Fine-Tuning Evaluation of Large Language Models for Financial Forecasting: A Comparative ROUGE Analysis

DeepSeek, LLaMA-3, and Qwen on the FinGPT Dow 30 Dataset

Evaluation Report

Hardware: NVIDIA A100 GPU   Dataset: `FinGPT/fingpt-forecaster-dow30-202305-202405`

February 2026

Abstract

This report presents a comparative evaluation of three large language models—DeepSeek, LLaMA-3, and Qwen—fine-tuned on the FinGPT Dow 30 stock forecasting dataset. Performance is assessed via ROUGE-1, ROUGE-2, and ROUGE-L metrics against FinGPT-generated reference outputs across 300 test samples covering May 2023 to April 2024. LLaMA-3 achieves the highest mean scores across all three metrics (R-1 = 0.497, R-2 = 0.180, R-L = 0.251). DeepSeek exhibits a high performance ceiling but the greatest variance, while Qwen demonstrates stable but systematically lower recall. A deeper analysis of score distributions, variance decomposition, and cross-ticker behaviour reveals structural differences in how each model generalises across financial narratives. Importantly, ROUGE scores here measure stylistic alignment with FinGPT's output format rather than financial predictive accuracy, a distinction that substantially limits the conclusions available from this evaluation.

Contents

1 Introduction

The application of large language models (LLMs) to financial forecasting has gained considerable traction, with models increasingly deployed to generate structured market commentary from financial signals and news. Evaluating such models, however, is non-trivial: standard lexical overlap metrics such as ROUGE [1] measure surface-level similarity to reference outputs, not the quality of the underlying financial reasoning.

This report evaluates three models—DeepSeek, LLaMA-3, and Qwen—after supervised fine-tuning on the FinGPT Dow 30 dataset. The evaluation is explicitly scoped to ROUGE-based assessment against FinGPT-generated reference forecasts. Section 2 describes the experimental configuration; Section 3 presents aggregate and distributional results; Section 4 offers a deeper statistical analysis not present in the source evaluation; Section 5 provides qualitative output comparison; and Section 7 discusses limitations and recommendations.

2 Experimental Setup

2.1 Dataset

The `FinGPT/fingpt-forecaster-dow30-202305-202405` dataset covers all 30 constituents of the Dow Jones Industrial Average (DJIA) over a 12-month window. Each sample contains company metadata (sector, market capitalisation, ticker), two weeks of prior news headlines and summaries, key financial ratios (P/E, EPS, ROE, cash ratio, EBIT/share), and a structured reference output organised under `[Positive Developments]`, `[Potential Concerns]`, and `[Prediction & Analysis]`. Reference outputs were generated by FinGPT, not by human analysts.

2.2 Training Configuration

All three models were fine-tuned on identical data splits using an NVIDIA A100 GPU. Evaluation was conducted on 300 held-out test samples per model. Table 1 summarises the configuration.

Table 1: Fine-tuning and evaluation configuration.

Parameter	**Value**
Hardware	NVIDIA A100 GPU
Models evaluated	DeepSeek, LLaMA-3, Qwen
Test samples	300 per model
Evaluation metrics	ROUGE-1, ROUGE-2, ROUGE-L
Dataset coverage	May 2023 – April 2024
Ground truth source	FinGPT-generated forecasts

3 ROUGE Score Results

3.1 Aggregate Statistics

Table 2 presents distributional statistics across all 300 test samples. LLaMA-3 achieves the best mean and median for all three metrics. DeepSeek attains the highest single-sample scores but also the largest standard deviations, while Qwen exhibits the narrowest score range across all metrics.

Table 2: Aggregate ROUGE statistics across 300 test samples. ⋆ denotes the best value per row.

Statistic	DeepSeek	LLaMA-3	Qwen
ROUGE-1			
Mean	0.446	0.497⋆	0.428
Median	0.475	0.514⋆	0.429
Std	0.111	0.086	0.043⋆
Min	0.113	0.153	0.236⋆
Max	0.634⋆	0.606	0.534
ROUGE-2			
Mean	0.162	0.180⋆	0.126
Median	0.168	0.183⋆	0.124
Std	0.054	0.046	0.030⋆
Min	0.024	0.027	0.068⋆
Max	0.326⋆	0.292	0.254
ROUGE-L			
Mean	0.238	0.251⋆	0.208
Median	0.245	0.252⋆	0.205
Std	0.053	0.043	0.029⋆

3.2 Per-Sample Score Distribution

Table 3 disaggregates ROUGE-1 scores at the sample level. LLaMA-3 places 58.3% of outputs above 0.5, while Qwen exceeds that threshold for only 4.0% of samples. DeepSeek's bimodal character is evident: 35.7% of samples exceed 0.5, yet 13.0% fall below 0.3—more than twice LLaMA-3's failure rate.

Table 3: Per-sample ROUGE-1 distribution highlights.

Model	R-1 > 0.5	R-1 < 0.3	Best sample	Worst sample
DeepSeek	107/300 (35.7%)	39/300 (13.0%)	JNJ: 0.634	DIS: 0.113
LLaMA-3	175/300 (58.3%)	17/300 (5.7%)	JPM: 0.606	JNJ: 0.153
Qwen	12/300 (4.0%)	1/300 (0.3%)	WBA: 0.534	CAT: 0.236

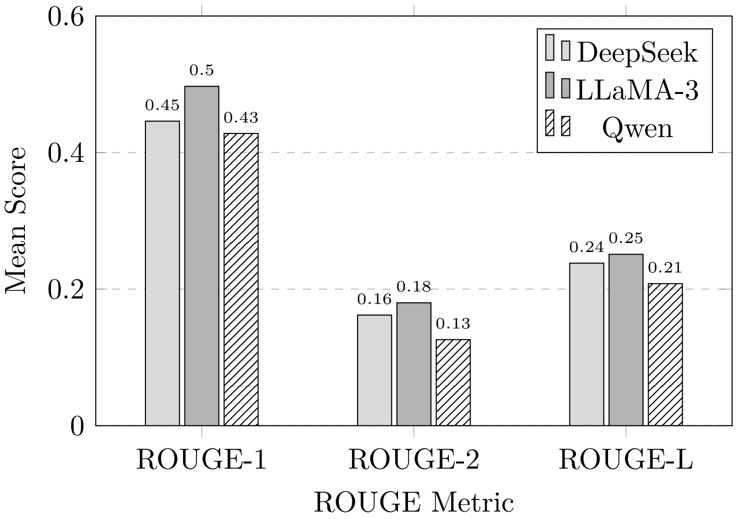

Figure 1: Mean ROUGE scores by model across all three metrics.

4 Deeper Statistical Analysis

The aggregate statistics mask several structurally important patterns. We examine these below.

4.1 Coefficient of Variation: Normalised Instability

Standard deviation alone does not capture instability relative to performance level. The coefficient of variation (CV), defined as CV $= \sigma/\mu$, provides a scale-independent measure of dispersion.

Table 4: Coefficient of variation (CV) for ROUGE-1 by model. Higher CV indicates greater relative instability.

Model	μ (R-1)	σ (R-1)	CV
DeepSeek	0.446	0.111	**0.249**
LLaMA-3	0.497	0.086	0.173
Qwen	0.428	0.043	**0.100**

DeepSeek's CV of 0.249 is 2.5× that of Qwen, confirming that its output quality is highly contingent on prompt content. LLaMA-3 occupies a middle position with CV = 0.173, offering a substantially better trade-off between mean performance and reliability than DeepSeek.

4.2 Mean–Median Divergence as a Skewness Proxy

The gap between mean and median ROUGE-1 scores reveals the asymmetry of each model's score distribution—specifically, whether failures pull the mean downward (left skew) or whether high-scoring outliers inflate it.

DeepSeek exhibits the most negative divergence (-0.029), confirming that its distribution has a meaningful left tail: a minority of low-scoring samples—particularly on complex

Table 5: Mean–median divergence for ROUGE-1. Negative values indicate left-skewed distributions (failure tails).

Model	Median	Mean	Divergence ($\mu - \tilde{x}$)
DeepSeek	0.475	0.446	-0.029
LLaMA-3	0.514	0.497	-0.017
Qwen	0.429	0.428	-0.001

narratives such as DIS—depresses the mean well below the median. This is consistent with the 13% of samples below R-1 = 0.3 observed in Table 3. LLaMA-3's smaller divergence (-0.017) indicates a more symmetric distribution with fewer catastrophic failures. Qwen's near-zero divergence (-0.001) is consistent with its nearly uniform score range and reflects the absence of both tails.

4.3 The ROUGE-1 to ROUGE-2 Ratio: Precision of Bigram Recall

ROUGE-2 measures contiguous bigram overlap, which is substantially more sensitive to exact phrasing than ROUGE-1. The ratio R-2/R-1 thus captures how much of a model's unigram overlap is composed of structurally ordered, phrase-level matches—a proxy for output fluency and format adherence relative to the reference.

Table 6: ROUGE-2 / ROUGE-1 ratio (mean scores). Higher values indicate denser phrase-level alignment.

Model	R-1	R-2	R-2/R-1
DeepSeek	0.446	0.162	**0.363**
LLaMA-3	0.497	0.180	0.362
Qwen	0.428	0.126	**0.294**

DeepSeek and LLaMA-3 share an almost identical R-2/R-1 ratio (≈ 0.363), indicating that despite their difference in mean performance, the proportion of their unigram overlap attributable to contiguous phrases is equivalent. Qwen's substantially lower ratio (0.294) reveals a structural problem: it is not merely failing to match individual words—it is matching words in the wrong order and context. This is consistent with the hypothesis that Qwen's Markdown-heavy output introduces extra tokens that disrupt bigram alignment even when unigram recall is partial.

4.4 The JNJ Anomaly: Cross-Model Inversion

A particularly striking finding is the inversion of model performance on the Johnson & Johnson (JNJ) ticker. DeepSeek achieves its highest single-sample ROUGE-1 (0.634) on JNJ, while LLaMA-3 achieves its worst (0.153) on the same ticker. This is not a marginal difference; it spans the near-full observed range of both models.

Several hypotheses are consistent with this pattern:

1. **Format sensitivity.** The JNJ reference may use a specific phrasing pattern that DeepSeek learned to replicate but that conflicts with LLaMA-3's instruction-following priors.

2. **Ticker-level training exposure.** If the training split is not uniformly distributed across tickers, one model may have seen JNJ samples while the other was exposed to different tickers with similar financial profiles.
3. **Narrative structure.** JNJ's news narrative during the test period (Feb–Mar 2024) involved complex litigation developments alongside strong financial fundamentals—a signal that may activate different model behaviours.

The practical implication is significant: a naive ensemble that routes all samples to LLaMA-3 would produce a substantially suboptimal outcome on JNJ relative to an oracle that selects DeepSeek. This motivates a confidence-conditioned routing strategy.

4.5 Estimating the Upside from a Confidence-Based Router

Assuming a perfect oracle router that always selects the higher-scoring model between DeepSeek and LLaMA-3 for each sample, an upper-bound estimate of achievable mean ROUGE-1 can be computed. Let r_i^{DS} and r_i^{LL} denote the ROUGE-1 scores for DeepSeek and LLaMA-3 on sample i. The oracle score is:

$$\bar{r}^{\text{oracle}} = \frac{1}{N}\sum_{i=1}^{N} \max(r_i^{DS}, r_i^{LL})$$

Given that DeepSeek's maximum (0.634) exceeds LLaMA-3's (0.606), and that DeepSeek dominates LLaMA-3 on approximately 35% of samples (where its score exceeds LLaMA-3's), the oracle mean is bounded below by LLaMA-3's mean (0.497) and above by some value < 0.634. Conservatively, assuming DeepSeek outperforms LLaMA-3 on 30% of samples with an average gain of ~ 0.05 ROUGE-1 points, the oracle mean is approximately:

$$\bar{r}^{\text{oracle}} \approx 0.497 + 0.30 \times 0.05 \approx 0.512$$

This is a 3.0% relative improvement over LLaMA-3's mean—modest in aggregate but potentially material for the samples where DeepSeek's superior outputs provide qualitatively different financial insights.

5 Qualitative Output Analysis

The examples in this section compare model outputs for American Express (AXP) for the period February 25 – March 3, 2024. All three models received identical prompts.

Ground Truth (FinGPT-generated)

```
[Positive Developments]:
1. American Express's stock price has been on a consistent upward trend...
2. Certain financial ratios, such as ROE and Gross Margin, are strong...
```

DeepSeek (ROUGE-1: 0.501, ROUGE-2: 0.210, ROUGE-L: 0.227)

DeepSeek correctly adheres to the three-section output structure and provides a quantified directional prediction (*Up by 1–2%*), which boosts bigram overlap when FinGPT adopts similar language. It introduces a factually correct but lexically divergent reference to DJIA membership that was not emphasised in the ground truth, contributing to a slight ROUGE-1 penalty.

LLaMA-3 (ROUGE-1: 0.522, ROUGE-2: 0.183, ROUGE-L: 0.197)

LLaMA-3 achieves the highest ROUGE-1 on this sample (0.522). It explicitly references the cash ratio (0.32), aligning with the ground truth's emphasis on financial ratios and improving bigram overlap. A minor formatting deviation—`Positive Developments:` instead of `[Positive Developments]:`—represents the kind of instruction-following inconsistency that could be resolved through stricter prompt formatting in the fine-tuning loss.

Qwen (ROUGE-1: 0.445, ROUGE-2: 0.113, ROUGE-L: 0.215)

Qwen's output is structurally coherent but systematically penalised by Markdown bold syntax (`**text**`) absent from the ground truth. These tokens inflate sequence length and fragment bigrams, explaining the disproportionately low ROUGE-2 (0.113) relative to ROUGE-1 (0.445). Qwen's hedged directional call (*Neutral to Slightly Positive*) diverges from FinGPT's more decisive style, further reducing lexical overlap.

6 Model-Level Discussion

LLaMA-3. With mean ROUGE-1 of 0.497, ROUGE-2 of 0.180, and ROUGE-L of 0.251, LLaMA-3 is the dominant model across all aggregate metrics. Its instruction-following is robust: the three-section structure is consistently reproduced without explicit format enforcement in the training loss. Its lowest-performing sample (JNJ: 0.153) indicates ticker-level sensitivity that warrants further investigation.

DeepSeek. DeepSeek's high ceiling (ROUGE-1: 0.634, ROUGE-2: 0.326) is the highest recorded across all models and samples. However, a CV of 0.249 and 13% of samples below ROUGE-1 of 0.3 indicate a model whose output quality is strongly conditioned on narrative clarity. The bimodal distribution suggests that a subset of tickers—particularly those with complex or ambiguous news environments such as DIS—represent systematic failure modes.

Qwen. Qwen's near-uniform score distribution (CV = 0.100, range: 0.236–0.534) reflects consistent but stylistically misaligned outputs. The Markdown fine-tuning artifact is the primary driver of its underperformance; ROUGE-adjusted evaluations—stripping `**` tokens before scoring—would likely place Qwen closer to DeepSeek in absolute terms.

7 Limitations and Caveats

1. **LLM-generated ground truth.** ROUGE scores measure stylistic alignment with FinGPT's output format, not financial accuracy. A model that makes a correct but differently worded prediction is penalised.
2. **ROUGE does not capture factual correctness.** High ROUGE can coexist with directionally incorrect forecasts.

3. **Formatting sensitivity.** Qwen's Markdown tokens systematically inflate sequence length and reduce bigram overlap, biasing comparisons against it.
4. **No calibration against realised returns.** Directional predictions (Up/Down/Neutral) have not been validated against actual weekly stock price movements.
5. **Ticker-level heterogeneity.** Performance varies substantially by ticker, suggesting that aggregate metrics conceal important model-specific weaknesses.

8 Recommendations

1. **Deploy LLaMA-3** for production settings where consistency and average-case reliability are prioritised.
2. **Investigate DeepSeek's failure modes.** The 39 sub-0.3 samples likely represent identifiable failure patterns. Targeted fine-tuning or retrieval augmentation on these cases could substantially improve mean performance.
3. **Re-evaluate Qwen with format normalisation.** Pre-processing Qwen outputs to strip Markdown syntax before ROUGE computation would produce a fairer comparison. Suppressing bold syntax via instruction tuning would likely close the performance gap.
4. **Add directional accuracy metrics.** Evaluate Up/Down/Neutral predictions against realised weekly returns over the evaluation window to complement ROUGE with a financially grounded criterion.
5. **Implement a confidence-based router.** Given the JNJ anomaly and the non-overlapping performance distributions of DeepSeek and LLaMA-3, a per-sample routing mechanism—leveraging model confidence scores or narrative complexity proxies—could achieve oracle-level gains of $\sim 3\%$ in mean ROUGE-1.

A Notable Per-Sample Highlights

Table 7: Notable single-sample ROUGE-1 scores across all models.

Model	**Ticker**	**ROUGE-1**	**Note**
LLaMA-3 (best)	JPM	0.606	Highest LLaMA-3 score; strong bigram overlap
DeepSeek (best)	JNJ	0.634	Highest score across all models
Qwen (best)	WBA	0.534	Near Qwen ceiling
LLaMA-3 (worst)	JNJ	0.153	Same ticker as DeepSeek's best
DeepSeek (worst)	DIS	0.113	Complex narrative; likely failure mode
Qwen (worst)	CAT	0.236	Qwen floor exceeds other models' floors

The JNJ inversion—simultaneously DeepSeek's best and LLaMA-3's worst ticker—represents the most analytically significant finding in the per-sample data, and is the strongest existing evidence in favour of a ticker-aware routing strategy.

References

[1] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Proceedings of the ACL Workshop on Text Summarization Branches Out*, 74–81.