QuadCore ML task Marinos Report

Technical Choices

○ Technical choices made (models, fine-tuning strategies, data sources, etc.).

- Final model chosen: Bart-large-cnn
- Fine tuning strategies: LoRA (Low-Rank Adaptation of Large Language Models)
- Data sources: I scraped yahoo finance news articles for 50 different tickers (TSLA, AMZN, etc..) and getting the <summary> that exists there from metadata as my ground truth

○ Reasoning behind these choices (trade-offs considered, resource constraints handled).

Picking BART-Large-CNN

- We went with BART-Large-CNN because it's the go-to news summarizer in the Hugging Face world. It's got the encoder-decoder to handle articles and still spit out coherent, human-style summaries. Plus, everyone's written docs and tutorials for it, so we weren't reinventing the wheel. In terms of resource constraints, on a T4 GPU the performance was still in good shape. Other considerations and tests I did with T5-small, distilbart-cnn-12-6, pegasus xsum but I ended up choosing this.

Why LoRA Adapters

If we were to fully fine tune **BART** on a 16 GB T4 we could be OOM so quickly, therefore I chose to use LoRA to fine tune my system.

- Instead of updating all 400 M+ parameters, we only tweak a couple million.
- With 8-bit weights, and even a single T4 can go through multiple epochs.
- Changing a rank or learning rate takes a fraction of the time, so we could try different settings without waiting all day.

Why we picked LoRA instead of QLoRA

- **Resource Constraints**
  On a single T4 (16 GB), 8-bit+LoRA already let us train comfortably with a quite high batch size, so there wasn't the need to use Q-LoRA

- **Speed**
  Quick iterations, LoRA+8-bit gave us faster step-times than a comparable 4-bit setup, so our grid sweeps finished in a few hours instead of dragging overnight.

Scraping Yahoo Finance Summaries

Our goal was to create an as realistic as possible dataset including high quality summaries. The data size was quite high ~800-900 samples for 50 different unique tickers.

Trade-offs & Reality Checks

- **One T4 on Colab** = not a lot of headroom. LoRA was our only smooth path to speed and accuracy.
- **Dataset size**: We collected up to 100 articles per ticker, but settled on ~40 each so we didn't spend all day training.
- **Grid search**: We tested small adapter sizes (r=8,16,32) and learning rates (1e-4 vs 2e-4) first, then zoomed in on the best combo. Quick wins over "perfect" wins.

Getting all into account, the summariser we built was able to run on a single T4 GPU.

○ Description of challenges faced and how they were addressed.

1. Dataset Wrangling

- **Messy HTML & Boilerplate**
  Yahoo Finance pages are full of "Subscribe now!", "More From…", disclosures, ads, etc. We wrote regex-based filters and cut-off patterns to strip out that noise and stop when actual content ends.
- **Continuation Links**
  Some articles hide the last few paragraphs behind "Continue reading" buttons. We detect those, follow the link, scrape the extra `<p>`s, and append them—so we never miss the ending.
- **Minimum Length & Quality**
  We set an 80-word floor on the cleaned body and dropped anything shorter, giving us more robust training/evaluation examples.

2. Model Choice Trade-offs

- **Memory vs. Performance**
  BART-Large-CNN is 400 M parameters—awesome for summaries but brutal on GPUs. Full fine-tuning just wouldn't fit on a single T4 (16 GB).
- **LoRA fine tuner**
  Instead of updating all parameters, we only learn a few million low-rank matrices.

Together with 8-bit quantization, this slashed VRAM enough to run multi-epoch sweeps.

- **Why Not QLoRA?**
  We briefly tested 4-bit QLoRA, but the extra de-quant overhead and occasional encoder–decoder instability made our jobs harder. 8-bit + LoRA was faster, simpler, and rock-solid.

3. Resource Constraints on Colab

- **Single T4, Limited Time**
  We only had one T4 on Colab with a 12-hour session limit. That meant no huge datasets or interminable grid searches.
- **Smaller Data, Smart Splits**
  We fetched up to 100 articles per ticker because this was the most allowed by yahoo finance api, then capped at ~40 high-quality ones like include more than 100 tokens, etc.
- **Gradient Checkpointing & FP16**
  To squeeze every last gigabyte, we turned on gradient checkpointing and mixed-precision (FP16). This dropped peak memory by several GB without hurting stability.

4. Time constraints and improvements on the code base

Due to time constraints there are no:

- Logging modules (to log every step), try/except etc.
- No configuration files and parsing arguments for the scripts.
- No clear structure between files and functions and each module. As in the code are 4 different modules:
    - **Create dataset**
    - **fine tune**
    - **evaluation on dataset**
    - **inference and UI**

  and as a result, we might find duplicate code in some parts of the code base.

- While fine tuning, we didn't consider the chunking. So for a long article, we truncate it.

○ Evaluation results and discussion (quantitative and qualitative).

Quantitative:

800 sample dataset → Train / Eval

1. **Hold-out Evaluation (200 examples)**

    - We froze these 200 examples entirely until final testing—no peeking (**used only for final evaluation)**.
2. **Training pool (600 examples)**
    - **Inside each run**, we split that 600 into:
        - **Train**: 450 (75% of 600)
        - **Validation**: 150 (25% of 600)
    - At every epoch, we monitor the 150-example validation set to early-stop or pick the best checkpoint according to bert score because we wanted to focus on **semanticaly** better summaries.

Why ROUGE + BERTScore?

- **ROUGE** is the classic go-to for summarization: it checks how many words and n-grams your model gets exactly right, so you see if you're covering the same surface content as the human summary.

- **BERTScore** goes deeper: it measures semantic similarity via contextual token embeddings, so you catch cases where the model paraphrases perfectly or drifts off into unrelated territory.

- So we are measuring, did we get the same words, **and** did we keep the same meaning?

Model	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**BERTScore-F1**
Baseline	0.421	0.300	0.364	0.289
3rd Best FT	0.474	0.384	0.429	0.356
2nd Best FT	0.472	0.380	0.423	0.355
Best Fine-Tuned	0.478	0.388	0.432	0.357

Absolute Improvements vs. Baseline

Metric	**3rd Best Δ**	**2nd Best Δ**	**Best Δ**
ROUGE-1	+0.053	+0.051	+0.057
ROUGE-2	+0.084	+0.080	+0.088
ROUGE-L	+0.065	+0.058	+0.068
BERTScore-F1	+0.067	+0.066	+0.068

Key Takeaways

- Best fine-tuned model boosts ROUGE-1 by ~5.7 pts and ROUGE-2 by ~8.8 pts over baseline—solid improvements in both word- and bigram-overlap.
- BERTScore-F1 jumps from 28.9 % to 35.7 % (+6.8 pts), showing the LoRA adapters not only increase surface overlap but also produce semantically richer summaries.
- All three top LoRA configurations outperform baseline grid search and reliably find stronger models.

Overall, fine-tuning with LoRA on your Yahoo Finance data yielded clear, measurable gains across both n-gram and embedding-based metrics, confirming that your adapters successfully specialized BART-Large-CNN for finance-domain summarization.

```
experiments = [
    ( 8, 16, 4, 1e-4, 5),   # exp1
    ( 8, 16, 4, 2e-4, 4),   # exp2
    ( 8, 16, 8, 2e-4, 3),   # exp3
    (16, 32, 4, 1e-4, 5),   # exp4
    (16, 32, 4, 2e-4, 4),   # exp5
    (16, 32, 8, 2e-4, 3),   # exp6
]
# Format: (LoRA rank r, alpha α, batch size bs, learning rate lr, epochs ep)
```

Best -> exp6, 2nd Best -> exp5, 3rd Best -> exp2

Qualitative

Giving 3 examples to explain what happened in the summary and our prediction

1. ADP (Implied Volatility Surging)

Reference summary

> Investors need to pay close attention to ADP stock based on the movements in the options market lately.

Model prediction

> Investors in Automatic Data Processing, Inc. ADP need to pay close attention to the stock based on moves in the options market lately. That is because the Jun 20,

2025 $150 Call had some of the highest implied volatility of all equity options today.Implied volatility shows how much movement the market is expecting in the future. Options with high levels of implied volatility suggest that investors in the underlying stocks are expecting a big move in one direction or the other.

Analysis

- **Fluency**: Mostly smooth, though there's a missing space in "today.Implied", or need should be needs.
  Coverage: Hits the core "pay attention" sentence and explains implied volatility—good deeper context.
- **Faithfulness**: All details come directly from the body; no inventions.
- **Conciseness**: It's quite a bit longer than the one-line reference. It omits the Zacks-rank data but maybe over-explains volatility more than needed.

AT&T (Down 1.1% Since Last Earnings)

Reference summary

AT&T (T) reported earnings 30 days ago. What's next for the stock? We take a look at earnings estimates for some clues.

Model prediction

It has been about a month since the last earnings report for AT&T (T). Shares have lost about 1.1% in that time frame, underperforming the S&P 500. Will the recent negative trend continue leading up to its next earnings release, or is AT&T due for a breakout? Before we dive into how investors and analysts have reacted as of late, let's take a quick look at its most recent earnings report in order to get a better handle on the important drivers.

Analysis

- **Fluency**: Very natural, reads like a human tease.
  Coverage: Mirrors the reference's "30 days" hook and "what's next?" question—plus adds the S&P 500 comparison, which is factual and relevant.
- **Faithfulness**: No hallucinations; every statement matches the article.
- **Conciseness**: A bit wordier than the reference, but still focused on the same two questions.

P&G (Heavily Search Procter & Gamble)

Reference summary

P&G (PG) has received quite a bit of attention from Zacks.com users lately. Therefore, it is wise to be aware of the facts that can impact the stock's prospects.

Model prediction

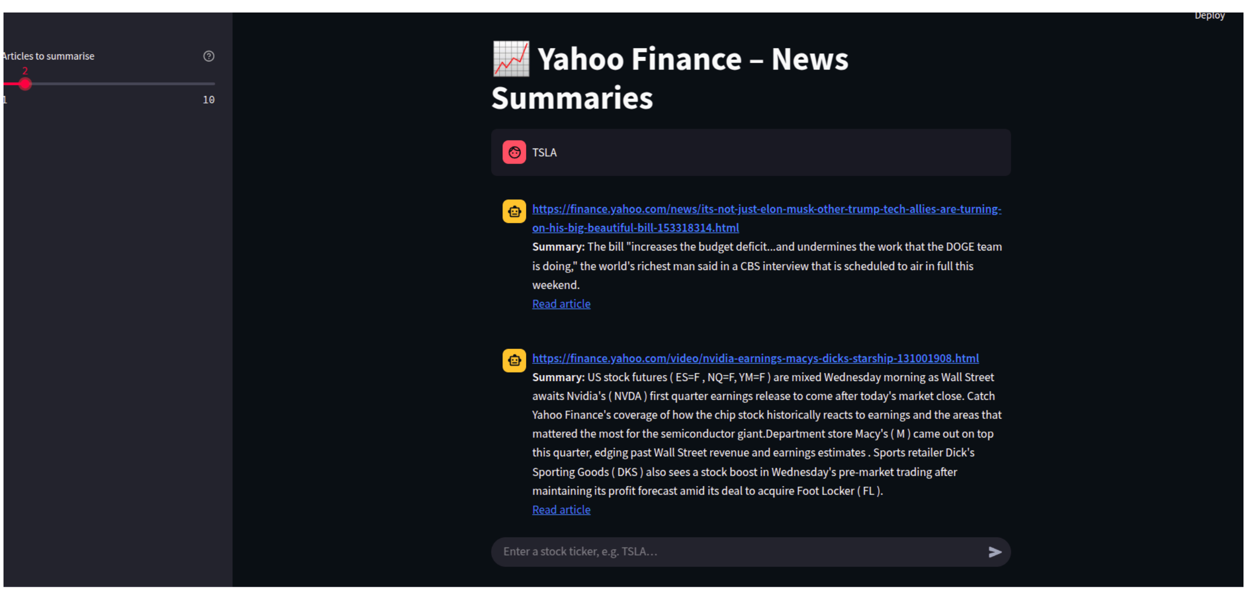
Procter & Gamble (PG) has been one of the most searched-for stocks on Zacks.com lately. So, you might want to look at some of the facts that could shape the stock's performance in the near term. Shares of this world's largest consumer products maker have returned -5.5% over the past month versus the Zacks S&P 500 composite's +3.8% change.

Analysis

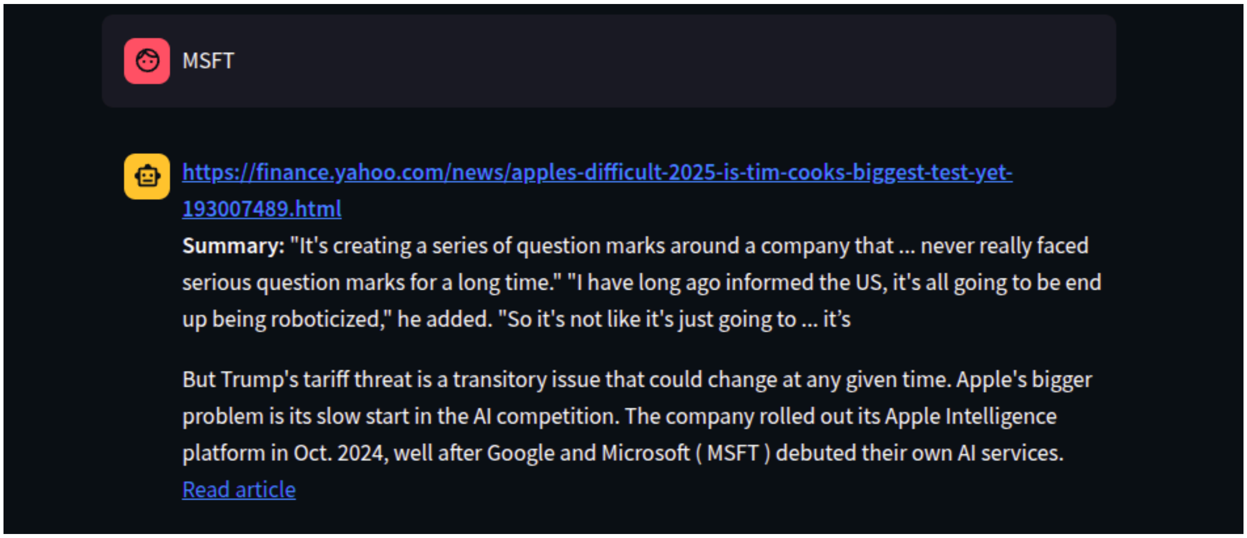
- **Fluency**: Clean and conversational.
- **Coverage**: Exactly captures the "attention" hook and advisement—then sensibly adds the month-return stat (correct and from the body).
- **Faithfulness**: No invented details; all pulled straight from the article.
- **Conciseness**: Slightly longer than the two-sentence reference, but that extra return figure is factual and arguably valuable.

○ Some examples of the summarized news results

Ticker: Tesla, 2 articles retrieved



Ticker: MSFT, 1 article retrieved:

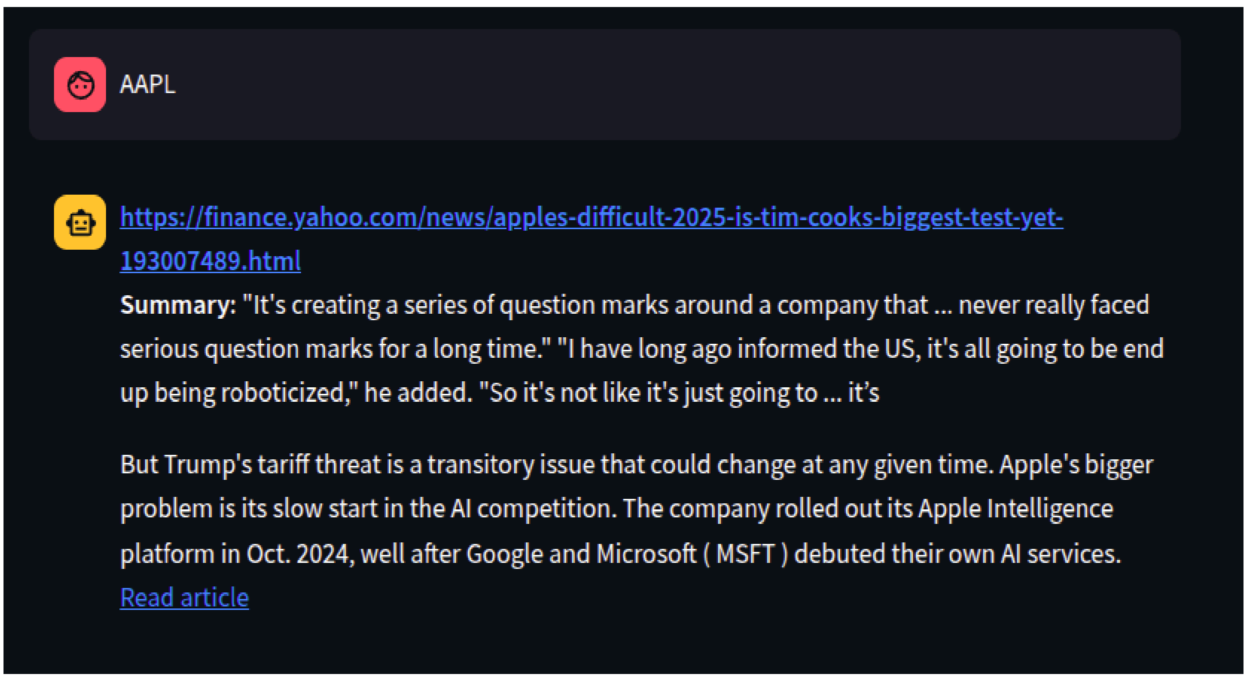


MSFT

https://finance.yahoo.com/news/apples-difficult-2025-is-tim-cooks-biggest-test-yet-193007489.html
Summary: "It's creating a series of question marks around a company that ... never really faced serious question marks for a long time." "I have long ago informed the US, it's all going to be end up being roboticized," he added. "So it's not like it's just going to ... it's

But Trump's tariff threat is a transitory issue that could change at any given time. Apple's bigger problem is its slow start in the AI competition. The company rolled out its Apple Intelligence platform in Oct. 2024, well after Google and Microsoft (MSFT) debuted their own AI services.
Read article

Ticker: AAPL, 1 article retrieved:



AAPL

https://finance.yahoo.com/news/apples-difficult-2025-is-tim-cooks-biggest-test-yet-193007489.html
Summary: "It's creating a series of question marks around a company that ... never really faced serious question marks for a long time." "I have long ago informed the US, it's all going to be end up being roboticized," he added. "So it's not like it's just going to ... it's

But Trump's tariff threat is a transitory issue that could change at any given time. Apple's bigger problem is its slow start in the AI competition. The company rolled out its Apple Intelligence platform in Oct. 2024, well after Google and Microsoft (MSFT) debuted their own AI services.
Read article