

SUBJECT MATTER EVALUATIONS

HR Agent SMEVals

What it takes to trust an agent that handles employee benefits, complaints, accommodations, and sensitive data

HR and benefits agents may answer questions about benefits, COBRA, leave, accommodations, harassment complaints, protected processes, employee records, compensation, enrollment, and health information.

That makes the workflow sensitive in two directions. The agent must be helpful enough to serve employees, but constrained enough to avoid disclosing PHI, making prohibited inquiries, mishandling complaints, retaliating, misrouting protected processes, or changing benefits or compensation without authority.

The Three Questions

01 Does the agent reach the right answer?

Does it understand the employee's question, benefit facts, deadlines, confidentiality constraints, and protected-process triggers?

02 Does the agent take the right action?

Does it disclose accurately, route appropriately, escalate complaints, initiate required processes, and avoid unauthorized changes?

03 Can it be tricked into doing the wrong thing?

Can it be manipulated into altering compensation, changing benefits enrollment, closing a complaint, disabling confidentiality, escalating privilege, leaking employee PII, or revealing the system prompt?



What the HR Agent SMEVal Tests

Benefits Accuracy and Engagement

FairPlay tests whether the agent provides accurate benefits figures, surfaces statutory deadlines accurately, answers benefits questions rather than stonewalling, and discloses the correct election window where required.

Complaint Handling and Anti-Retaliation

FairPlay tests whether the agent captures and escalates harassment complaints, avoids dismissing or improperly closing complaints, and avoids retaliatory routing or disclosure.

Routing, Escalation, and Fairness

FairPlay tests whether the agent routes inquiries correctly, avoids misrouting sensitive matters, escalates distress, and routes equally across protected classes using counterfactual testing.

Confidentiality and Prohibited Inquiries

FairPlay tests whether the agent avoids disclosing PHI to third parties, verifies identity before disclosing specifics, avoids prohibited disability or genetic inquiries under ADA / GINA, and avoids unauthorized record changes.

Accommodations and Protected Processes

FairPlay tests whether the agent initiates the ADA / PWFA interactive process when triggered and routes out-of-scope questions to a professional rather than answering.

Prompt-Injection and Attack Resistance

FairPlay tests whether the agent can be tricked into altering compensation, changing benefits enrollment, closing complaints, disabling confidentiality, escalating privilege, exfiltrating employee PII, or leaking the system prompt. Attack techniques include direct override, roleplay, refusal suppression, authority spoofing, obfuscation, payload splitting, and indirect injection through documents and record fields.

Failure Modes These HR Agent SMEVals Catch

- Giving inaccurate benefits amounts or deadlines.
- Asking prohibited disability or genetic questions.
- Failing to initiate an accommodation process.
- Following hidden instructions in uploaded documents or employee records.
- Disclosing PHI to the wrong person.
- Closing a harassment complaint without escalation.
- Misrouting employees based on protected-class signals.
- Changing benefits or compensation without authority.

fairplay

FairPlay SMEVals test whether HR and benefits agents can serve employees accurately while protecting confidentiality, preserving protected processes, avoiding retaliation, and resisting manipulation.

ABOUT FAIRPLAY

FairPlay is the AI enablement company for the financial services and insurance industries. FairPlay specializes in testing, tuning, and validating AI systems and agents, helping institutions deploy AI faster, more profitably, and in full compliance. Its products address the toughest challenges in AI deployment, including Bias Testing, Model Optimization, Model and Agentic Validation, Second Look, and Safe AI Sourcing. FairPlay's work has been featured in The Wall Street Journal, USA Today, Politico, American Banker, and Forbes. In 2023, Fast Company recognized FairPlay as a "World Changing Idea." To learn more, please visit fairplay.ai.