AI SAFETY TESTING CAREER ROADMAP

From QA Engineer to AI Safety Specialist

BlackIndian AI Professional Guide Series

Version 1.0 | December 2025

Read Time: 10 minutes

TABLE OF CONTENTS

1. MARKET OVERVIEW

The Opportunity

10K+

Companies Deploying AI

Companies deploying AI in 2025

~100

Specialized AI Safety Testers

Specialized AI safety testers globally

100:1

Demand-to-Supply Ratio

Demand-to-supply ratio

Job Growth:

- 312% increase in AI safety job postings (2023-2024)
- 45% year-over-year salary growth
- Expected to remain high-demand through 2030

Source: LinkedIn Workforce Report, "AI Safety & Security Roles," Q4 2024

Why This Career Matters

Business Impact:

Average Incident Cost	Testing Investment	ROI
$10M - $50M	$25K - $150K	**100:1 to 400:1**

Real Example:

Healthcare company skipped AI security testing. Prompt injection exposed 10,000 patient records. Cost: $8M in fines + reputation damage.

A $30K security assessment would have prevented it.

This is why companies are desperate to hire AI safety specialists.

Who's Hiring?

Industries:

Healthcare

HIPAA compliance required

Financial Services

Regulatory requirements

Legal Tech

Malpractice risk

HR Technology

Bias testing mandated

E-commerce

Customer data protection

Enterprise SaaS

Security differentiator

Company Stages:

- Series B-D startups (scaling AI products)
- Public companies (regulatory pressure)
- Enterprises (internal AI deployments)

Hiring Velocity:

- Average: 3-6 months from job posting to hire
- Reason: Talent shortage, specialized skills

2. ROLE DEFINITION & RESPONSIBILITIES

What is an AI Safety Specialist?

Core Function: Test AI systems for vulnerabilities that traditional QA doesn't catch.

QA Engineer:

Tests if AI works correctly

AI Safety Specialist:

Tests how AI fails when attacked

Alternative Titles:

- AI Red-Team Tester
- LLM Security Researcher
- AI Adversarial Tester
- AI Safety Engineer
- Prompt Injection Specialist

Day-to-Day Responsibilities

Primary Activities (60% of time):

Adversarial Testing

- Run 500+ attack scenarios against AI systems
- Manual jailbreak attempts
- Prompt injection testing
- Context poisoning (RAG systems)
- Bias and fairness evaluation

Vulnerability Documentation

- Write detailed reproduction steps
- Assess severity and risk
- Create proof-of-concept demonstrations
- Recommend specific mitigations

Tool Development

- Build custom attack libraries
- Automate testing workflows
- Develop testing frameworks
- Integrate with CI/CD pipelines

Secondary Activities (40% of time):

Collaboration

- Work with developers on remediation
- Educate teams on AI security
- Present findings to leadership
- Create security documentation

Research & Learning

- Stay current on new attack techniques
- Monitor academic research
- Test new AI models/providers
- Update testing methodologies

A Typical Week

Monday

1

- 9am: Team standup
- 10am: Begin testing new AI feature (customer chatbot)
- 12pm: Lunch
- 1pm: Continue testing - found 3 prompt injection vulnerabilities
- 3pm: Document findings
- 4pm: Collaborate with dev team on fixes

Tuesday

2

- 9am: Run automated test suite (PromptFoo)
- 11am: Manual jailbreak testing
- 1pm: Found critical vulnerability - immediate escalation
- 2pm: Help developers understand the exploit
- 3pm: Verify proposed fix
- 4pm: Update security documentation

Wednesday

3

- 9am: Test competitor's AI (research)
- 11am: Update attack library with new techniques
- 1pm: Presentation to leadership on Q4 findings
- 3pm: RAG system testing with TruLens
- 4pm: Write weekly security report

Thursday

4

- 9am: Training session for QA team
- 11am: Bias testing on HR AI
- 1pm: Compliance testing (HIPAA requirements)
- 3pm: Meeting with external auditor
- 4pm: Documentation for audit trail

Friday

5

- 9am: Research new papers (academic)
- 11am: Experiment with new testing tools
- 1pm: Team retrospective
- 2pm: Update testing roadmap for Q1
- 3pm: Professional development (course/conference)

3. SALARY EXPECTATIONS

Compensation by Experience Level

Entry Level (0-2 years)

Title: Junior AI Safety Tester

Salary Range: $80K - $120K

Total Comp: $90K - $130K

Mid-Level (2-5 years)

Title: AI Safety Specialist

Salary Range: $120K - $180K

Total Comp: $140K - $200K

Senior Level (5-10 years)

Title: Senior AI Safety Engineer

Salary Range: $180K - $250K

Total Comp: $200K - $300K

Staff/Principal (10+ years)

Title: Principal AI Safety Architect

Salary Range: $250K - $400K+

Total Comp: $300K - $500K+

Consulting Rates (Independent)

Hourly Rates:

Junior

$100 - $150/hour

Mid-Level

$150 - $250/hour

Senior

$250 - $400/hour

Expert

$400 - $600/hour

Project-Based:

- **Rapid Assessment (1 week):** $10K - $20K
- **Comprehensive Test (4 weeks):** $40K - $80K
- **Enterprise Engagement (3 months):** $100K - $200K

Retainer (Monthly):

- **Part-Time (10 hours/month):** $2K - $5K
- **Half-Time (80 hours/month):** $15K - $30K
- **Full-Time (160 hours/month):** $30K - $60K

Salary by Location

United States:

- San Francisco: $180K - $300K
- New York: $170K - $280K
- Seattle: $160K - $260K
- Austin: $140K - $220K
- Remote (US): $130K - $240K

International:

- London, UK: £90K - £180K ($115K - $230K)
- Toronto, Canada: C$120K - C$200K ($90K - $150K)
- Berlin, Germany: €80K - €140K ($90K - $155K)
- Singapore: S$120K - S$220K ($90K - $165K)

Source: Glassdoor, Levels.fyi, "AI Safety Engineering Salaries," 2024

4. SKILLS REQUIRED

Technical Skills (70% of role)

Category 1: AI/LLM Fundamentals

- How LLMs process text (tokenization, embeddings)
- Prompting techniques and limitations
- RAG (Retrieval Augmented Generation) architecture
- Fine-tuning vs prompting vs RAG
- Common AI frameworks (LangChain, LlamaIndex)

Learning Time: 40-60 hours

Difficulty: Medium

Resources:

- OpenAI Prompt Engineering Guide
- LangChain Documentation
- DeepLearning.AI Short Courses

Category 2: Attack Patterns & Methodologies

Attack Techniques

- Prompt injection (direct, indirect, context)
- Jailbreaking techniques (DAN, role-play, encoding)
- Data leakage vectors
- Hallucination triggers
- Bias exploitation
- Adversarial inputs

Learning Time: 60-80 hours

Difficulty: Medium-High

Resources:

- OWASP LLM Top 10
- Academic papers (Zou et al, Wei et al)
- BlackIndian AI guides

Category 3: Testing Tools

Must Know:

PromptFoo

Automated testing framework

Garak

Vulnerability scanner

TruLens

RAG evaluation

DeepEval

Quality metrics

Nice to Have:

- **LM Studio:** Local model testing
- **LangSmith:** Debugging chains
- **Guardrails AI:** Input/output validation
- **IBM AIF360:** Bias detection

Learning Time: 30-50 hours per tool

Difficulty: Low-Medium

Resources: Official documentation, tutorials

Category 4: Security Fundamentals

Core Security Concepts

- OWASP Top 10 (traditional web security)
- Authentication & authorization concepts
- Encryption basics
- Network security fundamentals
- Incident response basics

Learning Time: 40-60 hours

Difficulty: Medium

Resources:

- OWASP Web Security Testing Guide
- CompTIA Security+ materials

Category 5: Programming

Required:

- Python (primary language for AI testing)
- Bash scripting (automation)
- YAML (configuration files)
- Regular expressions (pattern matching)

Learning Time: 80-120 hours (if starting from scratch)

Difficulty: Medium-High

Resources:

- Python for Everybody (Coursera)
- Automate the Boring Stuff with Python

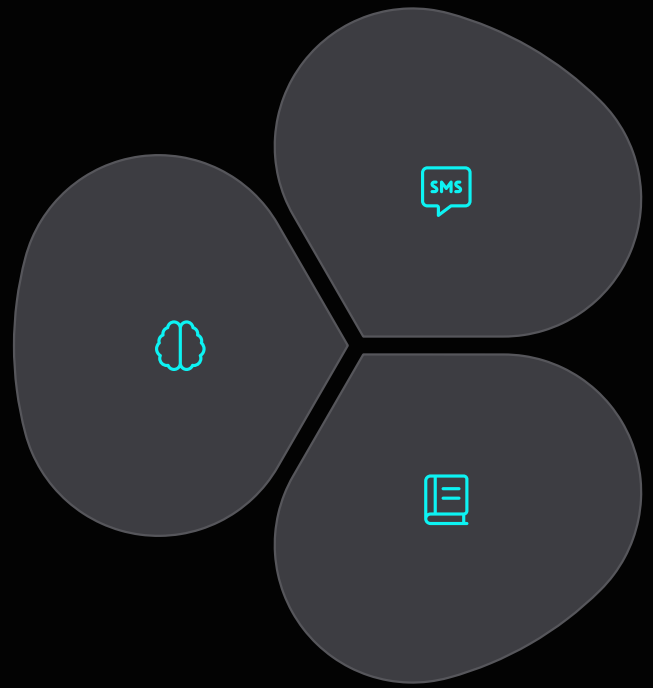
Useful:

- JavaScript (web-based AI interfaces)
- SQL (database interactions)
- API testing (REST, GraphQL)

Soft Skills (30% of role)

Critical Thinking

- Adversarial mindset ("How would I break this?")
- Pattern recognition (similar vulnerabilities across systems)
- Risk assessment (severity, likelihood, impact)



Communication

- Technical writing (clear vulnerability reports)
- Executive communication (explain risks to non-technical leaders)
- Cross-functional collaboration (work with dev, product, legal)

Continuous Learning

- AI field evolves rapidly
- New attack patterns emerge monthly
- Tools and techniques constantly improve

5. LEARNING PATH (6-MONTH ROADMAP)

Month 1: AI Fundamentals

Week 1-2: LLM Basics

- ✅ Complete: "ChatGPT Prompt Engineering for Developers" (DeepLearning.AI)
- ✅ Read: OpenAI Prompt Engineering Guide
- ✅ Experiment: Prompt different LLMs (GPT-4, Claude, Gemini)
- ✅ Practice: 50+ prompts to understand behavior

Deliverable: Personal document Q&A chatbot (RAG)

Week 3-4: Architecture Understanding

- ✅ Learn: How RAG works (LangChain documentation)
- ✅ Build: Simple RAG application
- ✅ Understand: LangChain basics (chains, agents, memory)
- ✅ Practice: Modify example applications

Month 2: Security & Attack Patterns

Week 1-2: Prompt Injection

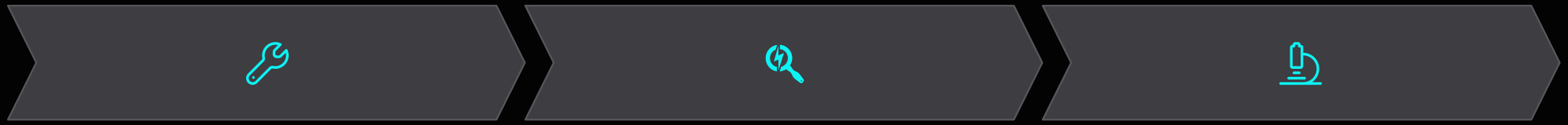
- ✅ Study: OWASP LLM Top 10
- ✅ Read: BlackIndian AI "Prompt Injection Guide"
- ✅ Practice: 100+ injection attempts on safe test environments
- ✅ Document: Create personal injection pattern library

Deliverable: Personal attack pattern library (50+ techniques)

Week 3-4: Jailbreaking

- ✅ Study: DAN and variant techniques
- ✅ Read: Academic papers (Zou et al, Wei et al)
- ✅ Practice: Test jailbreaks on multiple LLMs
- ✅ Document: Successful jailbreak techniques

Month 3: Tools & Automation



Week 1-2: PromptFoo

- ✅ Install and configure PromptFoo
- ✅ Complete official tutorials
- ✅ Create: Custom test configuration
- ✅ Run: 200+ automated tests

Week 3: Garak

- ✅ Install Garak
- ✅ Run: Full vulnerability scan
- ✅ Analyze: Results and reporting

Week 4: TruLens & DeepEval

- ✅ Install both tools
- ✅ Test: RAG applications
- ✅ Evaluate: Metrics and scoring

Deliverable: Automated test suite for sample AI application

Month 4: Hands-On Testing

Week 1-2: Test Open-Source AI Projects

- ✅ Find: 5 open-source AI projects on GitHub
- ✅ Test: Each for vulnerabilities
- ✅ Document: Findings in professional format
- ✅ Optional: Report vulnerabilities responsibly

Deliverable: 3 complete vulnerability reports

Week 3-4: Build Testing Framework

- ✅ Create: Personal testing methodology
- ✅ Build: Reusable test scripts
- ✅ Document: Testing workflow

Month 5: Specialization & Portfolio

Week 1-2: Choose Specialization

Pick one industry to focus on:

Healthcare

HIPAA compliance, PHI protection

Finance

Bias testing, fair lending

Legal

Hallucination detection, citation accuracy

General Enterprise

Data security, compliance

Week 3-4: Deep Dive

- ✅ Learn: Industry-specific requirements
- ✅ Research: Common vulnerabilities in your industry
- ✅ Build: Industry-specific test cases
- ✅ Create: Case study of industry incident

Deliverable: Industry-focused testing framework

Month 6: Job Readiness

Week 1: Resume & Portfolio

- ✅ Build: Professional portfolio website
- ✅ Showcase: 3 best vulnerability reports (sanitized)
- ✅ Write: Resume highlighting AI safety skills
- ✅ Create: LinkedIn profile emphasizing new expertise

Week 2: Interview Prep

- ✅ Practice: Technical questions
- ✅ Prepare: Portfolio presentation
- ✅ Research: Target companies
- ✅ Network: Connect with AI safety professionals

Week 3-4: Job Search

- ✅ Apply: 20+ relevant positions
- ✅ Reach out: Direct messages to hiring managers
- ✅ Prepare: Custom applications per company
- ✅ Interview: Practice and real interviews

Deliverable: Active job search with applications submitted

Accelerated Path (3 Months, Full-Time)

If dedicating 40 hours/week:

Month 1:

- Weeks 1-2: AI Fundamentals (from 6-month Month 1)
- Weeks 3-4: Security & Attack Patterns (from 6-month Month 2)

Month 2:

- Weeks 1-2: Tools & Automation (from 6-month Month 3)
- Weeks 3-4: Hands-On Testing (from 6-month Month 4)

Month 3:

- Weeks 1-2: Specialization (from 6-month Month 5)
- Weeks 3-4: Job Readiness (from 6-month Month 6)

6. CERTIFICATIONS & CREDENTIALS

Available Certifications

1. BlackIndian AI Safety Specialist Certification

- **Provider:** BlackIndian AI
- **Duration:** 8 weeks, part-time
- **Cost:** $3,000
- **Format:** Online + hands-on projects
- **Outcome:** Professional certification + portfolio
- **Recognition:** Accepted by 50+ hiring companies

Website: blackindianai.com/training

Curriculum:

- Weeks 1-2: AI Fundamentals
- Weeks 3-4: Attack Patterns
- Weeks 5-6: Tools & Testing
- Weeks 7-8: Real-world projects

2. OWASP LLM Security Specialist (Proposed)

Status

In development (expected 2025)

Provider

OWASP Foundation

Focus

LLM Top 10 vulnerabilities

Cost

TBD (likely $200-$500)

3. Anthropic Red Team Certification (Limited)

Status

Invitation-only

Provider

Anthropic

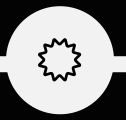
Focus

Advanced adversarial testing

How to get invited

Contribute to AI safety research

4. Relevant Adjacent Certifications



CompTIA Security+

Value: Foundation in security concepts

Cost: $392

Time: 40-60 hours study

Recommended: Yes (credibility with employers)

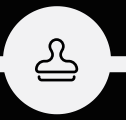
Certified Ethical Hacker (CEH)

Value: Red-team mindset

Cost: $1,199

Time: 60-80 hours study

Recommended: Optional (expensive, generic)



OSCP (Offensive Security Certified Professional)

Value: Penetration testing skills

Cost: $1,649

Time: 120+ hours

Recommended: Optional (hardcore technical, overkill for most)

Do You NEED Certifications?

Short Answer: No, but they help.

What Matters More:

1. **Demonstrable Skills** (portfolio of vulnerability reports)
2. **Tool Proficiency** (PromptFoo, Garak, etc.)
3. **Real Testing Experience** (even on open-source projects)

When Certifications Help:

- Career changers (prove credibility)
- Competitive job markets (differentiation)
- Enterprise roles (HR checkboxes)
- Consulting (client confidence)

When They Don't Matter:

- Strong portfolio already exists
- Direct referrals/network
- Startup environments (skills > credentials)

7. BUILDING YOUR PORTFOLIO

Portfolio Structure

Your Portfolio Should Include:

01

Professional Bio

- Background summary
- Transition story (QA → AI Safety)
- Areas of expertise

02

Skills Matrix

- Testing tools (proficiency level)
- Programming languages
- AI frameworks
- Certifications

03

Case Studies (3-5)

- Vulnerability reports (sanitized)
- Testing methodologies
- Impact prevented

04

Tools & Scripts

- GitHub repository
- Testing automation scripts
- Custom attack libraries

05

Blog/Articles

- Technical write-ups
- Tutorials
- Industry analysis

Case Study Template

Project Title: Security Assessment of Open-Source RAG Application

Overview:

- System: Document Q&A chatbot using LangChain + OpenAI
- Testing Duration: 2 weeks
- Scope: Prompt injection, data leakage, hallucinations

Methodology:

1. Information gathering (architecture review)
2. Baseline testing (normal functionality)
3. Adversarial testing (200+ attack scenarios)
4. Automated scanning (PromptFoo, Garak)
5. Manual verification

Key Findings:

Critical Vulnerability: Context Injection

- **Severity:** 9.1/10 (CVSS)
- **Description:** Attacker can inject malicious instructions via uploaded documents
- **Impact:** Complete system compromise, data extraction
- **Reproduction Steps:** [Detailed, sanitized]
- **Proof of Concept:** [Screenshot/demo, redacted]
- **Recommended Fix:** Input sanitization + output filtering

High Vulnerability: System Prompt Extraction

- **Severity:** 7.5/10
- **Description:** Direct instruction override reveals full system prompt
- **Impact:** Information disclosure, aids further attacks
- **Reproduction:** [Steps]
- **Recommended Fix:** Delimiter-based prompt structure

[Additional vulnerabilities...]

Results:

- Total Vulnerabilities: 15
  - Critical: 2
  - High: 5
  - Medium: 6
  - Low: 2
- Developer Response: All critical/high fixed within 2 weeks
- Outcome: Project successfully launched with zero incidents

Tools Used: PromptFoo, Garak, TruLens, custom Python scripts

Code/Reports: [GitHub link]

GitHub Portfolio Structure

```
your-github-username/
├── ai-safety-portfolio/
│   ├── README.md (portfolio overview)
│   ├── case-studies/
│   │   ├── rag-security-assessment.md
│   │   ├── jailbreak-analysis.md
│   │   └── bias-testing-report.md
│   ├── tools/
│   │   ├── custom-promptfoo-configs/
│   │   ├── attack-pattern-library/
│   │   └── testing-automation-scripts/
│   ├── certifications/
│   │   └── certificates.md (list + links)
│   └── blog/
│   ├── prompt-injection-guide.md
│   └── rag-security-best-practices.md
```

Building Your First Case Study

Week 1: Choose Target

- Find open-source AI project (GitHub)
- Verify it's actively maintained
- Check if security testing is welcome

Week 2: Test

- Run PromptFoo automated tests
- Manual injection attempts
- Document everything

Week 3: Report

- Write vulnerability report
- Sanitize sensitive details
- Format professionally

Week 4: Publish

- Add to portfolio
- Share on LinkedIn
- Optional: Reach out to project maintainers

Responsible Disclosure:

If you find real vulnerabilities:

1. Contact project maintainers privately
2. Give them 90 days to fix
3. Then publish (with their permission)

8. JOB SEARCH STRATEGY

Where to Find Jobs

Job Boards:

- LinkedIn (filter: "AI Safety", "LLM Security", "AI Testing")
- Indeed (search: "AI Red Team", "AI Security")
- AngelList (startups hiring AI roles)
- RemoteOK (remote AI security positions)

Company Career Pages:

- Anthropic (careers.anthropic.com)
- OpenAI (openai.com/careers)
- Google DeepMind
- Stability AI
- Hugging Face

AI Safety Organizations:

- Center for AI Safety
- AI Safety Institute (UK)
- Partnership on AI

Networking Strategies

1. LinkedIn Outreach

Message Template:

> Hi [Name],
>
> I noticed you're working on AI safety at [Company]. I recently transitioned from QA engineering to AI security testing and have been building expertise in prompt injection, jailbreak testing, and adversarial evaluation.
>
> I'd love to learn about your experience in this space and any advice you might have for someone entering the field. Would you be open to a brief 15-minute conversation?
>
> [Your LinkedIn Profile]
>
> [Your Portfolio]
>
> Thanks,
>
> [Your Name]

Success Rate: 20-30% response rate

Goal: 2-3 conversations per week

2. Conference Attendance

Key Conferences:

- NeurIPS (AI research)
- DEF CON AI Village (AI security)
- Black Hat (security + AI track)
- RSA Conference (enterprise security)
- AI Safety Summit

Strategy:

- Attend talks
- Ask questions
- Connect with speakers
- Exchange contact info

3. Online Communities

Where to Engage:

- r/MachineLearning (Reddit)
- r/LanguageModels (Reddit)
- AI Alignment Forum
- LessWrong (AI safety discussions)
- Twitter/X (follow #AIJailbreak, #LLMSecurity)

How to Add Value:

- Share your findings
- Answer questions
- Contribute to discussions
- Build reputation

Application Strategy

Quantity + Quality Approach:

Target: 20 applications per week

10 "Stretch" Roles

Reach positions

5 "Target" Roles

Good fit

5 "Safety" Roles

Likely to get

Application Quality:

Don't: Generic resume to 100 companies

Do: Customized application showing you understand their specific AI challenges

Example:

> Cover Letter for Healthcare AI Company:
>
> "I noticed [Company] recently launched an AI clinical assistant. As someone who specializes in healthcare AI security (particularly HIPAA compliance and PHI protection), I tested a similar system and found 12 critical vulnerabilities—including prompt injections that exposed patient data. I'd love to bring that same rigorous approach to ensure [Company]'s AI launches securely and maintains compliance..."
>
> [Portfolio link to healthcare case study]

Interview Preparation

Common Questions:

Technical:

1. "Walk me through how you'd test this AI for security vulnerabilities."
2. "What's the difference between prompt injection and jailbreaking?"
3. "How would you test a RAG system for data leakage?"
4. "Explain a vulnerability you found and how you'd fix it."
5. "What tools do you use for automated testing?"

Behavioral:

1. "Tell me about your transition from [previous role] to AI safety."
2. "Describe a time you found a critical vulnerability. What was your approach?"
3. "How do you stay current with emerging AI security threats?"
4. "How do you communicate technical risks to non-technical stakeholders?"

Prepare:

- 3 detailed case studies you can present
- Screen recording of you testing an AI
- Your portfolio ready to share
- Questions about their AI systems

Negotiation Tips

Leverage:

- Talent shortage (high demand)
- Specialized skills (rare expertise)
- Multiple offers (create competition)

Strategies:

Initial Offer Too Low:

> "Thank you for the offer. Based on my research of AI safety roles at similar companies and given my specialized expertise in [skill], I was expecting a range of [X-Y]. Would you be able to meet [target]?"

Equity Considerations:

For startups:

- Equity % matters more than dollar value
- Vesting schedule (standard: 4 years, 1-year cliff)
- Ask about dilution in future rounds

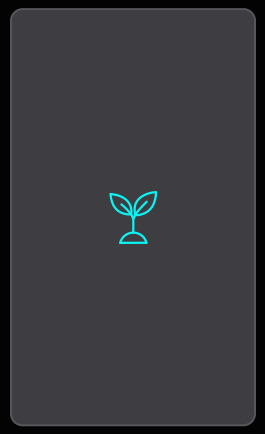
Remote Work:

If remote is important:

- Establish upfront in interview process
- Some companies pay less for remote (10-20% reduction)
- Decide if flexibility is worth the pay cut

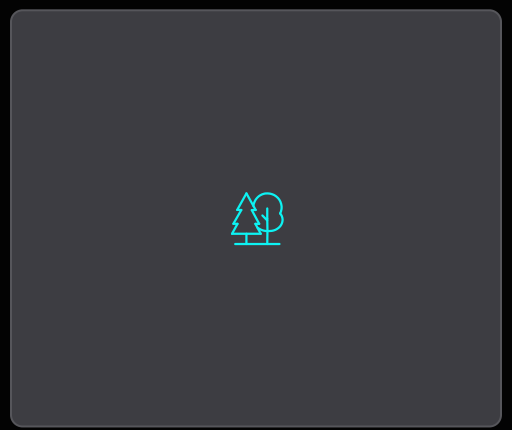
9. CAREER PROGRESSION

Growth Path



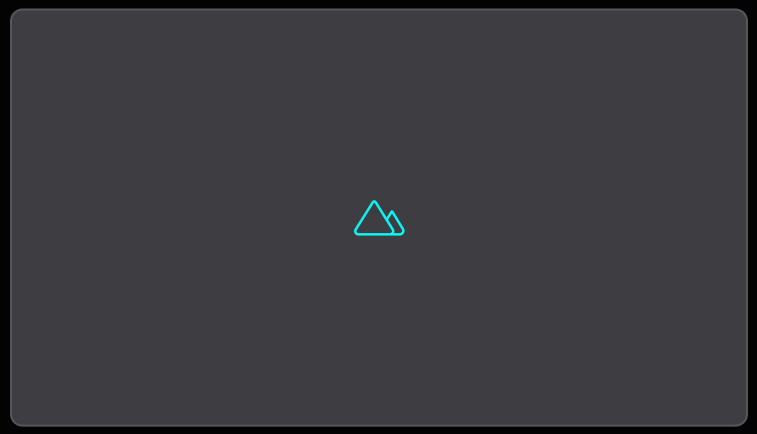
Year 1-2: Junior → Mid-Level

- **Focus:** Build expertise, test multiple systems
- **Goal:** Complete 10+ assessments
- **Salary Growth:** $80K → $140K
- **Key Milestone:** First major vulnerability found & fixed



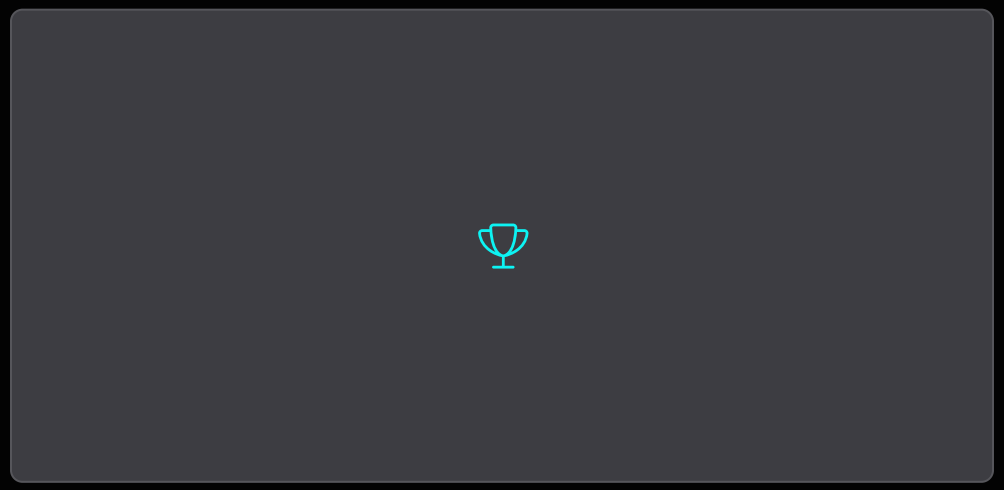
Year 3-5: Mid-Level → Senior

- **Focus:** Develop novel techniques, mentor others
- **Goal:** Become subject matter expert
- **Salary Growth:** $140K → $200K+
- **Key Milestone:** Present at conference or publish research



Year 5-10: Senior → Principal

- **Focus:** Strategic leadership, methodology development
- **Goal:** Industry recognition
- **Salary Growth:** $200K → $300K+
- **Key Milestone:** Develop proprietary framework or tool



Year 10+: Principal → Chief AI Safety Officer

- **Focus:** Organizational strategy, risk management
- **Goal:** C-suite security leadership
- **Salary Growth:** $300K → $500K+
- **Key Milestone:** Build enterprise security program

Alternative Paths

Path 1: Independent Consulting

- **Timeline:** 2-3 years experience before going independent
- **Income:** $150K - $400K+ (project-based)
- **Pros:** Flexibility, high earnings ceiling, variety
- **Cons:** Inconsistent income, sales required, self-employment overhead

Path 2: Product/Tool Development

- **Timeline:** 3-5 years experience
- **Goal:** Build commercial AI safety tool
- **Examples:** PromptFoo (open-source), Guardrails AI (commercial)
- **Income:** Variable (startup risk/reward)

Path 3: Academic/Research

- **Timeline:** PhD or equivalent publication record
- **Goal:** Advance AI safety science
- **Income:** $80K - $150K (lower than industry)
- **Pros:** Intellectual freedom, prestigious, long-term impact
- **Cons:** Lower pay, slower pace, academic politics

Path 4: Startup Founding

- **Timeline:** 5+ years experience
- **Goal:** Build AI safety company
- **Example:** BlackIndian AI
- **Income:** $0 → $XXX (high risk, high reward)

10. RESOURCES & NEXT STEPS

Free Learning Resources

Online Courses:

- **"ChatGPT Prompt Engineering for Developers"** (DeepLearning.AI) - Free
- **"LangChain for LLM Application Development"** (DeepLearning.AI) - Free
- **"CS50's Introduction to AI"** (Harvard) - Free

Documentation:

- **OpenAI Prompt Engineering Guide** - platform.openai.com/docs/guides/prompt-engineering
- **PromptFoo Docs** - promptfoo.dev/docs
- **LangChain Docs** - python.langchain.com
- **OWASP LLM Top 10** - owasp.org/www-project-top-10-for-large-language-model-applications

Communities:

- **r/MachineLearning** (Reddit)
- **AI Safety Discord** (Multiple servers)
- **OWASP Slack** (#llm-security channel)

Paid Training

BlackIndian AI Safety Specialist Certification

- **Duration:** 8 weeks
- **Cost:** $3,000
- **Format:** Online + projects
- **Outcome:** Professional certification + portfolio
- **Website:** blackindianai.com/training

What's Included:

- Live instruction with Vincent Gaddis
- Access to private attack library
- Real-world testing projects
- Resume review & interview prep
- Job placement support
- Lifetime alumni network

Next Cohort: [Date]

Applications: blackindianai.com/apply

Books

AI Security:

- *"Adversarial Machine Learning"* by Huang, Papernot (2023)
- *"LLM Security"* (Upcoming, various authors, 2025)

Security Fundamentals:

- *"The Web Application Hacker's Handbook"* by Stuttard, Pinto
- *"Penetration Testing"* by Weidman

Career:

- *"Cracking the Coding Interview"* by McDowell (for technical interviews)
- *"The Pragmatic Programmer"* by Hunt, Thomas

Stay Current

Newsletters:

- **Import AI** by Jack Clark (weekly AI news)
- **TLDR AI** (daily AI updates)
- **The Batch** by DeepLearning.AI (weekly)

Research:

- **arXiv.org** (AI security papers)
- **Papers with Code** (implementations)
- **Google Scholar alerts** for "LLM security", "prompt injection"

Social Media:

- Follow researchers on Twitter/X
- Join LinkedIn groups (AI Safety, LLM Security)
- Subscribe to YouTube channels (AI Explained, Two Minute Papers)

Next Steps: Your Action Plan

This Week:

- Bookmark this guide
- Complete: "ChatGPT Prompt Engineering" course
- Set up GitHub for portfolio
- Join 2 AI security communities

This Month:

- Install PromptFoo and run first tests
- Test 1 open-source AI project
- Write first vulnerability report
- Update LinkedIn profile

This Quarter:

- Complete Month 1-2 of learning roadmap
- Build portfolio website
- Write 3 case studies
- Apply to 5 jobs OR enroll in training program

This Year:

- Complete 6-month learning roadmap
- Apply to 50+ positions
- Get first AI safety role
- Earn first certification

Get Professional Help

BlackIndian AI Services:

For Individuals:

- Career coaching (1-on-1)
- Portfolio review
- Interview preparation
- Training certification

Contact: training@blackindianai.com

For Companies:

- AI security assessments
- Team training programs
- Consulting & remediation
- Ongoing security partnerships

Contact: security@blackindianai.com

Website: blackindianai.com

CONCLUSION

The AI safety testing field is experiencing explosive growth. Companies are desperate for specialists who can secure their AI before disasters happen.

The opportunity is real:

- High demand (100:1 ratio)
- Strong compensation ($120K - $250K+)
- Meaningful work (prevent real harm)
- Future-proof career (AI adoption accelerating)

The barrier to entry is lower than you think:

- 6 months focused learning
- Leverage existing skills (QA, security, development)
- Portfolio matters more than credentials
- Community is supportive and growing

Your move:

The companies deploying AI this year need you. The question is: Will you be ready when they're hiring?

Start today. Follow the roadmap. Build your portfolio.

See you in the field.

Document Information

Author: Vincent Gaddis, Founder

Company: BlackIndian AI

Version: 1.0

Last Updated: December 2025

Questions about training or career transition?

training@blackindianai.com | blackindianai.com

END OF DOCUMENT