Level 2: Advanced

15 modules 38+ 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. Advanced RLHF Concepts

Deep dive into RLHF failure modes, reward hacking, specification gaming, and alignment faking. Learn to detect and prevent these issues.

3 h

02. Inter-Annotator Agreement

Master IAA metrics including Cohen's Kappa and weighted Kappa. Learn platform-specific thresholds and strategies to improve agreement.

3 h

03. Model Failure Prompting & Adversarial Testing

Design prompts that systematically expose model failures. Categorize failures, test abstention with logic traps, and use the knowledge check follow-up technique.

3 h

04. Dimension Tensions

Navigate HHH (Helpful, Harmless, Honest) conflicts using the priority hierarchy. Master helpfulness vs safety trade-offs.

3 h

05. Ambiguous Prompt Interpretation

Apply the CLEAR framework for resolving unclear intent. Distinguish user error from intentional ambiguity and evaluate fairly.

2 h

06. Complex Safety Scenarios

Handle dual-use content, professional claims, cross-cultural safety, and novel harm recognition in challenging scenarios.

3 h

07. Hierarchical Criteria Structures

Master the GCM framework (Global Category Meta-Criteria Unit Properties) for building complex rubrics systematically.

4 h

08. Criterion Tension Resolution

Learn the SAFE framework for resolving conflicts between criteria when atomicity, instance-specificity, and other properties conflict.

3 h

09. Novel Task Rubric Creation

Master the five-step process for creating rubrics for unfamiliar task types. Build a pattern library for rapid rubric development.

3 h

10. Expert Speed Optimization

Achieve 10-15 minute complex task completion at 95%+ quality. Master speed techniques while avoiding time sinks.

3 h

11. Advanced Source Evaluation

Handle conflicting sources with the CONFLICT framework. Detect misinformation and evaluate controversial topics objectively.

2 h

12. Reviewer & QA Fundamentals

Master the quality rubric for reviewing contributors, feedback frameworks (sandwich, evidence-based, STAR), and SBQ decisions.

3 h

13. Calibration & Drift Detection

Master Cohen's Kappa for inter-annotator agreement, detect personal drift patterns, and develop self-calibration techniques.

3 h

14. Task Difficulty Assessment

Apply the Marimba checklist for project assessment. Recognize time sinks and calibrate task difficulty before committing.

2 h

15. Cross-Platform Optimization

Develop efficient multi-platform workflows and a sustainable evaluation career. Master rapid onboarding, tool proficiency, reviewer-track progression, and specialization.

2 h