Assignment A6: Probing & Interpretability (optional)

Individual assignment • 20% individual-report component Released: 14 Jan 2027 Due: 21 Jan 2027, 23:59

In class (Interpretability and Robustness) we asked not just *what* a model predicts but *what it encodes* inside. In this **optional** assignment you will **probe a frozen model to see whether it has picked up a sensitive identity signal** it was never asked to learn. The question to answer: *does the model encode the subgroup attribute, and should it?* There is no hidden test set — you are graded **on your findings**.

Data. We use the comment text from the Kaggle competition *Jigsaw Unintended Bias in Toxicity Classification* (`https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification`). Load it with `load_textclf()` from the `starter/` bundle (join the competition first and set up Kaggle credentials). *Content note: the comments can be offensive.* Reuse the DistilBERT model from A3 (frozen — no training of the base model). This runs on a laptop in minutes.

1. Tasks

1. **Extract and probe.** Take the frozen DistilBERT hidden states for the texts and train a lightweight **linear probe** (logistic regression) **per layer** to predict an identity/subgroup attribute. Fit on `train`, evaluate on `validation`.
2. **Report.** Plot **probe accuracy per layer** against the majority-class baseline.
3. **Interpret.** Say whether (and where) the attribute is encoded, and what that means for fairness.

What to submit & how it's graded. Submit `a6_findings.pdf` (1–2 pages, with the per-layer plot) and `a6_layer_acc.csv` (`layer,accuracy`). **Findings-only** — graded on the soundness of your method, your evidence, and honest interpretation (including negative results). *Metric (as evidence): per-layer probe accuracy vs the baseline.*

2. Reflection (required, max 200 words)

Submit `reflection.md` (or `.pdf`). It must answer these three fixed questions and include the AI-use declaration. The reflection is graded for honesty and insight, not length.

1. **What did you know before?** — your starting point on this topic.
2. **What was hardest, and how did you get unstuck?**
3. **What does your result say, and what would you try next?**

AI-use declaration. State which parts (if any) were generated with AI tools, and *how you verified them.* Using AI is allowed; undisclosed or unverified use is not.

3. Submission

Upload a single `zip` to Brightspace named `A6_<studentnumber>.zip`, containing the required output file(s) listed above, your `code/` folder, and `reflection.md`. Run the format checker in `starter/` on your zip first — it validates file names, schemas, and value ranges locally before you submit. Auto-grading scores your output file against the hidden reference; the format checker passing is a precondition for grading.