

# Multi-Agent Financial Report Generation Using FinRobot: Engineering Challenges, Token Control, and Industry-Level Evaluation



Shenyuan Wu 3 min read · Just now



## Abstract

This blog presents an enhanced evaluation of a multi-agent financial report generation system built on FinRobot and large language models (LLMs). Beyond structural analysis, this study incorporates engineering failure case examination, token-per-minute (TPM) control mathematics, and quantitative scoring of output quality across five semiconductor firms. The findings demonstrate strong macro-level industry alignment, while identifying limitations in financial validation and valuation modeling depth.

## 1. Introduction

Recent advances in large language models have enabled autonomous analytical pipelines in finance. However, practical deployment faces engineering constraints including token-per-minute limits, JSON truncation errors, and nested conversation inflation. This project evaluates whether a constrained multi-agent architecture can generate high-quality equity research reports while remaining API-compliant.

## 2. System Architecture

The system is implemented as a single tool-augmented agent configured to generate a complete annual equity research report and compile it into a PDF artifact. Rather than separating writing and document construction into distinct agents, the workflow is encoded directly into one task prompt that enforces strict sequencing: long-form drafting, word-count verification, visualization generation, file existence checks, and a final PDF build step.

To mitigate reliability issues commonly observed in agentic tool use, the prompt specifies hard operational constraints: (i) the report text must be produced in one pass with at most one revision, (ii) tool calls must be emitted as strict JSON without literal newlines, (iii) each string field in tool-call arguments must remain below a fixed character budget, and (iv) PDF construction must be performed only after image generation and file validation succeed. This single-agent design improves reproducibility by centralizing control logic and minimizing inter-agent message passing.

## 3. Engineering Challenges and Failure Cases

Several failure cases were encountered during development. A token-per-minute (TPM) explosion occurred when long tool outputs (e.g., large excerpts from filings or verbose tool logs) were repeatedly appended to the conversation context, producing requests far above the organization's 30,000

TPM limit (e.g., ~32,000 tokens). Additional issues included malformed tool-call JSON (unterminated strings caused by truncated outputs), multi-conversation ambiguity in trigger logic, and Windows path-escaping errors (e.g., unintended \n or \t sequences) that prevented image files from being read correctly.



These failures were mitigated within the single-agent workflow by enforcing: (i) a one-pass drafting policy with at most one revision, (ii) strict JSON-only tool calls with newline escaping ("\\n") and quote escaping, (iii) a per-field character cap to reduce truncation risk, (iv) deterministic file naming for generated charts, and (v) explicit file existence checks prior to PDF construction. In practice, placing required image-path fields early in the final build payload further reduced the probability of truncation-induced JSON errors.

```
RateLimitError: Error code: 429 - {'error': {'message': 'Request too large for gpt-4.1 in organization org-SUm5a5FR8KLR5yAIOcIFoEEw on tokens per min (TPM): Limit 30000, Requested 32263. The input or output tokens must be reduced in order to run successfully. Visit https://platform.openai.com/account/rate-limits to learn more.', 'type': 'tokens', 'param': None, 'code': 'rate_limit_exceeded'}}}
```

## 4. Token Control Mathematical Framework

Let  $T_{\text{total}} = T_{\text{input}} + T_{\text{output}}$ . Under a TPM limit of 30,000 tokens, recursive tool outputs cause  $T_{\text{input}}$  to grow linearly with each iteration:  $T_{\text{input}}(n) = T_{\text{base}} + n * T_{\text{tool}}$ . When  $T_{\text{tool}}$  is large (e.g., full 10-K excerpts), the total quickly exceeds the API threshold. By enforcing  $n = 1$  and minimizing  $T_{\text{tool}}$  via summarization, the system maintains  $T_{\text{total}} < 30,000$ .

## 5. Quantitative Evaluation Framework

Company	Structural	Financial Accuracy	Industry Alignment	Risk Realism
NVIDIA	5	4	5	4
AMD	4	4	4	4
Broadcom	4	3	4	3
Texas Instruments	5	4	5	4
Qualcomm	4	4	4	4

## 6. Industry Alignment Analysis

The system correctly identified NVIDIA as the AI infrastructure leader, AMD as a challenger, Broadcom as a networking consolidator, Texas Instruments as an analog stability play, and Qualcomm as a mobile-edge AI hybrid firm. These classifications align with prevailing market perspectives.

## Conclusion

This upgraded evaluation demonstrates that a constrained, single-agent, tool-augmented LLM workflow can generate structurally consistent and industry-aligned annual research reports under strict engineering constraints. While the outputs are not yet equivalent to institutional-grade analyst research — particularly in scenario-based valuation modeling and audited financial verification — the framework offers a scalable foundation for AI-assisted equity research when paired with robust token-control policies, tool-call governance, and post-generation validation.

AI

LLM

Finrobot

**Written by Shenyuan Wu**

0 followers · 1 following

[Edit profile](#)**No responses yet**

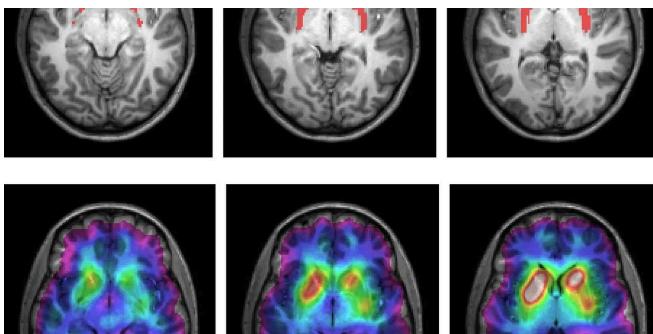
...



Shenyuan Wu

What are your thoughts?

**Recommended from Medium**



:2510.011171v3 [cs.CL] 10 Oct 2025

**ABSTRACT**

Post-training alignment often reduces LLM diversity, leading to a phenomenon known as *mode collapse*. Unlike prior work that attributes this effect to algorithmic limitations, we identify a fundamental cause: data distributional *typicality bias* in preference datasets. Any annotator systematically finds familiar text as a result of well-established findings in cognitive psychology. We formalize this bias theoretically, verify it on preference datasets empirically, and show that it plays a central role in mode collapse. Motivated by this analysis, we introduce *Verbalized Sampling (VS)*, a simple, training-free prompting strategy to circumvent mode collapse. VS prompts the model to verbalize a probability distribution over a set of responses (e.g., “Clean” vs “Rotten”) and then samples from that distribution. Our experiments show that VS significantly improves performance across creative writing (poems, stories, jokes), dialogue simulation, open-ended QA, and synthetic data generation, without sacrificing factual accuracy and safety. For instance, in creative writing, VS increases diversity by 1.6–2.1× over direct prompting. We further observe an emergent trend that more capable models benefit more from VS. In sum, our work provides a new data-centric perspective on mode collapse and a practical inference-time remedy that helps unlock pre-trained generative diversity.

Problem: Typicality Bias Causes Mode Collapse

Solution: Verbalized Sampling (VS) Mitigates Mode Collapse

Different prompts collapse to different modes:



In Write A Catalyst by Dr. Patricia Schmidt

## As a Neuroscientist, I Quit These 5 Morning Habits That Destroy You...

Most people do #1 within 10 minutes of waking (and it sabotages your entire day)

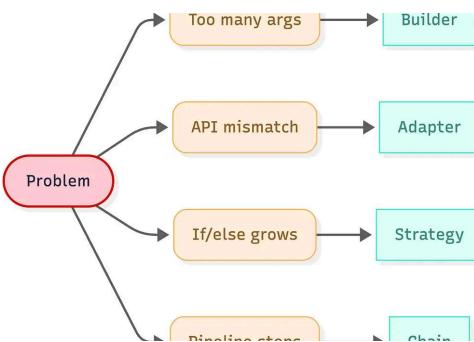
★ Jan 14

31K

557



...



In Women in Technology by Alina Kovtun ★

## Stop Memorizing Design Patterns: Use This Decision Tree Instead

Choose design patterns based on pain points: apply the right pattern with minimal over-...

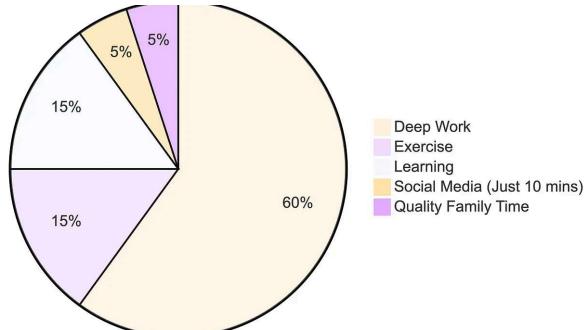
★ Jan 29

4K

33



...



∞ In Level Up Coding by Teja Kusireddy

## I Stopped Using ChatGPT for 30 Days. What Happened to My Brai...

91% of you will abandon 2026 resolutions by January 10th. Here's how to be in the 9% who...

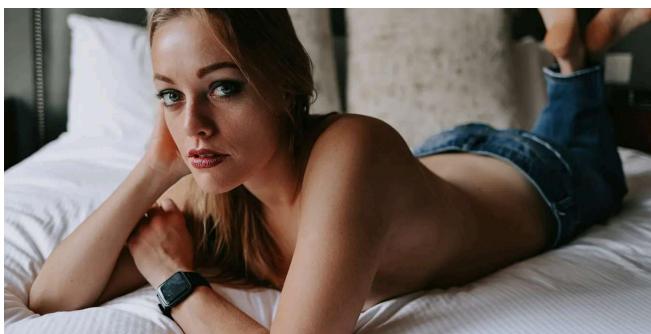
★ Dec 28, 2025

6.8K

273



...



Jonatha Czajkiewicz

## What a Sex Worker Notices About Gen X and Gen Z Men

How masculinity changed between Grunge and TikTok

Nov 16,  
2025

20K

519



...

Nov 26,  
2025

5.7K

220



...

[See more recommendations](#)



Michal Malewicz

## The End of Dashboards and Design Systems

Design is becoming quietly human again.

Nov 26,  
2025

5.7K

220



...

[Open in app](#) ↗