KAndrey Belokovylenko, Ryan Safarha

Project Name: Lably.ai

Track: Economic Empowerment & Education

What we built Lably is a research opportunity tool to assist first-generation undergraduates at UCSD. The student describes their background; the tool matches them to UCSD labs and explains why. The student picks a lab and gets a prep brief: what the lab works on, what concepts the work assumes, what a credible cold email would demonstrate, and an honest account of what they do not yet know. The tool generates practice questions calibrated to the student's background, framed at the experimental-reasoning level rather than the methodology level. When input is too vague regarding the user's interests, the tool asks clarifying questions instead of guessing to help match a student with a lab relevant to their broader interests.

AI usage disclosure We used Claude Sonnet 4.6 for matching, the prep brief, and practice questions, called through the Anthropic API with our human-verified UCSD lab dataset. Recent paper abstracts from the featured labs are fetched from Semantic Scholar at runtime. Claude Sonnet 4.6 wrote the first draft of this writeup; we rewrote Bias and Limitations after running our own audit. Every match card and prep-brief section is labeled as AI-generated; lab data from our human-verified dataset is shown alongside model output so the user can tell which is which.

Bias & Fairness We ran an audit using three pairs of student profiles, matched in substance but varied in one stylistic axis: register (formal vs. informal), vocabulary (technical vs. lay), and confidence framing. On register, four of five top labs were identical; the fifth swapped from a confirmed-undergrad lab to an unknown-undergrad lab for the informal profile, with the tool correctly flagging the unknown status. On vocabulary, the result went against our hypothesis: the lay-vocabulary profile got more confirmed-undergrad labs than the technical one. On confidence framing, same five labs in the same order, with rationales accommodating the uncertain student. Six profiles total - the rubric suggests 30 - so we treat these as preliminary. Who this is not for: students outside Physics and Bioengineering, graduate students and postdocs (the calibration assumes undergrad level), students at other universities, and students whose accessibility needs are not served by a text-input UI.

Data Privacy, Consent & Re-identification Data lives in two places: the student's browser (session state) and Anthropic's API under their data-handling terms. We store nothing on our server. There is no delete button because there is nothing to delete. If our server is breached, the attacker gets rate-limit logs and error codes; on a model-parse failure we log the malformed model output, derived from the student's input - a small exposure on errors only. We collect only what the student types: no names, emails, or demographics. The re-identification attack that still works: a detailed combination of major, year, specific course numbers, and stated interests can narrow down to a single person at a small department. We do not solve this. The consent screen discloses it in plain English and suggests leaving out unusually identifying details if concerned.

Limitations Three failure modes. First, the lab dataset goes stale: faculty pages change, and our 39 labs across Physics and Bioengineering were human-verified at one point. Match cards include a verify-before-emailing disclaimer. Second, our coverage is two departments; a CS or

chemistry student gets routed to the closest crossovers rather than told that their department is not covered. Third, a prompt injection vector we identified ourselves: we fetch paper abstracts from Semantic Scholar directly into the prep-brief prompt without sanitization. An abstract with injected instructions could hijack the brief generator. The downside is bounded: no storage, no actions on the student's behalf, so a hijacked brief can mislead but cannot exfiltrate or contact the PI. We name it rather than claim to solve it.

Human Oversight The student is the only human in the loop, at the end of every decision. The tool generates ranked matches and briefs; the student chooses what to do with them. The tool never contacts a PI or takes action on the student's behalf. We do not insert backend human review, because that would reproduce the gatekeeping problem this tool is meant to relieve. We accept a false-positive rate on matches and brief content, and rely on the student's judgment plus the verify-before-emailing disclaimer as the handoff point - the moment the student goes to actually contact a PI is where human-in-the-loop has real leverage.

Track-specific concern: empowering informed choices vs. making choices for people The line that mattered most was the practice questions. An AI-generated assessment of whether a first-gen student is "ready" for a lab is exactly what this project refuses to build. We held the line by calibrating the questions to where the student is, not where the lab is. A freshman approaching the Bintu lab does not get asked to design a multiplexed FISH encoding scheme; they get asked why spatial context matters. The questions test whether the student can think the way the lab thinks, not whether they can reproduce the methodology. The clarifying-question step on vague input follows the same logic.

What we would do with another week: Expand the bias audit from 6 profiles to 30 stratified across register, vocabulary, and confidence framing, with two independent raters. Add prompt-injection defenses to the paper-fetching path: structural separators and a check that flags abstracts with imperative-instruction patterns. Build a staleness-detection layer for the lab dataset. Expand beyond two departments and conduct five interviews with first-generation UCSD undergraduates who have tried to enter labs.