The Agentforce QA Checklist

A Developer's Framework for Testing AI Agents in Production

By A.I. Foil | ai-foil.vercel.app

About This Checklist

This checklist is designed for Salesforce developers and QA engineers deploying Agentforce AI agents. Use it to ensure your agents are production-ready, hallucination-resistant, and cost-effective.

Why Testing AI Agents Is Different:

- Traditional unit tests fail on non-deterministic outputs
- Context collapse happens in multi-turn conversations
- Hallucinations can fabricate data with high confidence
- API misuse leads to permission errors and data leaks

Pre-Production Checklist

✅ Unit Testing

- □ **Apex Test Coverage:** All agent actions have >75% test coverage
- □ **Action Validation:** Each action returns expected output structure
- □ **Data Validation:** Actions reject invalid inputs (malformed emails, null values)
- □ **Error Handling:** Actions gracefully handle Salesforce API failures

✅ Guardrails & Security

- □ **Out-of-Scope Rejection:** Agent refuses requests outside defined topics
- □ **Permission Enforcement:** Agent respects Salesforce sharing rules
- □ **Sensitive Data Protection:** Agent never exposes passwords, API keys, or PII
- □ **Adversarial Prompt Testing:** Agent resists jailbreak attempts

✅ Fallback Behavior

- □ **Uncertainty Detection:** Agent escalates to human when confidence is low
- □ **Graceful Degradation:** Agent provides helpful error messages (not "I don't understand")
- □ **Handoff Protocol:** Agent transfers context when escalating to support rep

Integration Testing Checklist

✅ Multi-Turn Conversations

- □ **Context Retention:** Agent remembers user info across 5+ turns
- □ **Entity Consistency:** Agent refers to the same Case/Account throughout conversation
- □ **Conversation Reset:** Agent correctly handles "start over" requests

✅ API & System Integration

- □ **Salesforce API Calls:** Agent correctly invokes SOQL/DML operations

- □ **External API Integration:** Agent handles third-party API timeouts gracefully
- □ **Data Consistency:** Agent updates CRM records without creating duplicates

✅ Topic Classification

- □ **Intent Accuracy:** Agent routes 90%+ of queries to correct topic
- □ **Cross-Topic Handling:** Agent recognizes when a query spans multiple topics
- □ **Ambiguity Resolution:** Agent asks clarifying questions when intent is unclear

---

Production Monitoring Checklist

✅ Hallucination Detection

- □ **Hallucination Rate:** <5% of responses contain fabricated data
  - Monitor for fake Case IDs, Account Names, or Product SKUs
  - Use RAG verification to cross-check against Salesforce data
- □ **Citation Accuracy:** Agent provides valid record IDs when referencing data

✅ Performance Metrics

- □ **Response Time:** 90% of queries respond in <3 seconds
- □ **Escalation Rate:** <10% of conversations require human handoff
- □ **Task Completion Rate:** >80% of conversations end with resolved query

✅ Cost Management

- □ **Einstein Credit Burn:** Monitor daily consumption vs. budget
- □ **LLM Token Usage:** Track tokens per conversation (aim for <2000 tokens)
- □ **API Call Volume:** Monitor Salesforce API limits (avoid hitting daily caps)

✅ Security & Compliance

- □ **Audit Logging:** All agent actions are logged for compliance review
- □ **Data Retention:** Conversation history follows GDPR/CCPA guidelines
- □ **Access Control:** Only authorized users can invoke sensitive agent actions

---

Critical Metrics: Benchmarks

Metric	Target	Red Flag
Hallucination Rate	<5%	>10%
Escalation Rate	<10%	>20%
Response Time	<3s (90th percentile)	>5s
Task Completion	>80%	<60%
Einstein Credit Burn	Within budget	150% of forecast
Topic Accuracy	>90%	<75%

Testing Tools Reference

Salesforce Native

- **Agentforce Testing Center:** Batch testing for topic classification and instruction adherence
- **Apex Test Classes:** Unit tests for agent actions
- **Einstein Trust Layer:** Monitor guardrail violations

Third-Party Frameworks

- **LangSmith:** Evaluation framework for LLM-based agents (trace logs, regression evals)
- **Promptfoo:** Open-source prompt testing with YAML configuration
- **RAGAS:** RAG-specific evaluation metrics (context precision, faithfulness)

When to Escalate to Production

Green Light Criteria:

1. All Pre-Production checks pass
2. Integration tests show >90% topic accuracy
3. Hallucination rate in testing is <3%
4. Cost per conversation is within budget
5. Security audit completed and signed off

Red Flags:

- Agent frequently hallucinates in testing
- Escalation rate >15% in pilot deployment
- Einstein Credit burn exceeds forecast by 50%+
- Security team identifies data leaks in logs

Why 40% of Projects Will Fail

Gartner predicts that over 40% of agentic AI projects will be canceled by 2027 due to:

- Unclear business value
- Inadequate risk controls
- **No testing strategy** ← You're fixing this right now

Don't be part of the 40%. Use this checklist to ship confidently.

Download the Full QA Guide

For the complete blog post with code examples (Apex + Python), hallucination detection patterns, and the 3-Layer Testing Model, visit:

ai-foil.vercel.app/blog/how-to-test-ai-agents-agentforce-qa-guide

---

About A.I. Foil

We help Salesforce teams build production-grade AI agents with governance, testing, and cost optimization. Follow us for more tactical guides on Agentforce development.

Website: ai-foil.vercel.app
Blog: ai-foil.vercel.app/blog
Contact: [Include your email]

Last Updated: January 2026
Version: 1.0