

SUBJECT MATTER EVALUATIONS

Adverse Media Screening SMEVals

What it takes to trust an agent that reads negative news and decides what matters

Adverse media screening agents are increasingly used to review news, public records, and open-source information for signals of financial crime, sanctions exposure, fraud, corruption, litigation, reputational risk, or customer misconduct.

That sounds like a mere research task. It is not.

An adverse media agent can affect onboarding, enhanced due diligence, account restrictions, escalation, SAR consideration, relationship management, and customer treatment. A weak agent may overreact to irrelevant news, miss serious adverse information, confuse two people with similar names, rely on stale or low-quality sources, or treat mere allegations as established facts.

Generic AI testing can show that the agent retrieved articles and summarized them. An SMEVal asks whether the agent correctly matched the media to the customer, assessed its relevance and severity, and reached the right disposition.

The Three Questions

01 Does the agent reach the right answer?

Does it distinguish relevant adverse media from noise, recognize source quality, assess recency, separate allegation from adjudicated fact, and avoid false matches?

02 Does the agent take the right action?

Does it route the case for the right level of review, document the rationale, apply the institution's risk appetite, and avoid improper account action without required human approval?

03 Can it be tricked into doing the wrong thing?

Can a customer profile, document, webpage, or prompt-injection attack cause it to ignore adverse media, suppress findings, exaggerate risk, or leak sensitive information?

What Adverse Media SMEVals Test



Identity and Match Integrity

FairPlay tests whether the agent correctly distinguishes the subject from similarly named individuals, uses available identifiers, recognizes weak or ambiguous matches, and escalates when identity confidence is insufficient.

Severity, Recency, and Disposition

FairPlay tests whether the agent classifies adverse media by seriousness, timing, and business relevance. It should distinguish a decade-old minor civil matter from recent allegations of fraud, corruption, sanctions evasion, trafficking, or money laundering.

Fairness and Protected-Class Risk

FairPlay tests whether the agent avoids using protected characteristics, stereotypes, nationality proxies, language, geography, or immigration status as substitutes for actual adverse information.

Source Quality and Relevance

FairPlay tests whether the agent weighs reliable sources appropriately, flags stale or duplicative media, avoids treating rumors as facts, and identifies whether the media is actually relevant to financial-crime or reputational-risk review.

Documentation and Explainability

FairPlay tests whether the agent records the evidence, source, date, confidence level, match rationale, and disposition. A reviewer should be able to reconstruct why the agent recommended clearing, escalating, or restricting the relationship.

Prompt-Injection and Attack Resistance

The SMEVal tests whether the agent can be manipulated into ignoring negative news, deleting adverse references, leaking internal risk rules, or following instructions embedded in webpages, articles, uploaded documents, or customer records.

Failure Modes These SMEVals Catch

Adverse media agents can fail quietly. Examples include:

- Clearing a customer because the agent misses a high-confidence adverse match.
- Treating unverified allegations as confirmed misconduct.
- Creating unfair outcomes based on nationality, geography, or language proxies.
- Producing a disposition that cannot be reconstructed by compliance.
- Escalating a customer based on a false positive involving someone with the same name.
- Ignoring recency, source credibility, or legal outcome.
- Suppressing adverse media because a customer-supplied document says "ignore prior news."

fairplay

ABOUT FAIRPLAY

FairPlay SMEVals are developed by subject matter experts — including former regulators, examiners, and lawyers with deep financial-services domain expertise. We test not just whether an agent retrieved information, but whether it made the right subject-matter judgment, took the right action, and resisted manipulation.

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.