# Lesson Plan 2: Evaluating AI Outputs for Debugging and Learning

**Duration:** 45–55 min | **Audience:** CS1/CS2 students | **Format:** Pair-based, hands-on

**Objectives**

By the end of this module, students will:

- Critically evaluate AI outputs rather than accept them at face value
- Identify differences in AI model responses and reasoning quality
- Develop verification habits when using AI for debugging
- Practice life-long learning competencies: questioning assumptions, seeking multiple sources, refining evaluation skills

## Materials

Access to at least two different AI chatbots (ChatGPT, Claude, Gemini, Copilot, DeepSeek); a buggy code snippet (instructor's choice of language); Comparison Worksheet; optional Evaluation Framework handout.

## Lesson Outline

### Opening (5 min)

Frame the session: tools change constantly; the durable skill is evaluating them. Ask:

- *Have you ever gotten code from AI that didn't work?*
- *How do you know if an AI's answer is correct?*

**Key message:** *Life-long learners don't stop at the first answer — human or AI.*

### Part 1 — Present the Buggy Code (5 min)

Display a short program with a realistic bug (e.g., a function that crashes on empty input due to unchecked array access). Walk through what the bug is and why it matters in a real program. Frame the upcoming activity as practicing evaluation skills that will outlast any specific AI tool.

### Part 2 — Generate AI Outputs (12–15 min)

Pair students: **Prompter** (asks the AI) and **Evaluator** (documents responses). Switch roles for the second model.
Provide two prompts on screen:

- **Standard:** *"This function crashes when the user enters an empty input. Explain why and suggest a fix."*
- **Better (humility-based, from Module 1):** *"I'm a beginner CS student. This function sometimes crashes. Can you explain what's wrong, why it happens, and how to fix it safely — step by step, without advanced jargon?"*

Each pair: Runs one prompt through their first AI model; documents the response | Switches roles and runs the same prompt through a second AI model | Reads both responses together.
Circulate and ask: *What criteria would you use to evaluate these?*

### Part 3 — Compare and Analyze (15–18 min)

**Pair analysis (8–10 min):** Students complete the Comparison Worksheet, judging each AI response on accuracy, clarity of explanation, teaching value, and any concerns.
**Whole-class discussion (7–8 min):** Facilitate with:

- *How many pairs got nearly identical solutions?*
- *Which model best explains WHY the crash happens?*
- *Did any AI suggest a fix that wouldn't actually work? How would you catch that?*
- *Which explanation helped you understand, not just fix, the problem?*

**Key insight:** The skill of comparing sources, identifying quality, and making informed judgments transfers to any future tool; that's what life-long learners do.

National Science Foundation

*This material is based upon work supported by the National Science Foundation under Grant No. 2111097*

### Part 4 — Build Your Evaluation Framework *(Optional, 5–7 min)*

Co-create with students a checklist they can apply to any learning resource (AI, tutorial, documentation, Stack Overflow):

- **Verify:** Can I test it?
- **Understand:** Can I explain why it works in my own words?
- **Compare:** How does this differ from other sources?
- **Question:** What edge cases or limitations should I consider?
- **Document:** What did I learn that I want to remember?

Students add this to their AI-use log from Module 1.

### Closing (3–5 min)

Brief reflection (oral or written):

- *How does critical evaluation connect to life-long learning?*
- *What evaluation criteria will you use next time you learn something new?*

**Key takeaway:** *Seek multiple sources · Test and verify—don't accept claims at face value · Focus on understanding principles, not memorizing tools · Continuously refine your judgment.*

## Adaptable Variations

- **Shorter (30 min):** Instructor pre-generates AI outputs and distributes them; skip generation, go straight to comparison.
- **Alternative bugs:** Off-by-one loops, type-conversion issues, logic errors in conditionals, inefficient algorithms.
- **Different language:** Any beginner-appropriate buggy snippet works — the comparison skills are language-independent.

### Model Comparison Worksheet

Name: Date:

**AI Models Tested:**

| | |
|---|---|
| Model A: | Model B: |

**The code:**

| |
|---|
| (Added by the instructor) |

**The prompt you used:**

**SECTION 1: Document the Responses**

| |
|---|
| Model A Response (copy key parts): |

| |
|---|
| Model B Response (copy key parts): |

### Model Comparison Checklist

| **Evaluation Criterion** | **Model A: _____** | **Model B: _____** | **My Choice** |
|---|---|---|---|
| 1. Is the fix correct? | ☐ Yes ☐ No | ☐ Yes ☐ No | Model A |
| 2. Explains WHY the bug happens? | ☐ Yes ☐ No | ☐ Yes ☐ No | Model A |

National Science Foundation

*This material is based upon work supported by the National Science Foundation under Grant No. 2111097*

| | | | |
|---|---|---|---|
| 3. Explains in a beginner-friendly language? | ☐ Yes ☐ No | ☐ Yes ☐ No | Model A |
| 4. Can I understand and explain the fix? | ☐ Yes ☐ No | ☐ Yes ☐ No | Model A |
| 5. Does it teach transferable concepts? | ☐ Yes ☐ No | ☐ Yes ☐ No | Model A |
| 6. Is the response lengthy? | ☐ Yes ☐ No | ☐ Yes ☐ No | Model A |

Which model would you trust more for learning in general? Why?

What evaluation criteria do you like most?

What evaluation criteria would you add to the list?

National Science Foundation

*This material is based upon work supported by the National Science Foundation under Grant No. 2111097*