Senior Engineer - AI Evaluator

Location
Remote

Employment Type
Full time

Location Type
Remote

Department
G2i Eng Team

Overview Application

Senior AI Evaluator (Codex / Claude Code)

Contract | $100–$200/hour | 10–20 hrs/week | Start ASAP (through mid May)

We're looking for **highly experienced software engineer** to help evaluate the quality of interactions with modern coding agents such as OpenAI Codex and Claude Code.

This is not a traditional engineering role.

You won't be writing production code.
You'll be evaluating something harder: **whether the model *thinks* like a great engineer.**

What This Role Actually Is

You will assess how AI coding agents behave in real-world scenarios — focusing on:

- Whether the **response makes sense**
- Whether the **preamble and reasoning are useful**
- Whether the **output reflects strong engineering judgment**
- Whether the **interaction feels right to an experienced developer**

This role is about **engineering taste** — not syntax correctness.

What You'll Be Doing

- Evaluate AI-generated coding interactions end-to-end
- Judge whether outputs are:
    - Useful
    - Correct (at a high level)
    - Aligned with how a strong engineer would think
- Assess the **quality of explanations and reasoning**, not just code
- Distinguish between different levels of response quality (e.g. what makes something a *2 vs 4*)
- Provide clear, opinionated feedback on:
    - What worked
    - What didn't
    - What felt "off" or misleading
- Help define what *great* looks like when interacting with tools like Cursor

What We Mean by "Taste"

We're specifically looking for engineers who can answer questions like:

- *Does this feel like something a strong engineer would actually say?*
- *Is this explanation helpful, or just technically correct?*
- *Is the model guiding the user well, or just dumping output?*

- *Would this interaction build or erode trust?*

You should be comfortable making **subjective but rigorous judgments**.

Who You Are

- Staff / Principal-level engineer (or equivalent experience)
- Strong background in one of the below:
    - TypeScript / JavaScript
    - Python
- Hands-on experience using:
    - OpenAI Codex
    - Claude Code
    - Cursor
- Deep familiarity with modern AI-assisted dev workflows
- Able to evaluate code **without needing to fully execute or deeply review every line**
- Comfortable giving **direct, opinionated feedback**
- High bar for what "good engineering" looks like

Nice to Have

- Experience with tools like Cursor or similar AI-first IDEs
- Prior exposure to prompt design or evaluation workflows
- Experience mentoring senior engineers or defining engineering standards

Engagement Details

- **Rate:** $100–$200/hour
- **Hours:** ~20 hours+
- **Duration:** Through early May (with possible extension)

- **Start:** ASAP
- **Process:**
    - Take-home evaluation exercise
    - Technical Test with the client directly

Apply for this Job