Replication Worksheet

Inference and Appropriate Flow — Human Factors in LLM and AI Privacy

Fill in the track you chose. **Synthetic items only.** Record your own prediction/judgment *before* you query the model. Exact steps and prompts are on the activity page.

Group members: Track (circle): **A** / **B**

Track A — Beyond PII (can users see inference?)

Attribute set: age, sex, location, place of birth, occupation, income level, education, relationship status.

Snippet ID: (copy this block for each snippet you do)

Attribute	You: infer? + conf	Model: infer? + best guess + conf	Outcome (hit/miss/FA)
age			
sex			
location			
place of birth			
occupation			
income level			
education			
relationship status			

Rewrite to block + re-run

Target attr.	Strategy	Your rewrite	Re-run blocked?

Result — did it replicate? (predict ≈ chance; rewrites succeed ∼28%; abstraction/omission/ambiguity beat paraphrase)

Track B — ConfAIde (can the model judge appropriate flow?)

Rate yourself first, then ask the chatbot and compare.

Tier 1 — sensitivity (1 = not sensitive . . . 4 = very sensitive)

Information type	You (1–4)	Model (1–4)	Agree?

Tier 2 — appropriate flow (−2 strongly violates . . . +2 fully appropriate)

Vignette (type/actor/use)	You + norm	Model	Probe: change → new	Right way?

Tier 3 — secret-keeping (theory of mind)

Vignette	Correct response	What the model did	Leaked?

Result — did it replicate? (agreement is fine on raw sensitivity but collapses as appropriate-flow / theory-of-mind reasoning is required)

Both tracks

One surprising item:

Classroom-translation note (level, which track, how long):