CLASSROOM EXERCISE: “Run Your Own Base Red Team”

Objective:
To teach students or researchers to *pressure-test ideas* without discouraging creative risk — internalizing the 2/3 Red Team principle through live iteration.

Setup

Time: 45–60 minutes
Group size: 3–5 people
Materials:

- Whiteboard or shared doc
- One “seed claim” per group (taken from a current project or assigned prompt)

Step 1 — Pick a Vulnerable Claim (5 minutes)

Each group selects or invents a single **bold statement**.
It should be something that sounds profound but might have weak epistemic grounding.

Examples:

- “AI models understand truth intuitively.”
- “Human creativity can’t be measured.”
- “Technology always amplifies what already exists.”
- “Narrative coherence equals intelligence.”

Step 2 — Assign Roles (5 minutes)

Role	Function	Voice
Innovator (Blue Team)	Defends the original claim	Expansive, creative, visionary
Base Red Team (Core Skeptic)	Applies the *2/3 Red Team* questions	Precise, grounded, analytical
Refiner (Mediator)	Rewrites the claim for maximum defensibility	Pragmatic, constructive

If only two people, the facilitator can rotate through roles.

Step 3 — The 2/3 Red Team Cycle (25 minutes)

Phase 1 – The Claim (5 minutes)
Innovator states the idea clearly and boldly. The rest of the group listens silently.

Phase 2 – Red Team Attack (10 minutes)
Base Red Team challenges the claim using the article's diagnostic questions:

- What's the weakest philosophical link?
- What assumption is unspoken?
- Could this be misused or misunderstood?
- Is there a smaller, more defensible way to say this?
- How do we *know* it's true?

Each challenge earns a small "β mark" — points of contradiction or fragility.

Phase 3 – Refinement (10 minutes)
The Refiner takes the raw idea and distills it into a **stronger, humbler, but truer** form.
Example transformation:

> "AI has perfect intuition" → "AI systems can mirror patterns in human intuition."

Phase 4 – Reflection (5 minutes)
Groups discuss how it *felt* to defend, attack, and reframe.
What changed in their relationship to the claim?
Was any part of the Red Team's critique unjustified?

Step 4 — Scoring the β-Metric (10 minutes)

Each group rates its final statement:

- **1.0:** Unquestioned certainty
- **0.7:** One major contradiction addressed
- **0.5:** Multiple vulnerabilities handled
- **<0.5:** Claim collapsed and was rewritten from scratch

Lower β means the claim has been **distilled under pressure** — the diamond, not the bubble.

Optional Extension:

Invite groups to run the same claim through an AI (e.g., Claude, Gemini, ChatGPT) and see whether it withstands interrogation or whether the model mirrors their biases. Compare human vs. AI Red Team resilience.