

Prepara il tuo agente alla valutazione.

Una checklist pratica delle evidenze per sanità e finanza. Per ogni controllo identifica il responsabile, le prove disponibili e le lacune. Registra soltanto informazioni che sei autorizzato a condividere.

Agente, versione e perimetro di utilizzo

La compilazione di questa checklist non costituisce un esito di valutazione o una garanzia assicurativa. Restano necessari un riesame indipendente e un perimetro definito.

Governance e perimetro

Responsabilità, uso previsto, esclusioni e inventario versionato dell'agente valutato.

G01 · 0.1

☐ Responsabilità nominativa

Nominare responsabile, operatore tecnico e decisore indipendente; chi implementa non certifica il proprio lavoro.

Evidenze da preparare

- Matrice dei ruoli firmata con diritti decisionali.
- Dichiarazioni sui conflitti e prova di competenza del revisore.

Responsabile / riferimento della prova

G02 · 0.1

☐ Uso previsto ed esclusioni

Definire utenti, decisioni, canali, dati, azioni ed esclusioni in modo applicabile nelle operazioni.

Evidenze da preparare

- Dichiarazione d'uso approvata e percorso utente.
- Avvisi, regole di instradamento e blocchi visibili.

Responsabile / riferimento della prova

G03 · 0.1

☐ Inventario di sistema versionato

Vincolare la valutazione a versioni immutabili di modello, prompt, fonti, strumenti, regole e deployment.

Evidenze da preparare

- Inventario firmato e distinta delle dipendenze.
- Hash o identificativi degli artefatti valutati.

Responsabile / riferimento della prova

Dati e riservatezza

Flussi, minimizzazione, accesso, conservazione e protezione da divulgazioni involontarie.

D01 · 0.1

☐ Flussi mappati e minimizzati

Mappare input, derivazioni, destinazioni e finalità; rimuovere i campi non necessari al compito dichiarato.

Evidenze da preparare

- Diagramma dei flussi con località dichiarate.
- Revisione per campo di finalità e minimizzazione.

Responsabile / riferimento della prova

D02 · 0.1

☐ Accesso e conservazione

Applicare accessi per ruolo, separazione tenant, conservazione e cancellazione documentate a prompt, tracce, cache e backup.

Evidenze da preparare

- Matrice degli accessi e progetto di isolamento.
- Piano di conservazione con prove di cancellazione.

Responsabile / riferimento della prova

D03 · 0.1

☐ Resistenza alla divulgazione

Applicare filtri, autorizzazione del retrieval e isolamento affinché ogni richiedente riceva solo dati consentiti.

Evidenze da preparare

- Regole di autorizzazione del retrieval.
- Casi red-team di divulgazione diretta e indiretta.

Responsabile / riferimento della prova

Sicurezza delle integrazioni

Input non affidabili, permessi degli strumenti, segreti, dipendenze e separazione dei tenant.

S01 · 0.1

☐ Gestione degli input non affidabili

Trattare i contenuti esterni come dati, preservare la priorità delle istruzioni e bloccare l'autoautorizzazione.

Evidenze da preparare

- Modello dei confini di fiducia e regole dei prompt.
- Corpus di injection con esiti attesi.

Responsabile / riferimento della prova

S02 · 0.1

☐ Strumenti con privilegio minimo

Limitare ogni strumento a scope, record, azioni e durata minimi; separare tenant e ambienti.

Evidenze da preparare

- Catalogo strumenti con scope e responsabili.
- Configurazione delle credenziali e dell'isolamento.

Responsabile / riferimento della prova

S03 · 0.1

☐ Segreti, dipendenze e confini di esecuzione

Tenere i segreti fuori da prompt e codice, analizzare le dipendenze e isolare comandi o codice con limiti di risorse.

Evidenze da preparare

- Registri di gestione e rotazione dei segreti.
- Inventario dipendenze e regole della sandbox.

Responsabile / riferimento della prova

Affidabilità

Casi rappresentativi, output non supportati e regressioni per l'uso dichiarato.

R01 · 0.1

☐ Casi prestazionali rappresentativi

Versionare un set con casi normali, limite e ad alto impatto per l'uso svizzero, incluse le lingue supportate.

Evidenze da preparare

- Motivazione, dimensione e composizione del test set.
- Esito atteso e risultato rivisto per caso.

Responsabile / riferimento della prova

R02 · 0.1

☐ Gestione degli output non supportati

Esigere fonti tracciabili quando servono fatti e rifiutare o rivedere risposte con supporto debole o contraddittorio.

Evidenze da preparare

- Regole di grounding e qualità delle fonti.
- Casi con prove assenti, obsolete e contraddittorie.

Responsabile / riferimento della prova

R03 · 0.1

☐ Controllo delle regressioni

Eseguire regressioni basate sul rischio prima del rilascio e definire i cambiamenti che richiedono nuova valutazione.

Evidenze da preparare

- Gate di rilascio e storico regressioni.
- Matrice tra cambiamenti e test.

Responsabile / riferimento della prova

Supervisione umana

Approvazione, escalation e arresto delle azioni con conseguenze rilevanti.

H01 · 0.1

☐ Approvazione delle azioni rilevanti

Prima di prenotazioni, invii esterni, cancellazioni e avanzamenti clinici, legare l'approvazione nominativa ad azione, obiettivo e contenuto esatti.

Evidenze da preparare

- Inventario azioni e policy di approvazione.
- Tracce anonimizzate di richieste accettate, negate e scadute.

Responsabile / riferimento della prova

H02 · 0.1

☐ Escalation a una persona competente

Definire trigger osservabili, destinazione, contesto trasmesso e responsabilità di risposta.

Evidenze da preparare

- Matrice di escalation e responsabilità.
- Esempi con contesto e risoluzione.

Responsabile / riferimento della prova

H03 · 0.1

☐ Arresto e contenimento

Fornire controlli testati e indipendenti dall'agente per bloccare azioni, revocare credenziali, isolare un tenant e conservare prove.

Evidenze da preparare

- Procedura di arresto e contenimento.
- Verbale di esercitazione con ruoli e passaggi.

Responsabile / riferimento della prova

Trasparenza e impatto

Informazione sull'IA, decisioni comprensibili ed effetti sulle persone coinvolte.

T01 · 0.1

☐ **Informazione sull'interazione con l'IA**

Al momento pertinente indicare IA, finalità, principali usi dei dati e percorso verso una persona.

Evidenze da preparare

- Testi d'interfaccia per ogni canale.
- Revisione di tempistica e accessibilità.

Responsabile / riferimento della prova

T02 · 0.1

☐ **Esiti comprensibili**

Presentare base materiale, incertezza, passo successivo e revisione al livello del pubblico interessato.

Evidenze da preparare

- Modelli di esito e standard di spiegazione.
- Esempi rivisti per pubblici e lingue diverse.

Responsabile / riferimento della prova

T03 · 0.1

☐ **Valutazione d'impatto e gruppi coinvolti**

Identificare gruppi, danni plausibili, gravità, probabilità e mitigazioni prima di usi ad alto impatto e dopo cambiamenti.

Evidenze da preparare

- Valutazione d'impatto con contributi consultati.
- Responsabili dei rischi, mitigazioni e decisioni di accettazione.

Responsabile / riferimento della prova

Operazioni

Tracce che rendono evidenti le manomissioni, gestione degli incidenti e cambiamenti controllati.

O01 · 0.1

☐ Tracce degli eventi a prova di alterazione

Registrare identità, versione, riferimenti input, decisioni di policy, strumenti, approvazioni ed esiti in storage append-only o con rilevazione delle modifiche.

Evidenze da preparare

- Schema eventi, progetto d'integrità e regole di accesso.
- Traccia collegata tra agente, approvazione e strumento.

Responsabile / riferimento della prova

O02 · 0.1

☐ Risposta agli incidenti dell'agente

Definire severità specifica, contenimento, conservazione prove, valutazione delle notifiche e responsabilità di apprendimento.

Evidenze da preparare

- Playbook d'incidente e albero dei contatti.
- Esercitazione o incidente reale con azioni successive.

Responsabile / riferimento della prova

O03 · 0.1

☐ Cambiamento e revisione trimestrale

In questa bozza di schema, rivedere le prove almeno ogni trimestre e subito dopo cambiamenti materiali o incidenti definiti.

Evidenze da preparare

- Calendario e criteri di cambiamento materiale.
- Revisioni trimestrali, storico release ed eccezioni aperte.

Responsabile / riferimento della prova

Evidenze

Provenienza, riproducibilità, limiti della valutazione e rischio residuo.

E01 · 0.1

☐ Provenienza e integrità delle evidenze

Etichettare ogni artefatto con origine, raccoglitore, data, ambiente, versione e riferimento d'integrità, oscurando dati superflui.

Evidenze da preparare

- Registro delle evidenze con catena di custodia.
- Originali o hash in storage ad accesso controllato.

Responsabile / riferimento della prova

E02 · 0.1

☐ Verifica riproducibile

Registrare precondizioni, input, procedura, atteso, osservato e valutatore perché un secondo esperto possa ripetere.

Evidenze da preparare

- Test eseguibili o protocolli precisi.
- Manifest dell'ambiente e risultati conservati.

Responsabile / riferimento della prova

E03 · 0.1

☐ Perimetro e rischio residuo

Pubblicare versione, ambiente, uso, esclusioni, aree fallite o non testate, rischi residui e trigger di revisione accanto all'esito.

Evidenze da preparare

- Sezione firmata su perimetro e rischio residuo.
- Controllo di coerenza tra rapporto, registro e dichiarazioni pubbliche.

Responsabile / riferimento della prova
