Financial Sentiment · Full FT vs. LoRA

A0328873N
Fang Yucheng

Github Link: https://github.com/CathySama/Assignment_lab3_Cathy

---

This report was written with the assistance of ChatGPT-5.
The AI was mainly used for:

- Drafting the report structure and wording.
- Refining explanations of algorithms and experimental results.
- Formatting content into clear Markdown style.

All code implementation, experiments, and results were conducted by the student.
The AI did not generate or modify the core experimental results.

---

1) Dataset & Motivation

Dataset. Plain-text `Sentences_*Agree.txt`, each line is `sentence + label` with robust parsing (supports `@`, tab, comma, or last-token labels). Labels map to `negative→0`, `neutral→1`, `positive→2`.
Motivation. Financial sentiment contains domain-specific cues (guidance up/down, margins, hedged phrasing), so we prioritize higher-agreement subsets (e.g., `AllAgree`, `75Agree`) to reduce label noise while enabling controlled noise ablations (`66/50Agree`).

---

2) Models & Fine-Tuning Strategies

- **Backbone:** `distilbert-base-uncased`.
- **Full fine-tuning (Full-FT):** All parameters trainable — strongest quality, higher memory/optimizer cost.
- **LoRA:** Low-Rank Adapters (rank `r=8`, `alpha=16`, dropout `0.05`) on attention projections (default `q_lin, v_lin`) with the backbone frozen — far fewer trainables.

Rationale. Full-FT forms a high-quality upper bound; LoRA targets efficient adaptation with small model deltas and faster iteration.

---

3) Experimental Setup

- **Tokenization:** DistilBERT tokenizer; truncation at `max_length=128`.
- **Splits:** Stratified — ~81%/9%/10% for train/valid/test.
- **Metrics:** Validation **Accuracy** & **Macro-F1** (`evaluate`).
- **Hyperparameters (shared unless noted):**
  - `learning_rate=1e-5`, `num_train_epochs=20`
  - `per_device_train_batch_size=16`, `per_device_eval_batch_size=32`

- Weight decay: **0.05** (Full-FT), **0.0** (LoRA)
- Mixed precision: `fp16=True` on CUDA (if available)
- Seed: `42`
- `load_best_model_at_end=True`, metric=`f1_macro`
- Eval cadence: **epoch** (Full-FT) vs **every 100 steps** (LoRA)

- **Trainer facts (from logs):** Total FLOPs — Full-FT **5.526e+14**, LoRA **5.621e+14** (≈ **-1.7%** lower).

---

4) Results (Tables & Plots)

4.1 Best Checkpoints (by Macro-F1)

Analysis. Full-FT peaks early and remains near-optimal; LoRA keeps improving and reaches its best at the final step — it likely benefits from slightly more steps or a warmup+cosine schedule.

Method	Best Step	Macro-F1	Accuracy	Eval Loss
Full-FT	460.0000	0.9665	0.9804	0.0841
LoRA	2300.0000	0.8612	0.9216	0.2646

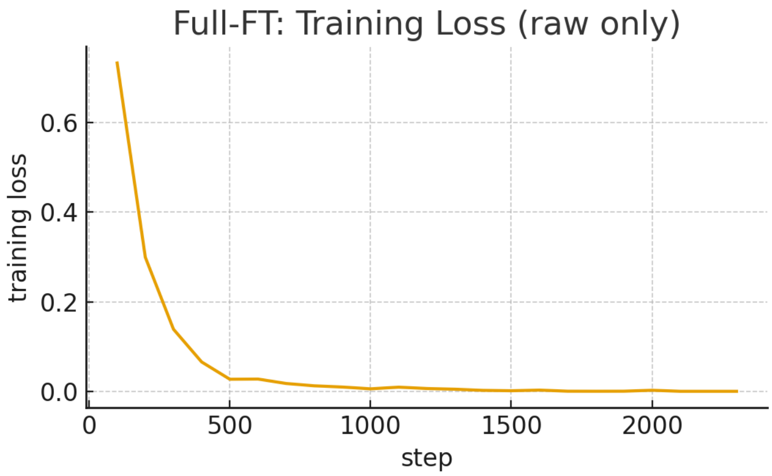
Analysis. Last-eval snapshot shows stability near the end; verify minority-class behavior with class-wise F1 & confusion matrices on the **test** split.

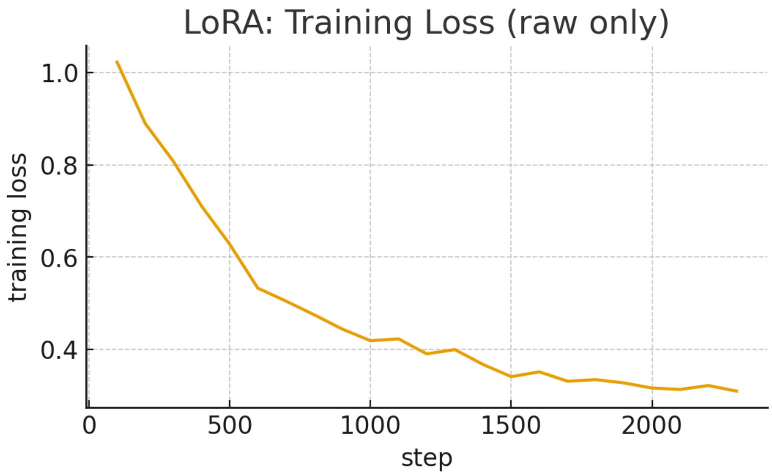
Method	Last Step	Macro-F1	Accuracy	Eval Loss
Full-FT	2300.0000	0.9564	0.9755	0.1296
LoRA	2300.0000	0.8612	0.9216	0.2646

Analysis（Macro-F1 vs step AUC）： Full-FT **0.9521**； LoRA **0.6751** (The higher the value, the better the average level of the entire training process)。

4.2 Training Loss (raw only)

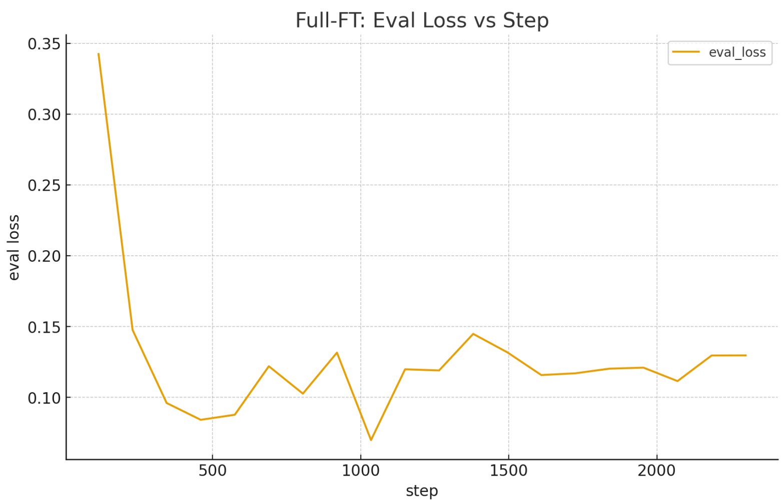
Analysis. Both runs descend; Full-FT approaches very small loss — together with its later eval-loss rebound this flags **mild overfitting** risk. LoRA's loss decreases steadily and aligns with its monotonic validation gains.

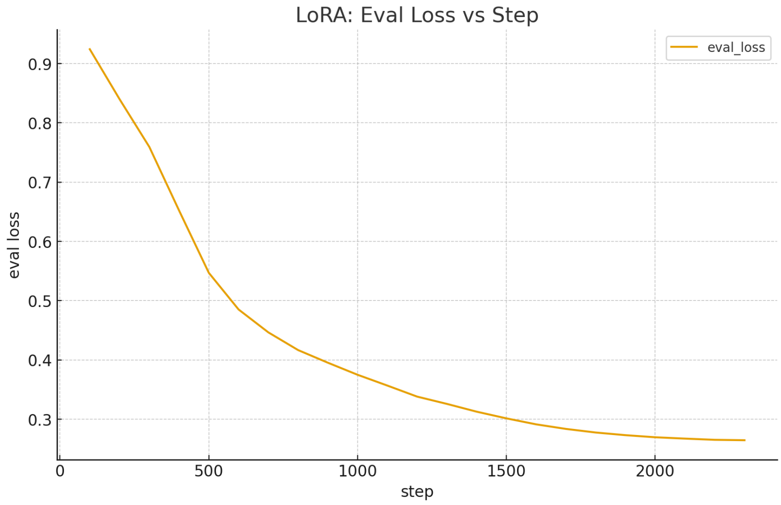




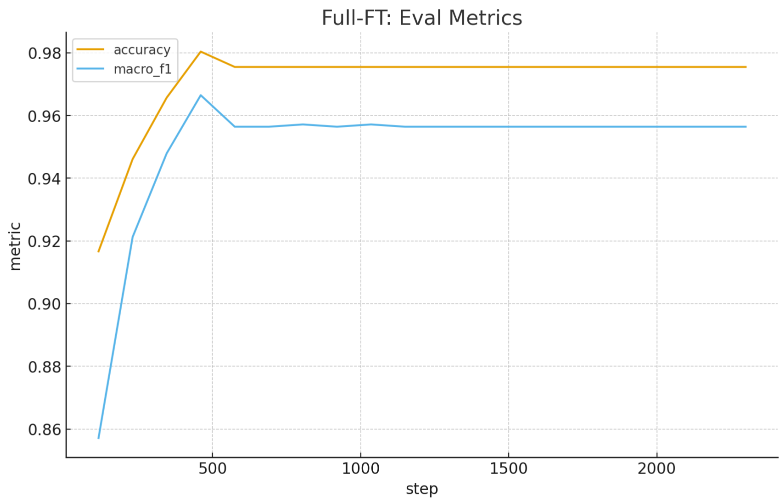
4.3 Validation Loss & Metrics

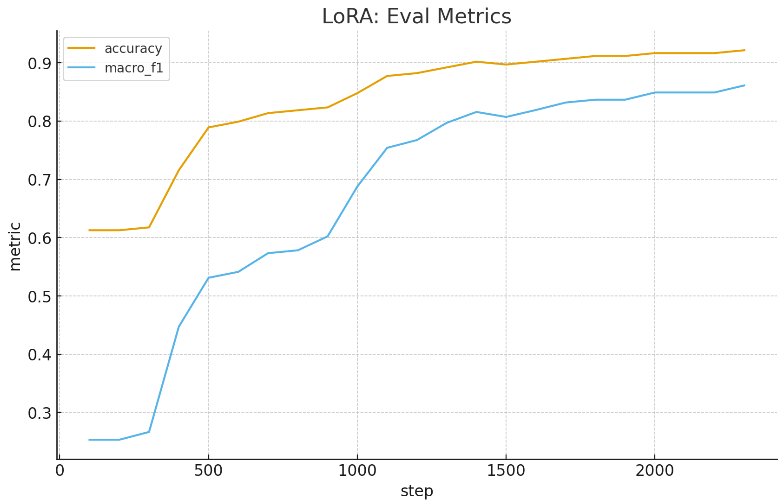
Analysis. Full-FT reaches a sharp optimum then slightly degrades; LoRA improves monotonically to the end — a strong hint that LR scheduling and/or more steps would help LoRA, while **early stopping** would help Full-FT.





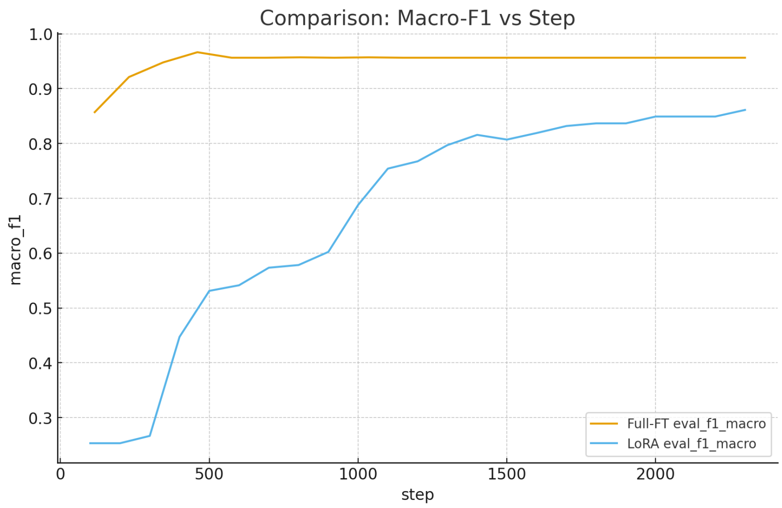
Analysis. Divergence between Macro-F1 and Accuracy late in Full-FT suggests minority-class degradation while overall accuracy remains high. Confirm with a per-class breakdown on the **test** set.

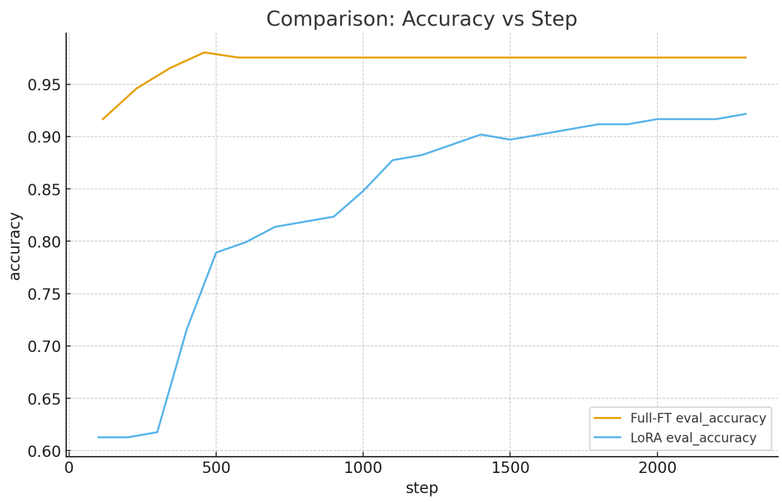




4.4 Head-to-Head Comparisons

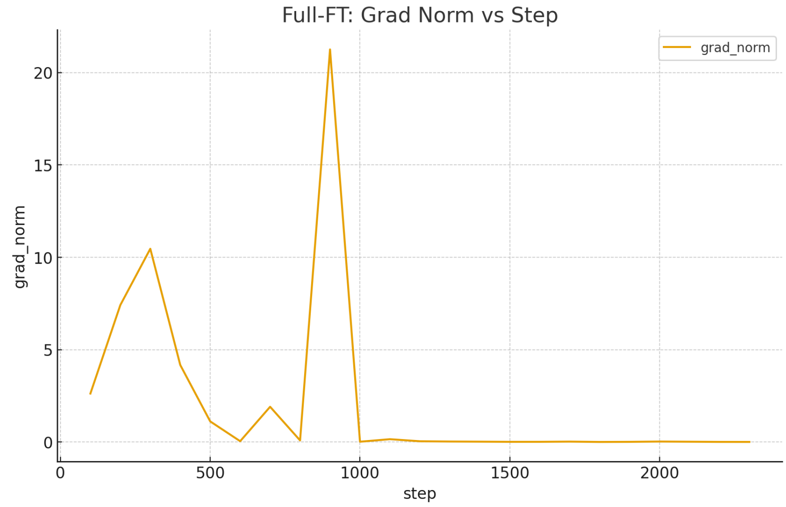
Analysis. Full-FT dominates peak performance; LoRA narrows the gap progressively. Eval-loss patterns support the same narrative (quality vs efficiency).

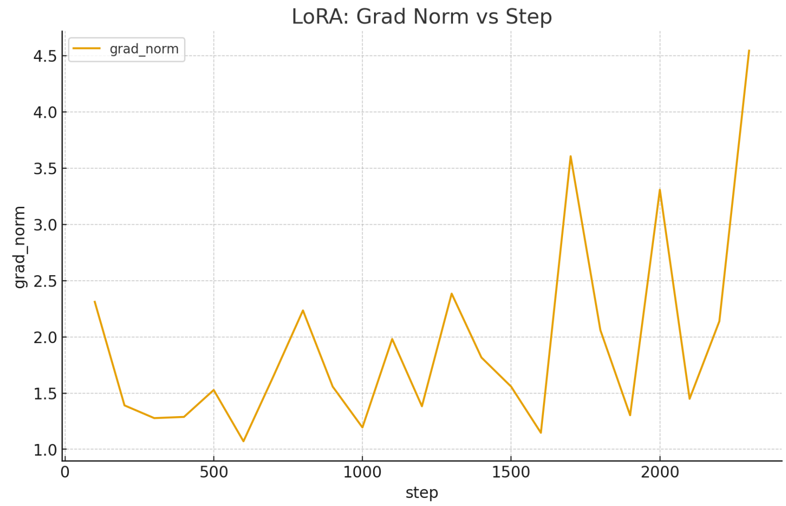




4.5 Optimization Diagnostics

Analysis. Full-FT shows larger gradient spikes; enable **gradient clipping (max-norm≈1.0)**. LoRA gradients remain moderate. Learning-rate traces are near-linear decay; adopt **warmup (5–10%) + cosine** to sharpen early learning (LoRA) and curb late overfitting (Full-FT).





---

5) Key Takeaways & Limitations

Takeaways.

- **Quality vs Efficiency.** Full-FT attains higher peak Macro-F1; LoRA is competitive with far fewer trainables and modest FLOP reduction.
- **Training dynamics.** Full-FT benefits from **early stopping** + **grad clipping**; LoRA benefits from **warmup+cosine** and slightly longer training.
- **LoRA capacity.** Expand targets (`q,k,v,out`) and consider higher rank (e.g., `r=16`) with mild dropout tuning.

Limitations.

- This report uses validation curves only; add **class-wise F1 & confusion matrices** on **test**.
- Noise robustness across `*Agree` subsets not quantified here.
- DistilBERT was selected for tractability; larger backbones may shift relative performance.
