Generation: A Practical Evaluation of Reliability, Output Quality, and Industry Alignment

Yang Yu 4 min read · Just now

To evaluate FinRobot's ability to generate investor-style company reports in a one-shot batch setting, I implemented a reliability-oriented pipeline and used it to produce the five reports specified by the instructor (INTC, NVDA, GOOGL, AMD, AAPL). In addition, I generated one extra report (PLTR) as an out-of-sample extension to test whether the workflow remains stable and whether the qualitative conclusions generalize beyond the required set. Although these companies span multiple sectors, they share a common market backdrop entering 2026: the key divergence is no longer whether to invest in AI, but the intensity of AI spending, the pathway to returns, and who can convert AI-related capital expenditures into sustainable cash flow. This macro lens provides a consistent basis to interpret the outputs across semiconductors, internet platforms, consumer hardware, and enterprise software.

To enable one-shot batch generation without conflicts, I introduced four reliability-oriented upgrades targeting reproducibility, scalability, isolation, and preflight validation. First, I pinned key dependency versions at the pipeline entry to prevent variations across machines or over time that could change tool-call behavior, file-generation workflows, or chart-rendering outputs, thereby improving reproducibility. Second, I extended the nested-chat trigger from reading a single instruction file to parsing and merging multiple instruction files, allowing users to provide multiple resources and constraints in one run and reducing context loss or missed requirements caused by multi-turn interactions, which improves multi-agent execution consistency. Third, I upgraded the workflow from single-company, single-run generation to looping over a company list, and assigned each company an independent output subdirectory (e.g., named by ticker/company) to isolate files and assets, preventing filename collisions, overwrites, and cross-report chart/reference leakage and ensuring sandboxed builds per company. Finally, I added a SEC/FMP data-source preflight check before launching the agent workflow to directly validate API keys and the stability of critical endpoints; if the preflight fails, the run stops or prompts for fixes, reducing

the risk that downstream agents fill in missing data with fabricated details and protecting the accuracy of the final conclusions.

After generation, output quality was evaluated qualitatively across four dimensions: structural consistency, financial reliability, industry alignment, and risk realism. Overall, the system performed well on structure and coverage: the reports generally followed a standardized section layout, making cross-company comparison straightforward and supporting the use case of fast screening. At the narrative level, the industry mapping was broadly consistent with mainstream views: NVDA and AMD were framed within the AI data-center compute/platform competition; INTC was treated as a turnaround story dominated by manufacturing capability and product-cadence recovery; GOOGL was characterized as an advertising cash-flow engine with Cloud/AI as incremental investment and growth; AAPL was described as a mature hardware base supported by services-driven resilience; and PLTR was presented as a data-platform expansion story driven by government and commercial adoption. In particular, within the same-industry semiconductor peer set (NVDA, AMD, INTC), the implied "leader–challenger–turnaround" structure aligns with prevailing investor narratives, suggesting the system can capture high-level industry positioning and relative competitive roles.

By contrast, the primary bottleneck lies in financial verifiability and the applicability of valuation fields. Some reports contained inconsistencies where growth-rate figures did not reconcile with underlying revenue trends, and extreme FCF-conversion or valuation multiples appeared in loss-making or highly volatile periods without sufficient definition or accounting context. Target-price ranges were also sometimes produced without meaningful constraints or explicit scenario assumptions, which makes them resemble unconstrained model artifacts rather than bounded valuation outputs. These missing sanity checks materially reduce research usability because credibility is typically judged first through numbers that can be audited and reconciled. In addition, the narrative occasionally showed repetition and template-like phrasing — adjacent sections reiterating similar points, competitor analysis leaning toward metric restatement rather than key-variable differentiation, and risk lists that were broad but insufficiently prioritized — indicating that section objectives may need to be tightened to increase incremental information density.

From the perspective of alignment with current market pricing, part of the valuation conclusions across the reports can be considered broadly consistent with prevailing market multiple ranges, most notably for NVDA, AAPL, GOOGL, and AMD. As of around mid-February 2026, NVDA traded at approximately $189.82 with a market EV/EBITDA of about 40.08 (2026–02–

10); AAPL traded at approximately $264.58 with a P/E of about 34.38x and an EV/EBITDA of about 26.85 (2026–02–11); GOOGL traded at approximately $314.98 with a P/E of about 23.65x and an EV/EBITDA of about 22.87 (2026–02–18); and AMD traded at approximately $200.15 with an EV/EBITDA of about 49.21 (2026–02–13). Therefore, if the reports' valuation views for these companies fall within comparable multiple ranges and interpret pricing through an "AI infrastructure vendor premium" (NVDA/AMD) and a "mature cash-flow-supported reinvestment/services premium" (GOOGL/AAPL), then this portion of the conclusions can be considered broadly consistent with the market's mainstream valuation framework, while PLTR serves as a useful extension case where high-multiple regimes make metric selection and post-generation validation especially important.

Taken together, the experiment suggests that FinRobot, under the proposed batch setup, functions reliably as a structured first-draft generator: it can produce standardized, industry-aligned company overviews and a baseline metric framework at scale, and it captures mainstream positioning reasonably well across both the required set and the added PLTR case. However, reaching a higher research standard likely requires stricter post-generation validation and stronger section-level objectives, including reconciliation checks between growth rates and level data, metric applicability gates (e.g., suppressing P/E in loss years and substituting appropriate multiples), scenario-bounded target-price construction, and redundancy controls that reduce repetition and force competitor analysis to focus on key variables rather than generic metric restatement.

Y

Written by Yang Yu

0 followers · 1 following

Edit profile

No responses yet

Y Yang Yu

What are your thoughts?

Recommended from Medium