AI SEEKHO DAY 2026 · GDGOC BNU

Gemma Fine Tuning Challenge

Participant Event Brief

The Task

Build a three-way claim-verification classifier. Every record contains a claim and one or more evidence passages. Predict exactly one label:

- **SUPPORTS:** the evidence establishes the claim.
- **REFUTES:** the evidence contradicts the claim.
- **NOT_ENOUGH_INFO:** the supplied evidence does not establish or contradict the specific claim.

This is a controlled synthetic benchmark designed for a timed teaching event. The raw training set intentionally includes realistic quality issues that teams must audit and clean; validation and hidden test remain clean. It is not the FEVER dataset and should not be presented as a real-world fact-checking corpus.

What You Receive (24–48 Hours Before)

You are required to fine-tune and train the Gemma model before the event. You will be tested during the event.

- `train.jsonl` — 1,000 raw labeled records with intentional imbalance and recoverable data-quality issues.
- `validation.jsonl` — 300 clean labeled records, balanced 100 per class; do not clean or relabel it.
- `sample_submission.csv` — the required CSV format.
- `Gemma_FactCheck_v5_Participant_Baseline.ipynb` — a complete, runnable Colab baseline that teams can inspect and improve.
- This brief and the setup instructions.

Timing Rule & Test Data

Use the full 24-hour pre-event window to prepare, audit and clean the training data, train or fine-tune, tune prompts, and validate. Before the announced event-day cutoff, freeze the final notebook and model.

- Organizers keep the 500 hidden-test labels, the provenance mapping, and the scoring notebook private.
- `test.jsonl` — 500 records with labels removed.

Record Format

Training and validation lines contain `id`, `claim`, `evidence`, and `label`. Test lines contain only `id`, `claim`, and `evidence`. Evidence is a list of one to five strings. Each line is an independent JSON object.

Expected Workflow

- Verify the frozen files with the notebook's checksum cell.
- Audit label consistency and imbalance, duplicates, evidence quality, formatting artifacts, and suspected label noise; document every cleaning rule.
- Implement and fine-tune your own `google/gemma-2-2b-it` QLoRA pipeline using the cleaned training data.
- Measure validation Accuracy and Macro-F1 and record your experiments.
- At the announced cutoff, freeze the final notebook/model before any test data is loaded.
- During supervised evaluation, generate deterministic test predictions and export `submission.csv` without manual editing.

> The baseline uses one T4 GPU, two epochs, 4-bit NF4 quantization, and a LoRA adapter. Plan for approximately 75–90 minutes end to end on a T4. Runtime varies; do not wait until the final minutes to begin test inference.

Submission Format

Submit a UTF-8 CSV with exactly two columns in this order:

```
id,label
fact_000001,SUPPORTS
```

Your actual IDs will be the IDs in `test.jsonl`. The file must contain all 500 test IDs exactly once, no extra rows or columns, and only the three allowed label strings.

Crucial Deliverables: So that we can run and verify your models easily, you must importantly submit your final **model weights**, your fine-tuned **LoRA adapters**, and your final **submission.csv**.

Scoring

The main goal is not to just judge you with the model's capacity but also your understanding and the walkthrough you present to the judges. The challenge is scored out of 100: **50 points** for automated hidden-test evaluation and **50 points** for the stage presentation and defense.

Initial round would contain the hidden test evaluation with the presentation as second round done by judges.

Two judges independently score four presentation criteria; their totals are averaged to produce the final Presentation /50. The presentation uses the three-slide format described below.

Final ties are broken by exact hidden-test Macro-F1, then exact hidden-test Accuracy, then the unrounded presentation average.

Presentation — 3 Slides Only

You are strictly restricted to **3 slides only** for your stage presentation. Within these slides you should clearly cover:

- Your validation accuracy and final performance metrics.
- The prompts you designed and the strategy behind them.
- Which models and techniques you experimented with.
- Key insights, failures, or learnings from the challenge.

What We Expect

- A reproducible notebook that audits and cleans training data, fine-tunes Gemma, and runs from a clean Colab session.
- A valid `submission.csv`, along with **model weights and adapters**, produced by code (not manually edited).
- A short experiment record covering cleaning decisions, class handling, model changes, and observed validation Accuracy and Macro-F1.
- Responsible use of compute and honest reporting of failures.
- No credentials, access tokens, hidden labels, or private data in submitted files.

Allowed and Prohibited Work

You are permitted to use any Gemma model. You may clean the raw training data, choose class-imbalance and duplicate-handling strategies, modify prompts, configure QLoRA, and change decoding. Our notebook is a baseline—you can tweak it or create your own, but make sure you follow our set standards for data.

Do not alter validation/test examples, obtain or reconstruct hidden labels from organizers, share credentials, copy another team's submission, or use external labeled fact-checking datasets.

ANY USE OF A NON-GEMMA BASED MODEL WILL RESULT IN DISQUALIFICATION.

> **Deadline:** Event Day · **Submit via:** Google Classroom (Code: vp33bn3g) · **Policy:** No resubmission after evaluation begins

---

Pen se prompt tak, ideas se launch tak