ANTHROPIC FELLOWS PROGRAM: AI SAFETY, ECONOMICS, AND RESEARCH METHODS PRACTICE TEST (SAMPLE)

STUDY GUIDE

SAMPLE STUDY GUIDE WITH LIMITED QUESTIONS

VISIT US ONLINE FOR THE FULL VERSION STUDY GUIDE

https://anthropicfellowsaisafetyecon.examzify.com

Table of Contents

Introduction

Preparing for a certification exam can feel overwhelming, but with the right tools, it becomes an opportunity to build confidence, sharpen your skills, and move one step closer to your goals. At Examzify, we believe that effective exam preparation isn't just about memorization, it's about understanding the material, identifying knowledge gaps, and building the test-taking strategies that lead to success.

This guide was designed to help you do exactly that.

Whether you're preparing for a licensing exam, professional certification, or entry-level qualification, this book offers structured practice to reinforce key concepts. You'll find a wide range of multiple-choice questions, each followed by clear explanations to help you understand not just the right answer, but why it's correct.

The content in this guide is based on real-world exam objectives and aligned with the types of questions and topics commonly found on official tests. It's ideal for learners who want to:

- Practice answering questions under realistic conditions,
- Improve accuracy and speed,
- Review explanations to strengthen weak areas, and
- Approach the exam with greater confidence.

We recommend using this book not as a stand-alone study tool, but alongside other resources like flashcards, textbooks, or hands-on training. For best results, we recommend working through each question, reflecting on the explanation provided, and revisiting the topics that challenge you most.

Remember: successful test preparation isn't about getting every question right the first time, it's about learning from your mistakes and improving over time. Stay focused, trust the process, and know that every page you turn brings you closer to success.

Let's begin.

How to Use This Guide

This guide is designed to help you study more effectively and approach your exam with confidence. Whether you're reviewing for the first time or doing a final refresh, here's how to get the most out of your Examzify study guide:

1. Start with a Diagnostic Review

Skim through the questions to get a sense of what you know and what you need to focus on. Your goal is to identify knowledge gaps early.

2. Study in Short, Focused Sessions

Break your study time into manageable blocks (e.g. 30 – 45 minutes). Review a handful of questions, reflect on the explanations.

3. Learn from the Explanations

After answering a question, always read the explanation, even if you got it right. It reinforces key points, corrects misunderstandings, and teaches subtle distinctions between similar answers.

4. Track Your Progress

Use bookmarks or notes (if reading digitally) to mark difficult questions. Revisit these regularly and track improvements over time.

5. Simulate the Real Exam

Once you're comfortable, try taking a full set of questions without pausing. Set a timer and simulate test-day conditions to build confidence and time management skills.

6. Repeat and Review

Don't just study once, repetition builds retention. Re-attempt questions after a few days and revisit explanations to reinforce learning. Pair this guide with other Examzify tools like flashcards, and digital practice tests to strengthen your preparation across formats.

There's no single right way to study, but consistent, thoughtful effort always wins. Use this guide flexibly, adapt the tips above to fit your pace and learning style. You've got this!

Questions

1. What best describes the Labor Market Effects of AI?

A. The measurable changes in employment caused by AI adoption, including which jobs are created, eliminated, augmented, or transformed.

B. The ethical implications of AI in the workplace.

C. The rate of AI hardware production.

D. The overall profitability of AI products.

2. What is the idea of Constitutional AI, and what problem does it attempt to solve in alignment?

A. Constitutional AI uses fixed reward signals derived from market data.

B. Constitutional AI uses a set of non-verifiable principles to guide outputs.

C. Constitutional AI uses a set of high-level, verifiable principles (a 'constitution') to guide model outputs and deference mechanisms, aiming to reduce value drift by constraining behavior with a formal, auditable rule set rather than opaque human preference signals.

D. Constitutional AI relies on post-hoc explanations to enforce safety.

3. Which term describes an AI model whose code and trained weights are publicly available for use?

A. Open-Source Model

B. API (Application Programming Interface)

C. Dataset

D. RLHF (Reinforcement Learning from Human Feedback)

4. Which components are recommended when designing evaluation suites to test interpretability and explainability of models?

A. Explanations tasks; fidelity and usefulness; consistency across inputs; counterfactual and causal justification components.

B. Only standardized accuracy metrics and user satisfaction surveys.

C. Data privacy techniques and encryption measures.

D. Model architecture choices and dataset size.

5. Which concept studies how humans and AI systems work together, including handoffs, trust, and productive partnerships?

A. Human-AI Collaboration

B. Technological Adoption

C. Labor Market Effects of AI

D. Digital Divide

6. The training method used to align language models with human preferences.

A. RLHF (Reinforcement Learning from Human Feedback)

B. Dataset

C. API (Application Programming Interface)

D. CAI (Constitutional AI)

7. Which term denotes a stage of the interview process where references are contacted without giving you notice?

A. Reference Check

B. External Infrastructure

C. Compute Budget

D. Public Output

8. Which term describes the final deliverable produced or expected from a fellowship project?

A. Compute Budget

B. Public Output

C. Reference Check

D. FTE / Full-Time Offer

9. Which AI Security mentor is known for work on model vulnerabilities and attacks?

A. Sam Bowman

B. Nicholas Carlini

C. Peter McCrory

D. Saffron Huang

10. What is the primary purpose of preference learning in AI alignment?

A. To generate synthetic data

B. To infer reward models from human judgments.

C. To reduce training time

D. To optimize a loss function

Answers

1. B
2. C
3. A
4. A
5. A
6. A
7. A
8. B
9. B
10. B

Explanations

1. What best describes the Labor Market Effects of AI?

A. The measurable changes in employment caused by AI adoption, including which jobs are created, eliminated, augmented, or transformed.

B. The ethical implications of AI in the workplace.

C. The rate of AI hardware production.

D. The overall profitability of AI products.

The key idea here is how AI changes the labor market in real, observable ways—who gets hired, what tasks people do, and how their skills and wages shift as AI is adopted. This focuses on the actual employment outcomes of AI: which jobs disappear, which are created, which become augmented, and how existing roles are transformed to incorporate AI tools. It also includes how skill requirements change, how workers move between occupations, and where these effects show up geographically or across industries. That makes the best choice the one that talks about measurable employment changes driven by AI, including creation, elimination, augmentation, and transformation of jobs. For example, routine or data-entry tasks may be automated, reducing demand for those tasks, while new roles in AI maintenance, data labeling, or model governance emerge. Some jobs evolve to require higher levels of data literacy or collaboration with AI systems. All of these are labor market outcomes tied to AI adoption. The other topics—ethical implications in the workplace, rates of hardware production, or overall profitability of AI products—cover important areas, but they describe different dimensions. Ethics deals with fairness, privacy, and governance; hardware production rates speak to supply chains and capacity; profitability relates to company finances. None of these are the direct measures of how AI changes employment, skills, and job opportunities that define the labor market effects.

2. What is the idea of Constitutional AI, and what problem does it attempt to solve in alignment?

A. Constitutional AI uses fixed reward signals derived from market data.

B. Constitutional AI uses a set of non-verifiable principles to guide outputs.

C. Constitutional AI uses a set of high-level, verifiable principles (a 'constitution') to guide model outputs and deference mechanisms, aiming to reduce value drift by constraining behavior with a formal, auditable rule set rather than opaque human preference signals.

D. Constitutional AI relies on post-hoc explanations to enforce safety.

Constitutional AI centers on a formal set of high-level, verifiable principles—the AI's constitution—that guide what the model should output and how it should defer to other judgments. This creates auditable constraints on behavior rather than relying on opaque, ever-shifting human preference signals used to shape rewards. The idea is to reduce value drift: by grounding decisions in a transparent rule set, the model stays within agreed norms even as tasks change or data shifts occur. Outputs are produced with the constitution in mind, and the system can defer or escalate when a principle is ambiguous or conflicts arise, keeping safety and alignment more predictable. This stands in contrast to approaches that depend on fixed reward signals, non-verifiable guidelines, or post-hoc explanations to enforce safety.

3. Which term describes an AI model whose code and trained weights are publicly available for use?

A. Open-Source Model

B. API (Application Programming Interface)

C. Dataset

D. RLHF (Reinforcement Learning from Human Feedback)

Open-source models are AI models whose code and trained weights are publicly available for anyone to inspect, run, modify, or build upon. This openness lets people reproduce results, customize the model for their own needs, and deploy it without relying on a single provider. An API, by contrast, gives access to a model's functionality through an interface hosted by someone else, without exposing the internal code or the trained weights. That's why an API isn't describing an open model. A dataset is the collection of data used to train the model, not the model itself. RLHF, or Reinforcement Learning from Human Feedback, is a training approach used to shape behavior, not a description of whether the model's code and weights are publicly available.

4. Which components are recommended when designing evaluation suites to test interpretability and explainability of models?

A. Explanations tasks; fidelity and usefulness; consistency across inputs; counterfactual and causal justification components.

B. Only standardized accuracy metrics and user satisfaction surveys.

C. Data privacy techniques and encryption measures.

D. Model architecture choices and dataset size.

Designing evaluation suites for interpretability centers on how explanations reflect the model's behavior and how helpful they are to humans. That means including tasks that require the model to produce explanations for its decisions (such as feature attributions, natural language rationales, or example-based reasoning). It also means measuring fidelity—the extent to which the explanation matches what the model actually used to reach its decision—and usefulness, i.e., whether the explanation helps a user diagnose errors, gain trust, or make better judgments. Explanations should be consistent across similar inputs, so explanations don't vary wildly for comparable cases, which would undermine reliability. Including counterfactual and causal justification components helps users understand how changing inputs would alter outcomes and how features causally relate to decisions, which strengthens intuition and transparency. The other options miss these core aspects. Focusing only on standardized accuracy metrics and user satisfaction surveys ignores whether explanations faithfully reflect the model's reasoning or assist in understanding and trust. Privacy techniques and encryption measures address security, not interpretability. Model architecture choices and dataset size influence performance and capacity, but they don't provide direct evaluation of how interpretable or explainable the model's decisions are.

5. Which concept studies how humans and AI systems work together, including handoffs, trust, and productive partnerships?

A. Human-AI Collaboration

B. Technological Adoption

C. Labor Market Effects of AI

D. Digital Divide

Understanding how humans and AI systems work together focuses on how people and machines share tasks, decide when to rely on automated outputs, and build trust to form a productive partnership. The best answer, Human-AI Collaboration, centers on the interaction, coordination, and joint problem-solving between people and intelligent systems, including handoffs where control or responsibility passes between them, how trust is earned and calibrated, and how to design workflows that maximize joint performance. This goes beyond mere adoption of technology or its economic effects; it's about the ongoing working relationship that lets humans leverage AI strengths while compensating for its limitations. In contrast, Technological Adoption deals with how quickly or widely technologies are adopted, not with the interactive dynamics themselves. Labor Market Effects of AI looks at employment and economic impact, rather than day-to-day collaboration and handoff design. Digital Divide focuses on unequal access to technology, not the collaborative dynamics between humans and AI.

6. The training method used to align language models with human preferences.

A. RLHF (Reinforcement Learning from Human Feedback)

B. Dataset

C. API (Application Programming Interface)

D. CAI (Constitutional AI)

Training method used to align language models with human preferences is RLHF, reinforcement learning from human feedback. It works by gathering human judgments on model outputs—through demonstrations or rankings—and then training a reward model to predict those judgments. The base language model is then fine-tuned with reinforcement learning to maximize the reward from that reward model, nudging outputs toward what people find helpful, safe, and desirable. This combines concrete examples with a signal that captures human preferences, helping the model generalize beyond the exact demonstrations. A plain dataset lacks this evaluative signal about alignment, so it can imitate but not optimally reflect human priorities. An API is just an interface to use the model, not a training method. Constitutional AI offers an alternative approach that constrains behavior with a set of rules, but RLHF specifically uses human feedback to shape the learning signal through rewards.

7. Which term denotes a stage of the interview process where references are contacted without giving you notice?

A. Reference Check

B. External Infrastructure

C. Compute Budget

D. Public Output

Reference checks are the stage of the hiring process where employers reach out to people who know you professionally to verify your background and experience. This often happens after you've made it past initial interviews and are nearing an offer, though some companies may conduct checks after you've given consent or at other late stages. The goal is to confirm details you've shared—roles, responsibilities, accomplishments—and to hear about your performance from someone who worked with you. References can shed light on how you work with colleagues, handle deadlines, communicate, and respond to challenges. During a reference check, interviewers typically ask about your strengths, areas for growth, reliability, teamwork, and reasons for leaving a position. References are usually former supervisors or teammates who can provide an independent perspective on your work. Sometimes checks happen with advance notice, and other times they occur without you being alerted in the moment; both scenarios are used to gather a candid, external view of your fit for the role. Other terms listed don't describe this part of the process. They refer to things like systems or resources (not related to interviewing) or outputs and budgets, which aren't stages in how employers verify your background through references.

8. Which term describes the final deliverable produced or expected from a fellowship project?

A. Compute Budget

B. Public Output

C. Reference Check

D. FTE / Full-Time Offer

The concept being tested is identifying the term that best names the final deliverable a fellowship project is expected to produce and share publicly. Public Output fits this role because it emphasizes the tangible product of the work that is intended to be shared with the wider community—things like a research paper, dataset, software artifact, or a final report. It signals not just what is produced, but that it is meant for public access and dissemination, reflecting the project's impact and usefulness. The other options describe things that are unrelated to the final product. A compute budget is about the resources needed for the work, not the deliverable itself. A reference check is a step in evaluating candidates or partners, not an outcome of the project. FTE / Full-Time Offer relates to employment terms, not what the project yields.

9. Which AI Security mentor is known for work on model vulnerabilities and attacks?

A. Sam Bowman

B. Nicholas Carlini

C. Peter McCrory

D. Saffron Huang

In AI security, understanding adversarial vulnerabilities and how to craft model-targeted attacks is a central focus because small, carefully designed input changes can cause a model to misbehave or reveal its internals. Nicholas Carlini is a leading figure in this area, known for developing powerful adversarial attacks against neural networks. His work, including the Carlini-Wagner attacks, shows how to produce misclassifications with minimal perturbations and across different threat models, which provides a rigorous way to test model robustness and evaluate defenses. This practical, rigorous approach to exposing weaknesses and benchmarking defenses has made him a prominent mentor-like figure in security research. The other names are known for work in areas outside this specific security focus, such as natural language understanding or broader AI research, so they aren't the figure most associated with model vulnerabilities and attacks.

10. What is the primary purpose of preference learning in AI alignment?

A. To generate synthetic data

B. To infer reward models from human judgments.

C. To reduce training time

D. To optimize a loss function

Preference learning in AI alignment focuses on turning human judgments into guidance for the agent by inferring a reward model from comparisons of preferred outcomes. Humans compare two outputs or behaviors and indicate which they prefer; this feedback is used to train a model that assigns higher reward to preferred results. The learned reward model then becomes the target the agent aims to maximize, typically inside a reinforcement learning loop. This approach ties the agent's goals to human values without needing explicit rules, making alignment more scalable for complex tasks. Generating synthetic data, reducing training time, or simply optimizing a generic loss function aren't the primary aim here; the core purpose is to derive a reward signal from human judgments that guides behavior.

Next Steps

Congratulations on reaching the final section of this guide. You've taken a meaningful step toward passing your certification exam and advancing your career.

As you continue preparing, remember that consistent practice, review, and self-reflection are key to success. Make time to revisit difficult topics, simulate exam conditions, and track your progress along the way.

If you need help, have suggestions, or want to share feedback, we'd love to hear from you. Reach out to our team at hello@examzify.com.

Or visit your dedicated course page for more study tools and resources:

https://anthropicfellowsaisafetyecon.examzify.com

We wish you the very best on your exam journey. You've got this!