Updated Week 4 Plan (assessor-ready)

🎯 Week 4 Goal

Produce a **final transformer benchmark (full training)** + **one controlled variant**, and update Chapter 4 so you can confidently explain progress in the assessor meeting.

🟢 MONDAY — Full transformer benchmark (final run) // DONE

Time: 2–4h
Coding

- Remove `max_steps=300` (full fine-tune)
- Keep settings fixed (same seed, same model, same split, same metrics)
- Evaluate on **valid + test**
- Save outputs **with "*full*" filenames** so you don't overwrite smoke test

Deliverables

- `results/transformer_full_valid_metrics.json`
- `results/transformer_full_test_metrics.json`
- `results/transformer_full_test_predictions.csv`
- `results/figures/cm_transformer_full.png`

✅ End-of-day: "I have my *real* transformer benchmark."

🟢 TUESDAY — One controlled variant (ONE knob only) // DONE

Time: 2–3h
Pick **one** (recommended: max length):

- **Variant A:** max_length **64** → **128** (best controlled comparison)
  or

- **Variant B:** learning rate **2e-5** → **3e-5**

Deliverables

- `results/transformer_var_valid_metrics.json`
- `results/transformer_var_test_metrics.json`
- `results/transformer_var_test_predictions.csv`
- `results/figures/cm_transformer_var.png`
- 5–7 lines: what changed + why (report-ready)

✅ End-of-day: “I can justify my transformer settings.”

🟢 WEDNESDAY — Final comparison table + short interpretation // DONE

Time: 1.5–2.5h
Tasks

- Update your comparison table to include:
  - Baseline (unigram)
  - Best classical variant (your best macro-F1 baseline)
  - Transformer full run
  - Transformer variant

Deliverables

- `results/model_comparison_week4.csv` (or update existing)

- Add a **new table** in Chapter 4 (or expand Table 4.X)
- 1 paragraph: what improved, what got worse, why (macro-F1 focus)

✅ End-of-day: “My results section looks like a dissertation.”

🟢 THURSDAY — Error analysis (DL vs baseline) // DONE

Time: 1.5–2.5h
Tasks

- Extract top 10–20 misclassified examples from:
  - baseline (you already have)
  - transformer full (new)
- Compare patterns:
  - Is “half-true” still a sink class?
  - Did “pants-fire” improve?
  - Any reduction in adjacent-label confusion?

Deliverables

- `results/transformer_error_examples.csv`
- 5 bullet insights to paste into Chapter 4

✅ End-of-day: “I can talk like a researcher, not just show numbers.”

🟢 FRIDAY — Assessor meeting pack (must-do) // DONE

Time: 1.5–2h
Make a **1–2 page meeting pack** (Google Doc is fine):

1. Project aim + what you're investigating (30 sec)
2. Baseline summary (numbers + what they mean)
3. Transformer summary (full run + variant)
4. 3 visuals:
    - label distribution
    - baseline confusion matrix
    - transformer confusion matrix (full)
5. What you'll do next (Week 5+)

Deliverables

- `assessor_meeting_pack.doc` (or Google doc)
- Repo committed + clean

✅ End-of-week: "I'm ready for the meeting."

🟡 SAT — Buffer

Finish anything unfinished.

🟡 SUN — Light review

Polish Chapter 4 wording + ensure all figures/tables are referenced in text.

Assessor meeting prep while working

Yes — we do both at the same time:

- Every Week 4 deliverable is **something you can show in the meeting** (results table, confusion matrices, final metrics, short plan).

If you paste your **current TrainingArguments cell**, I'll tell you the exact edits to:

- remove the cap
- save outputs under `_full_` names
- avoid overwriting smoke-test files
-