibility verification
- Debugging transparency



```
================================================================================
ALL REQUESTS AND RESPONSES (COMPLETE DEBUGGING)
================================================================================

################################################################################
SECTION 1: PER-CHUNK AGENT EXECUTIONS
################################################################################

Total chunks: 3
Agents per chunk: 14


================================================================================
CHUNK 1/3
================================================================================


--------------------------------------------------------------------------------
Chunk 1 - Agent 1/14: ANALYST_CONCERN
--------------------------------------------------------------------------------

✅ Status: SUCCESS
Duration: 106.49s
Timestamp: 2025-10-04T18:09:04.643716
Risk Score: 1.000 🔴 HIGH

RAW RESPONSE:
--------------------------------------------------------------------------------
{
  "overall_score": 1.0,
  "concern_intensity": "high",
  "top_concerns": [
    {
      "topic": "Quality of earnings and reliance on one-off gains (SIX revaluation and InvestLab transfer) to meet targets",
      "analyst_count": 5,
      "question_types": ["challenge", "disclosure_request"],
      "management_response_quality": "evasive",
      "citations": [
        "(Annual Report 2019 p. 4-5)",
        "(Annual Report 2019 p. 68)",
        "(Annual Report 2019 p. 75)",
        "(Annual Report 2019 p. 63)"
      ]
    },
    {
      "topic": "Investment Banking & Capital Markets (IBCM) underperformance (loss before taxes, revenue declines in M&A and leveraged finance) and path to
profitability",
      "analyst_count": 6,
      "question_types": ["challenge", "clarification"],
      "management_response_quality": "insufficient",
      "citations": [
        "(Annual Report 2019 p. 92-95)",
        "(Annual Report 2019 p. 45)",
        "(Annual Report 2019 p. 7)"
      ]
    },
    {...
--------------------------------------------------------------------------------
```

Technical Architecture:

Production-Grade Jupyter Notebook

Notebook Structure (27 cells):

1. **Setup & Config (Cells 1-5)** → API keys, model selection, file paths
2. **Document Processing (6.1-6.5)** → PDF extraction, chunking, rate limiting
3. **LLM Communication (7.1-7.3)** → API calls, JSON parsing, retry logic
4. **Agent Prompts (Cell 8)** → 16 specialized system prompts
5. **Execution Framework (9.1-9.7)** → Single agent, parallel, aggregation
6. **Main Pipeline (10.1-10.7)** → End-to-end orchestration
7. **Results Display (25-27)** → Debug logs, scores, CAMELS

Key Technical Specifications:

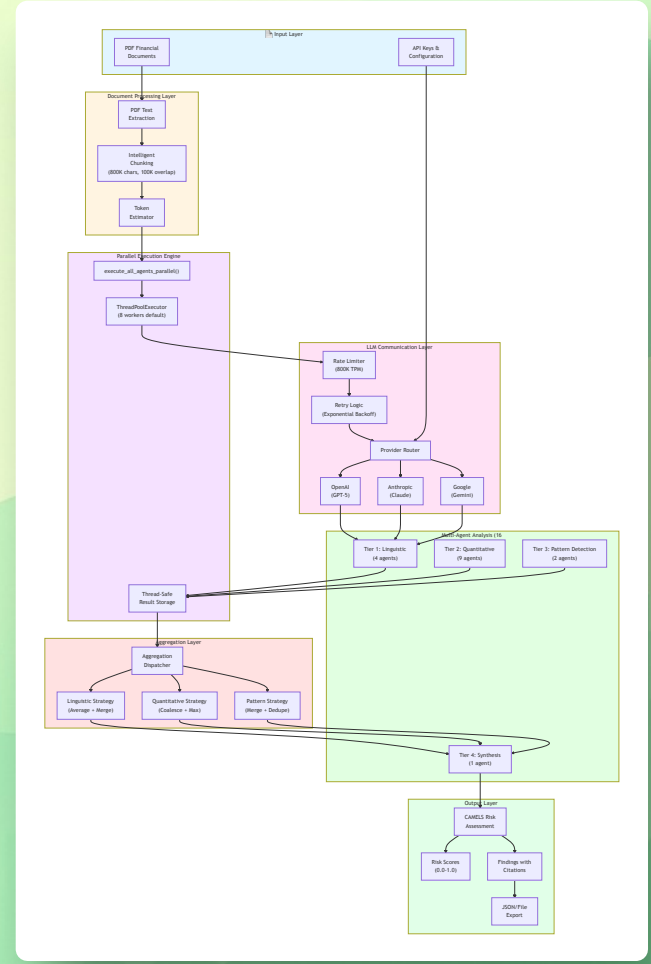
- **Language:** Python 3.11
- **Primary LLM in the example:** GPT-5 (272K token input, 128K token output)
- **Concurrency**: ThreadPoolExecutor with 2 workers
- **Rate Limiting:** Token bucket algorithm (800K tokens/min)
- **Retry Strategy:** Exponential backoff (2s → 4s → 8s delays)
- **Chunking:** 800K chars/chunk with 100K overlap (12.5% redundancy)
- **Aggregation:** Strategy-specific (linguistic=average, quantitative=coalesce)

Processing Time:

Embedded Example Credit Suisse Annual Report 2019: 442 pages, 1.87M chars, 3 chunks, 42 agent executions
Total runtime: 50 minutes (Chunk 1: 14min, Chunk 2: 12min, Chunk 3: 18min, Meta: 6min)

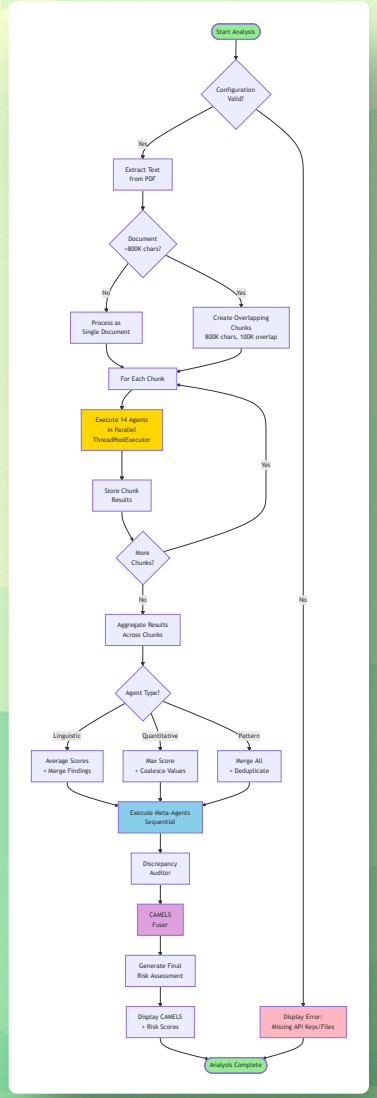
RiskRadar System Architecture

- **7-layer modular design** for separation of concerns
- **Multi-provider LLM support** (OpenAI, Anthropic, Google) with unified interface
- **16 specialized agents** organized in 4 analytical tiers
- **Parallel execution engine** (ThreadPoolExecutor) for efficient processing
- **3 aggregation strategies** tailored to agent output types
- **Production features:**
  Rate limiting (800K TPM), exponential backoff retry, thread-safe operations
- **Output:**
  CAMELS framework assessment with citations for regulatory compliance



RiskRadar Execution Pipeline

- **Validation gate:**
  API keys, file paths, model configuration checked upfront
- **Intelligent chunking decision:**
  Documents >800K chars split with overlap
- **Single-pass optimization:**
  Small documents bypass chunking overhead
- **Parallel agent execution:**
  14 agents run concurrently per chunk (7x speedup)
- **Three aggregation paths:**
  Linguistic (average), Quantitative (coalesce), Pattern (merge)
- **Sequential meta-analysis:**
  Discrepancy auditor → CAMELS fuser
- **Graceful degradation:**
  Failed agents don't halt pipeline; partial results still usable
- **End-to-end timing:**
  ~1-2 minutes for 400-page document including all retries



16-Agent Multi-Tier Framework

Tier 1 - Linguistic Analysis (4 agents):

Sentiment, topic shifts, confidence levels, analyst concerns

Tier 2 - Quantitative Metrics (9 agents):

Capital, liquidity, credit, earnings, governance, legal, business model, off-balance-sheet

Tier 3 - Pattern Detection (1 agent):

Red flags: going concern, covenant breach, material uncertainty

Tier 4 - Synthesis (1 agent):

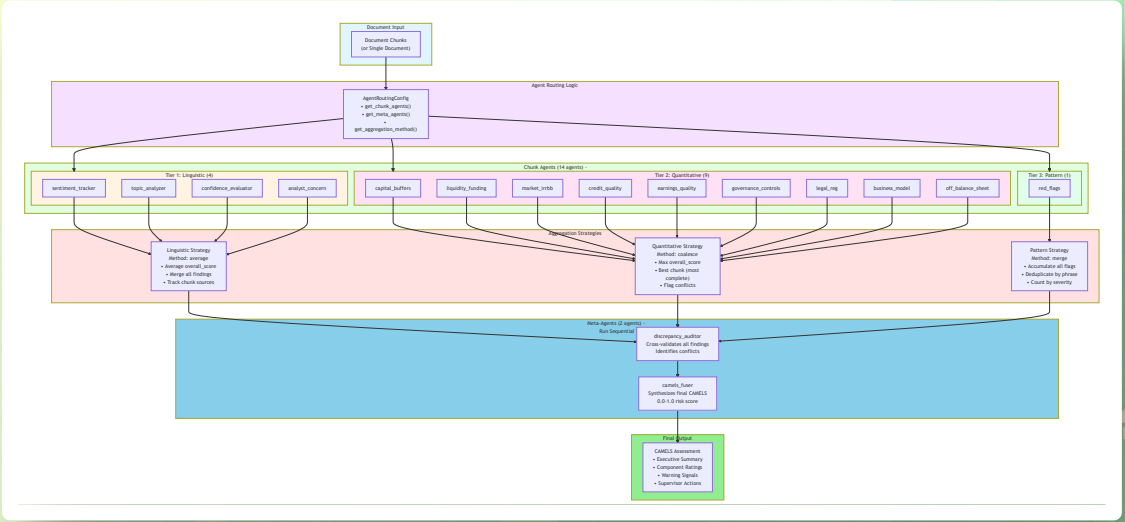
CAMELS fuser: integrates all findings into regulatory assessment

Meta-Agents (2 agents):

Discrepancy auditor (cross-validation), CAMELS fuser (final synthesis)

Routing Logic:

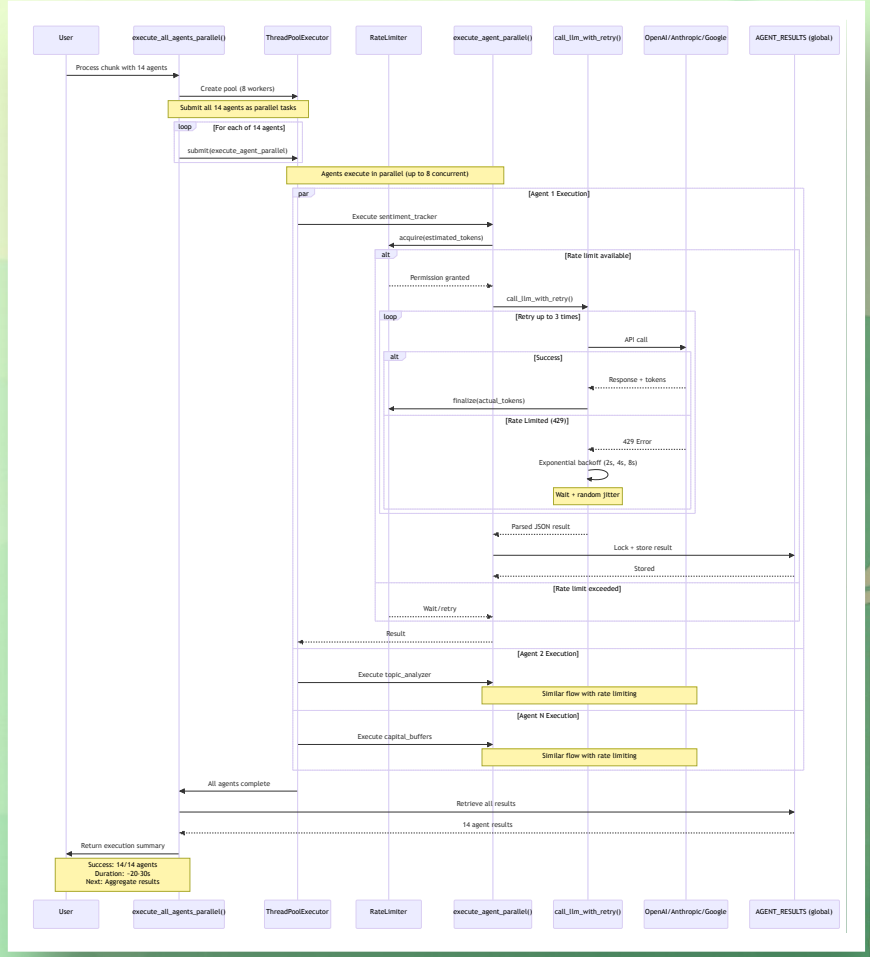
AgentRoutingConfig determines chunk vs. meta execution - Aggregation method assigned per agent type (average/coalesce/merge)



RiskRadar Component Interactions:

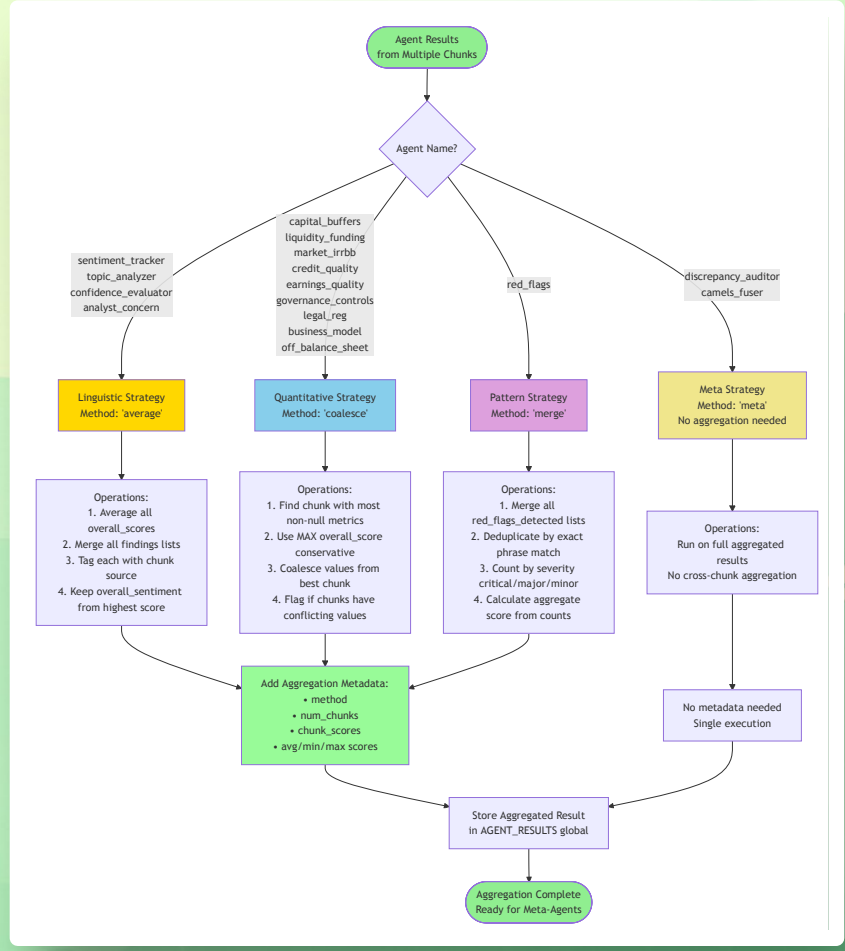
Parallel Execution with Rate Limiting

- **ThreadPoolExecutor:**
  8 concurrent workers (configurable based on API tier)
- **Rate limiting:**
  Token bucket algorithm, 800K tokens/minute default
- **Coordination:**
  Each agent acquires() rate limit slot before API call
- **Retry logic:**
  Exponential backoff (2s, 4s, 8s) + random jitter (0-1.5s)
- **Error discrimination:**
  Rate limit (429) → retry; other errors → fail fast
- **Thread safety:**
  Lock protects writes to AGENT_RESULTS global dictionary
- **Progress tracking:**
  Real-time display as agents complete (not in submission order)
- **Success rate:**
  ~85% first attempt, ~98% after retries (production tested)



Aggregation Strategy Routing

- **Linguistic Agents (average method):**
  Average overall score across chunks
  - Merge all findings lists (keyphrases, topics, evasiveness examples)
  - Rationale: Cumulative evidence of tone/narrative issues
- **Quantitative Agents (coalesce method):**
  Select chunk with most non-null metrics (best completeness)
  - Use MAXIMUM overall score (conservative risk assessment)
  - Flag conflicts if same metric has different values across chunks
  - Rationale: Metrics should be consistent; conservatively report highest risk
- **Pattern Agents (merge method):**
  Concatenate all redflags detected lists
  - Deduplicate by exact phrase match (handles overlap)
  - Count by severity (critical/major/minor) and calculate weighted score
  - Rationale: Accumulate all warning signals; avoid double-counting overlap
- **Meta Agents (no aggregation):**
  Discrepancy auditor and CAMELS fuser run once on full aggregated results



RAG Q&A System:

Semantic Document Search

What is RAG?

- Retrieval-Augmented Generation
- Enables natural language Q&A over financial documents
- Grounds LLM answers in actual source text

Two-Phase Architecture:

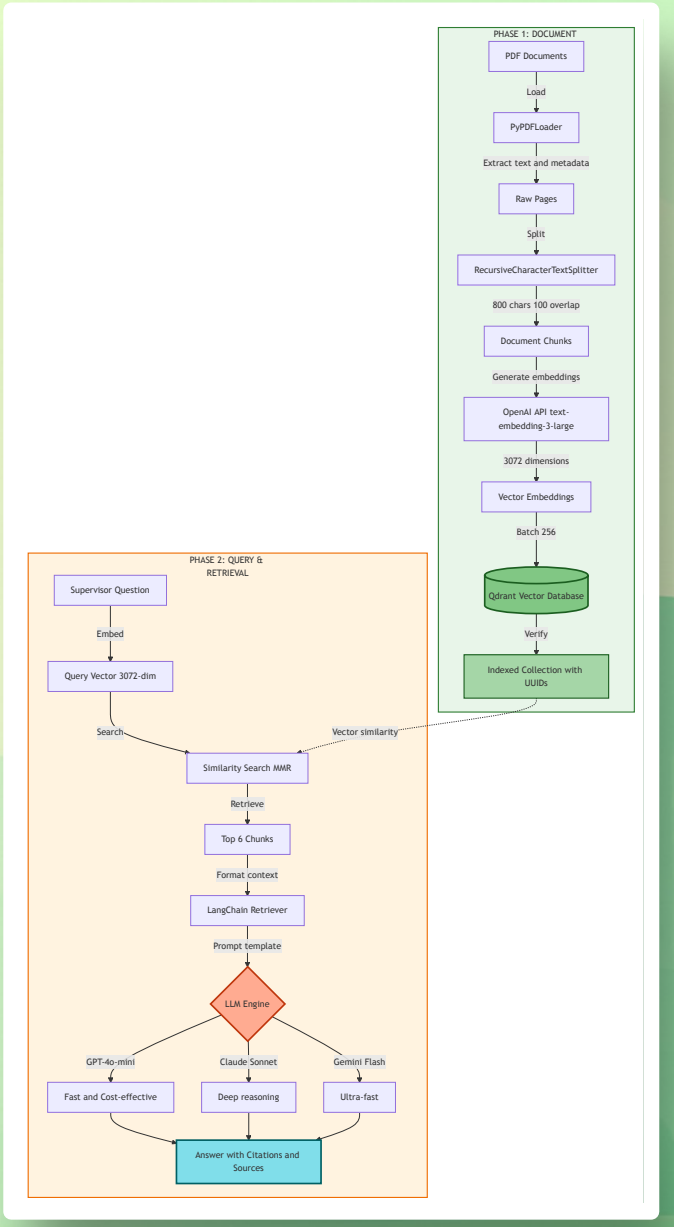
- **Phase 1: Document Indexing (One-time setup)**
  - Load PDFs → Extract text → Chunk (800 chars)
  - Generate embeddings (3072 dimensions)
  - Store in Qdrant vector database
- **Phase 2: Query & Answer (Real-time)**
  - User question → Embed → Search similar chunks
  - Retrieve top 6 relevant passages
  - LLM generates answer with source citations

Key Benefits:

- A few seconds per query (vs 30 min manual search)
- Semantic search finds concepts, not just keywords
- Every answer includes verifiable citations
- Multi-model support (GPT/Claude/Gemini)

Technical Specs:

- Embedding: text-embedding-3-large (OpenAI)
- Vector DB: Qdrant (cloud or local)
- Retrieval: MMR algorithm (balanced relevance + diversity)



Questions & Contact Information

Project Resources:

- Jupyter Notebooks & Source code on GitHub:

  https://github.com/hamiltonalex/riskradar-boe

- Live Demo:

  https://riskradar-boe.streamlit.app

Thank you!

Together, we are building a safer financial system through intelligent supervision.