Annotation Academy

AI Evaluator Certification

Level 1: Foundations

24 modules 54+ hours

Certification integrity: every certificate is tied to a verified identity (Stripe Identity) and earned through a proctored final exam. Anyone can confirm a certificate by its unique verification code.

01. Core Competencies & Mental Models

Master the 4 core competencies of AI evaluators and learn cognitive load management strategies for consistent evaluation work.

1 h

02. Modality-Specific Assessment

Evaluate AI outputs across text, code, image, and multi-modal modalities. Recognize modality-specific quality cues and failure patterns.

4 h

03. How AI Training Works

Learn about RLHF, the three stages of AI training, different evaluation task types, and the HHH framework.

2 h

04. Core Evaluation Skills

Master comparison methodology, identifying unrateable prompts, and reading & interpreting rubrics effectively.

2 h

05. Evaluation Dimensions

Learn to evaluate AI responses across key dimensions: accuracy, hallucination detection, instruction following, and failure type hierarchy.

2 h

06. Safety Fundamentals

Understand AI safety principles, harm categories, and how to evaluate responses for safety compliance.

2 h

07. Prompt Engineering & Writing

Master prompt decomposition, quality criteria identification, and structured evaluation approaches for different prompt types.

3 h

08. Justification Writing

Write clear, defensible justifications using the SPEC framework. Master evidence-based reasoning for evaluation decisions.

3 h

09. Data Annotation Fundamentals

Master annotation taxonomies, labeling guidelines, and quality control for structured AI training data.

3 h

10. Ideal Response Description & Rubric Properties

Create Ideal Response Descriptions through prompt analysis. Master criterion properties: atomicity, self-containment, objectivity, specificity, weighting, and golden responses.

4 h

11. Atomicity Bootcamp

Recognize and decompose non-atomic criteria. Write atomic criteria from scratch and apply atomicity reliably across any task type.

4 h

12. Instance-Specific Mastery

Write criteria tied to the specific instance, not to generic task templates. Surface what this prompt requires that others would not.

3 h

13. Self-Containment

Write criteria a reviewer can apply without consulting external materials. Eliminate hidden dependencies and implicit context.

2 h

14. Objectivity & Thresholds

Convert subjective judgments into concrete thresholds. Build criteria two reviewers would score the same way.

2 h

15. Applying Rubrics in Practice

Apply rubric properties to real tasks. Master modality-aware evaluation, platform patterns, speed drills, and integration practice.

3 h

16. Platform Rubric Patterns

Recognize the rubric patterns each major platform expects. Adapt your rubric writing to platform conventions without losing rigor.

2 h

17. Rubric Speed Drills

Build the muscle to draft sound rubrics under platform-realistic time limits. Cut deliberation time without dropping rubric quality.

2 h

18. Integration Practice

Combine rubric writing, evaluation, and justification on end-to-end tasks that mirror real platform workflows.

3 h

19. Citation & Fact-Checking Skills

Find and format citations rapidly. Master source reliability evaluation and platform citation formats.

2 h

20. Source Reliability

Distinguish strong, weak, and unreliable sources. Apply a repeatable check to every citation before you accept it as evidence.

2 h

21. Platform Citation Formats

Format citations to match each platform's expected style. Move fast without tripping on platform-specific format rules.

1 h

22. Platform Navigation & Tools

Navigate platform interfaces and tooling efficiently. Reduce friction so your evaluation time goes into the work, not the UI.

2 h

23. Time Management & Productivity

Run an evaluation session with a sustainable cadence. Protect attention across the hours where careless mistakes pile up.

2 h

24. Gating Test Simulations

Walk into the gating exam knowing what to expect. Practice on representative tasks under representative time pressure.

1 h