

9ème édition de la
Soirée du Test
Logiciel
Sophia -
Antipolis



15 septembre 2026
17h à 23h



ETSI, Valbonne



Soirée du
test logiciel



Sophia Antipolis

Comment tester un agent IA ?



Julien Van Quackebeke.
ALL4TEST

Sommaire

- L'essor des agents IA
- Tester un agent IA : un défi pour les équipes de test
- Stratégies et techniques adaptées aux agents IA
- Deux retours d'expériences
- Résumé et ressources

L'essor des agents IA

Agent IA = LLM de raisonnement + mémoire + outils + garde-fous



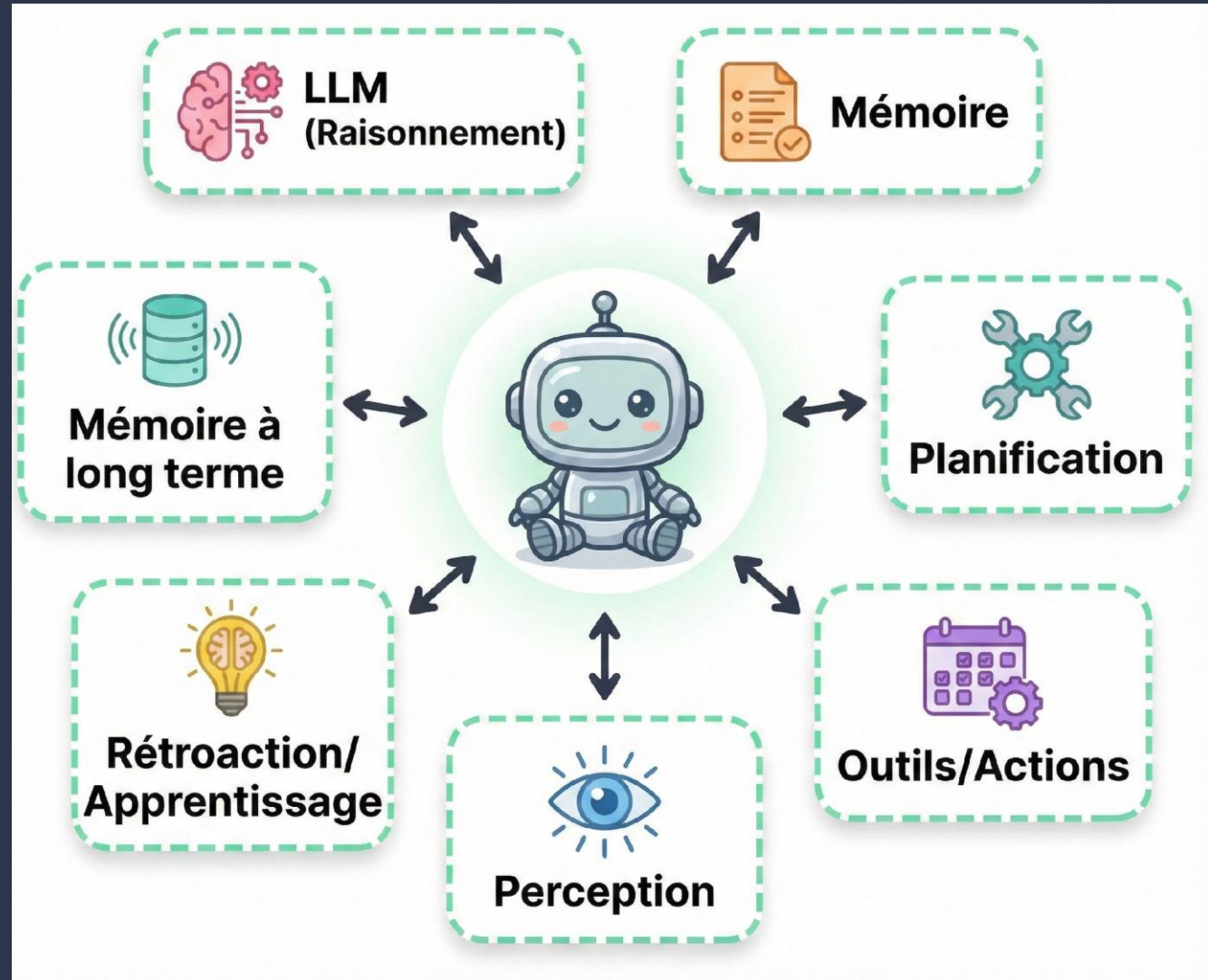
Capacités principales :

- Les agents IA utilisent des modèles IA avec raisonnement pour planifier et orchestrer la réalisation des tâches.
- Ils accèdent à des outils (API, outils tiers, navigateurs, OS) et les utilisent dans les limites fixées (les garde-fous).
- Ils possèdent une mémoire des actions réalisées et peuvent apprendre et s'améliorer.
- Ils peuvent établir un dialogue avec l'utilisateur pour poser des questions ou demander le consentement pour des actions sensibles.

Exemple d'agent IA – avec courte démo

Agent testeur

Exemple d'exécution
agentique d'un test



Exemple d'agent IA

Parcours d'inscription à la JFTL avec tutoriel

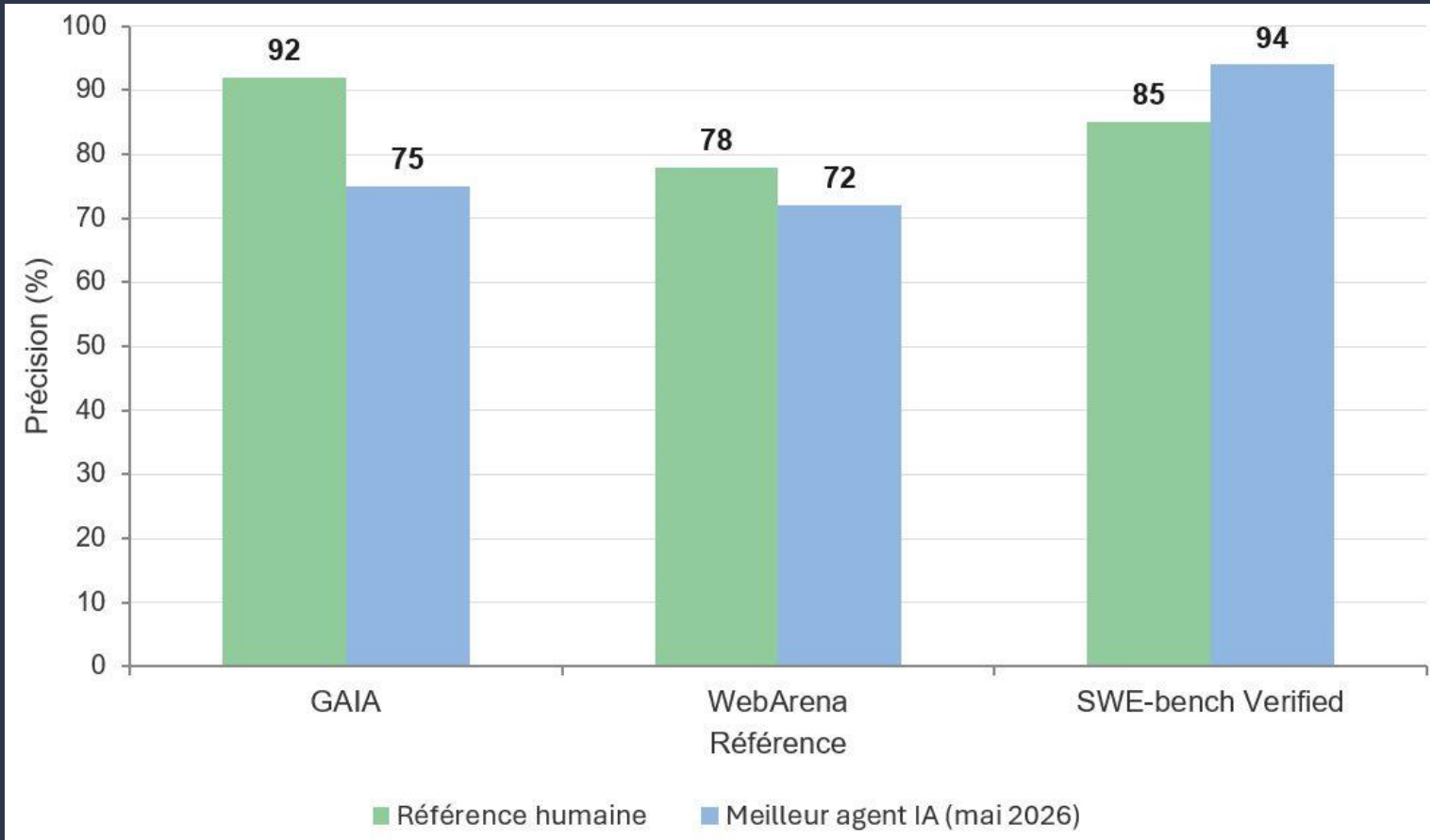
Tester le parcours d'achat d'un billet JFTL avec tutoriel

🌐 <https://cftl.fr/actualites/jftl/ticket/> 6 steps 🕒 v2

Steps

- 1. Accéder à la billetterie de la conférence JFTL (<https://cftl.fr/actualites/jftl/ticket/>)**
Expected: La page d'inscription s'affiche avec les différentes offres de billets disponibles et le stepper de progression (Commande, Coordonnées, Paiement, Récapitulatif).
- 2. Sélectionner 1 billet "Package JFTL + Tutoriels – Avec Certification (CFTL, GASQ, ISTQB, IQBBA, IBUQ, IREB, ReQB, TMMI)" pour les 8 & 9 juin 2026**
Expected: Le billet est ajouté au panier au tarif certifié de 730 € TTC. Le système permet de passer à l'étape suivante.
- 3. Renseigner les informations du participant : Sophie Martin, TestSoft Solutions, sophie.martin.test@example.com, 06 12 34 56 78**
Expected: Les informations sont acceptées sans erreur de validation.
- 4. Sélectionner les tutoriels de la journée du 8 juin : "IA fine-tuning" le matin et "Architecture d'agents IA" l'après-midi**
Expected: Les deux tutoriels sont enregistrés pour le participant.
- 5. Confirmer la possession d'une certification ISTQB (ou équivalent)**
Expected: L'éligibilité au tarif préférentiel est validée. Le système permet de passer au récapitulatif.
- 6. Vérifier le récapitulatif de la commande avant de procéder au paiement**
Expected: Le récapitulatif affiche fidèlement l'identité du participant, le billet sélectionné, les tutoriels choisis et le montant total de 730 € TTC.

Performances humaines vs performances des agents IA



**L'écart se réduit...
et s'inverse**

GAIA : de 65% à 75%

WebArena :
de 60% à 72%

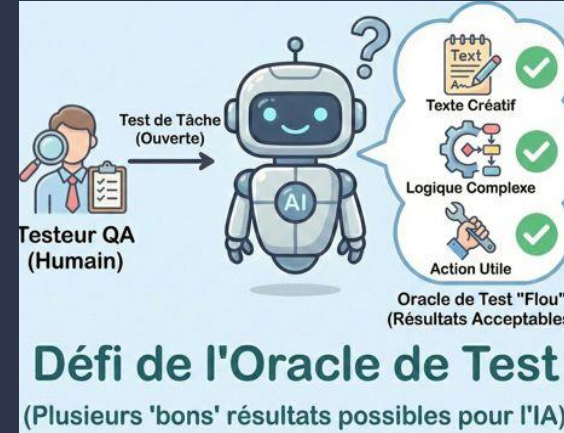
en 6 mois

GAIA, WebArena et SWE-bench sont trois benchmarks réputés de l'IA agentique.

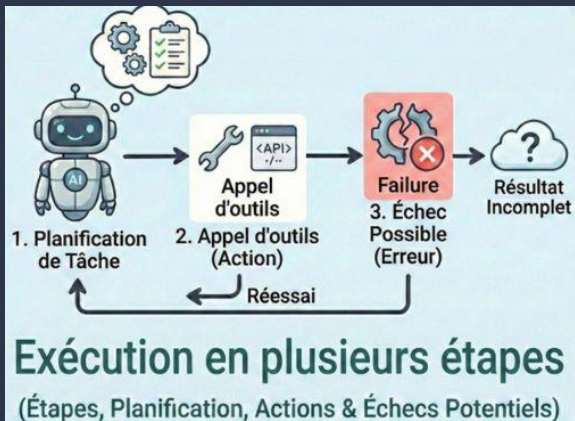
Tester les agents IA : 4 difficultés spécifiques



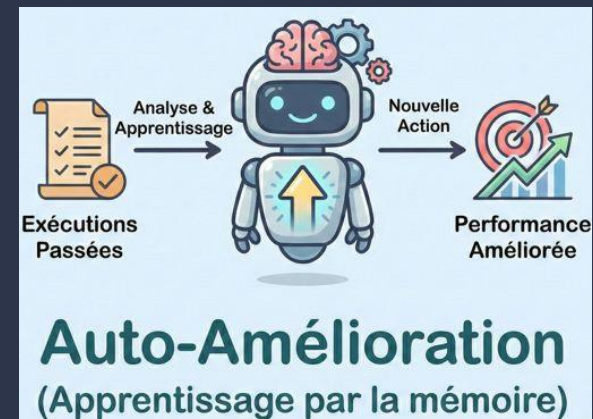
Usage des modèles de langage (LLM).



Comment établir le verdict de test ?



Un agent agit par étapes pour réaliser la tâche.



Le comportement de l'agent évolue.

Un changement de paradigme pour la QA

- Applications « traditionnelles »

Déterministe : mêmes entrées alors mêmes sorties / résultats

Niveaux de test : Composant (unitaire), Intégration, Système et Acceptation

Vérification d'un résultat attendu précis

- Agents IA

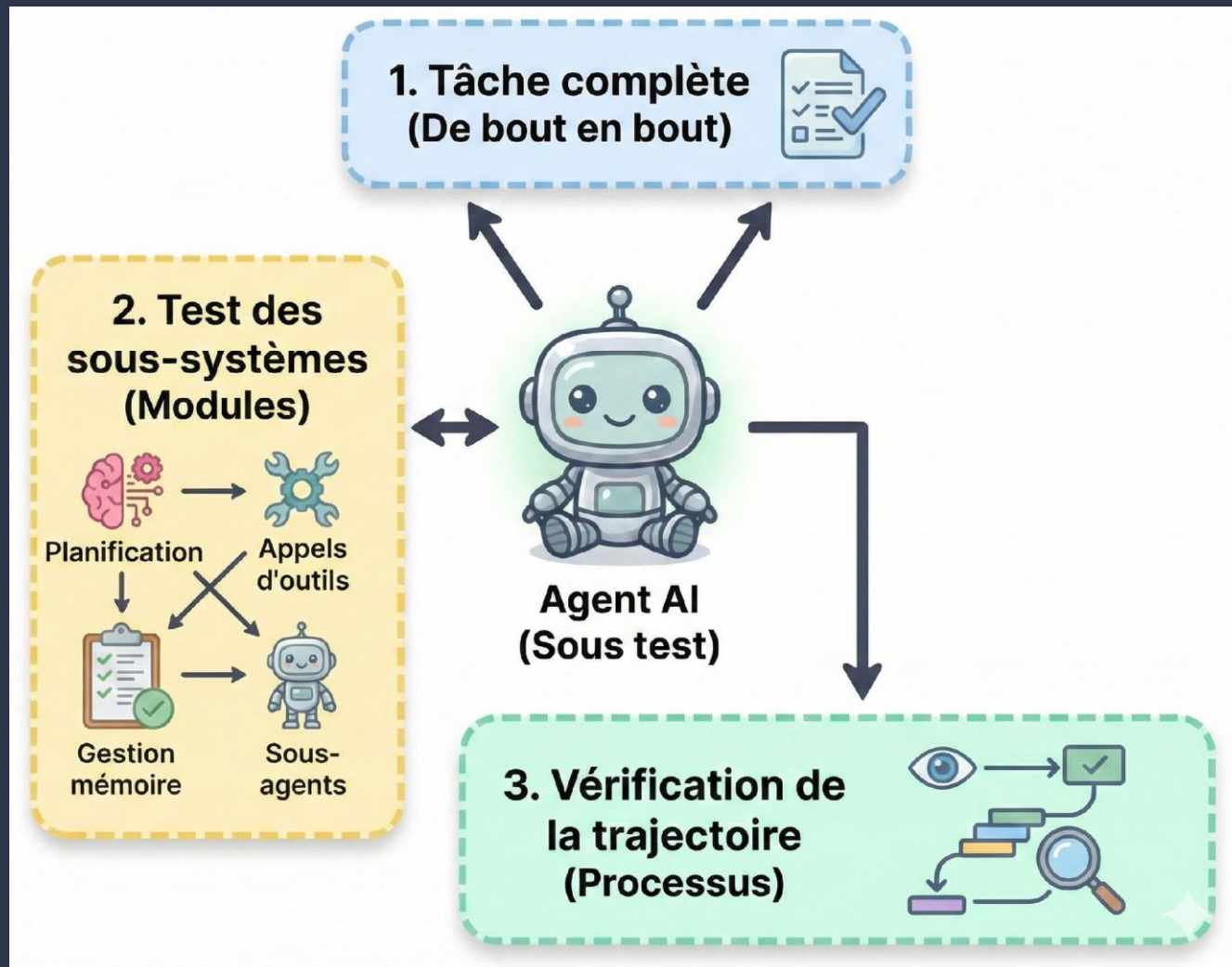
Non-déterministe : une requête mais des réponses acceptables « différentes »

Niveaux de test : Réalisation de la **tâche de bout-en-bout**, Test des **sous-systèmes** (planification, outils, mémoire), Vérification de la **trajectoire de l'agent**

Validation d'un comportement « acceptable »

Focus sur le comportement de l'agent – pas seulement le résultat

Tester les agents IA : besoin de méthodes et techniques spécifiques



Chacun des **niveaux de test** des agents IA met en œuvre des méthodes et techniques spécifiques.

Sommaire de la présentation

- L'essor des agents IA
- Tester un agent IA : un défi pour les équipes de test
- • Stratégies et techniques adaptées aux agents IA
- Deux retours d'expériences
- Résumé et ressources

Les 4 piliers de la qualité des agents



Ce que la QA peut mesurer (même lorsque les résultats varient) :

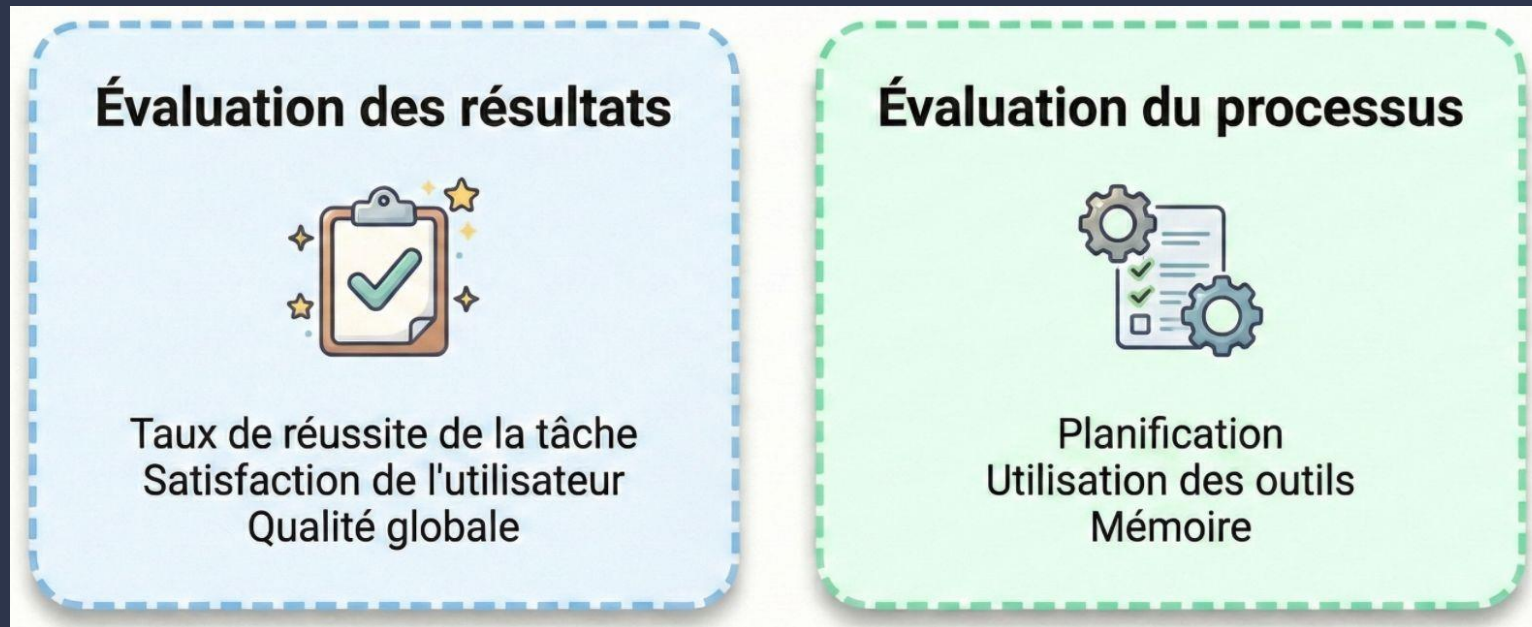
- **Efficacité** : réussite de la tâche, exactitude par rapport à l'intention de l'utilisateur
- **Efficience** : étapes, latence, coût par tâche

- **Robustesse** : nouvelles tentatives, dégradation progressive, questions de clarification vis-à-vis de la tâche à réaliser
- **Sécurité et alignement** : conformité aux politiques, confidentialité, résilience à l'injection de prompt malicieux

Stratégie de test des agents IA

Commencer par tester les résultats (boîte-noire) puis investiguer les composants et le processus de décision (ouvrir la boîte).

Niveau :
Bout-en-Bout



Niveaux :
Composants
• Trajectoire

Méthodes de test

Combiner plusieurs méthodes



- **Métriques automatisées** : le fondement pour tester les agents
- **Evaluation humaine** : pour qualifier la qualité sur un ensemble de tâches
- **Agent « as a judge »** : l'IA est un bon « critique »
- **Retours utilisateurs et préférences** : quel feedback de vos utilisateurs clés

Évaluation de la boîte noire (E2E) : les résultats d'abord

Définir le succès à partir de la **valeur délivrée par l'agent**, puis **automatiser l'évaluation** pour contrôler la qualité.

Indicateurs E2E courants :

- Taux de réussite des tâches (binaire ou noté)
- Satisfaction des utilisateurs (pouce vers le haut/vers le bas, note)
- Exhaustivité/précision des tâches délimitées
- Indicateurs clés de performance (usage, acceptation des résultats, ...)

Artefacts d'assurance qualité :

- Jeux d'essai de tâches utilisateur représentatives – « Golden set »
- Critères d'acceptation + grille d'évaluation
- Critères de sortie des tests (par exemple, aucune régression sur les 50 tâches principales)
- A/B pour mesurer les résultats réels des utilisateurs

Évaluation de la boîte de verre : la trajectoire

Lorsqu'un agent échoue, vous déboguez le chemin, pas seulement la réponse.

Analyse du processus de l'agent



1. **Planification** de la tâche



2. **Exécution** des actions



3. Utilisation de la **mémoire**



Investiguer les traces de l'agent :

- Outil mal choisi ou paramètres incorrects (problèmes JSON/schéma)
- Analyse du raisonnement
- Hallucinations ou pollution du contexte
- Nombre d'étapes/reprises excessives (coût + latence)

Stratégies de test du non-déterminisme

1 LLM/Agent en tant que juge

Utiliser un autre LLM/Agent pour évaluer la similarité sémantique au lieu de la correspondance exacte.

Indicateurs :

- Note pour l'exactitude
- Note pour la pertinence
- Note pour l'exhaustivité

2 Test basé sur les propriétés

Valider les caractéristiques de sortie plutôt que le contenu exact.

Vérifier des propriétés :

- Structure JSON valide
- Respect de la limite de caractères
- Ton approprié

3 Variabilité du comportement

Effectuer plusieurs tests et suivre les seuils de variance.

Indicateurs :

- Note moyenne sur l'ensemble des essais
- Seuils de variance
- Intervalles de confiance

4 Seuils de défaillance

Remplacer la réussite/l'échec par des fourchettes de notation graduées.

Échelles de notation :

- 0,8-1,0 : réussite
- 0,5-0,8 : avertissement
- Inférieur à 0,5 : Échec

Évaluation humaine et par agent IA

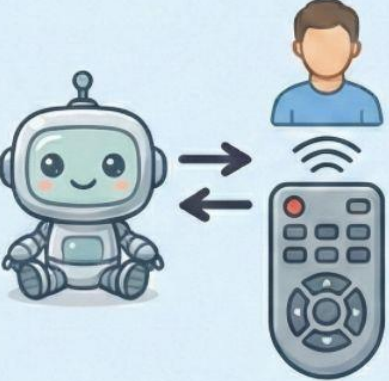
Complémentaire : l'évaluation humaine « incarne » le ressenti humain, et l'évaluation IA facilite l'extension des cas testés.



La supervision humaine des agents IA

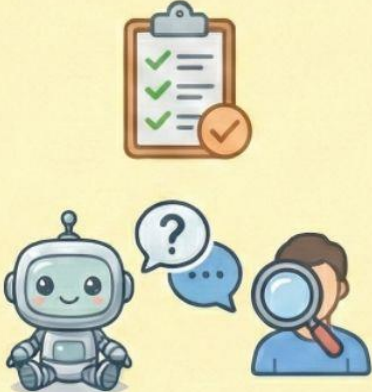
Trois modes : synchrone, asynchrone et hybride

SUPERVISION SYNCHRONE



Contrôle en temps réel du comportement de l'agent par l'humain.

SUPERVISION ASYNCHRONE



Dialogue après la tâche pour clarifications ou corrections.

SUPERVISION HYBRIDE



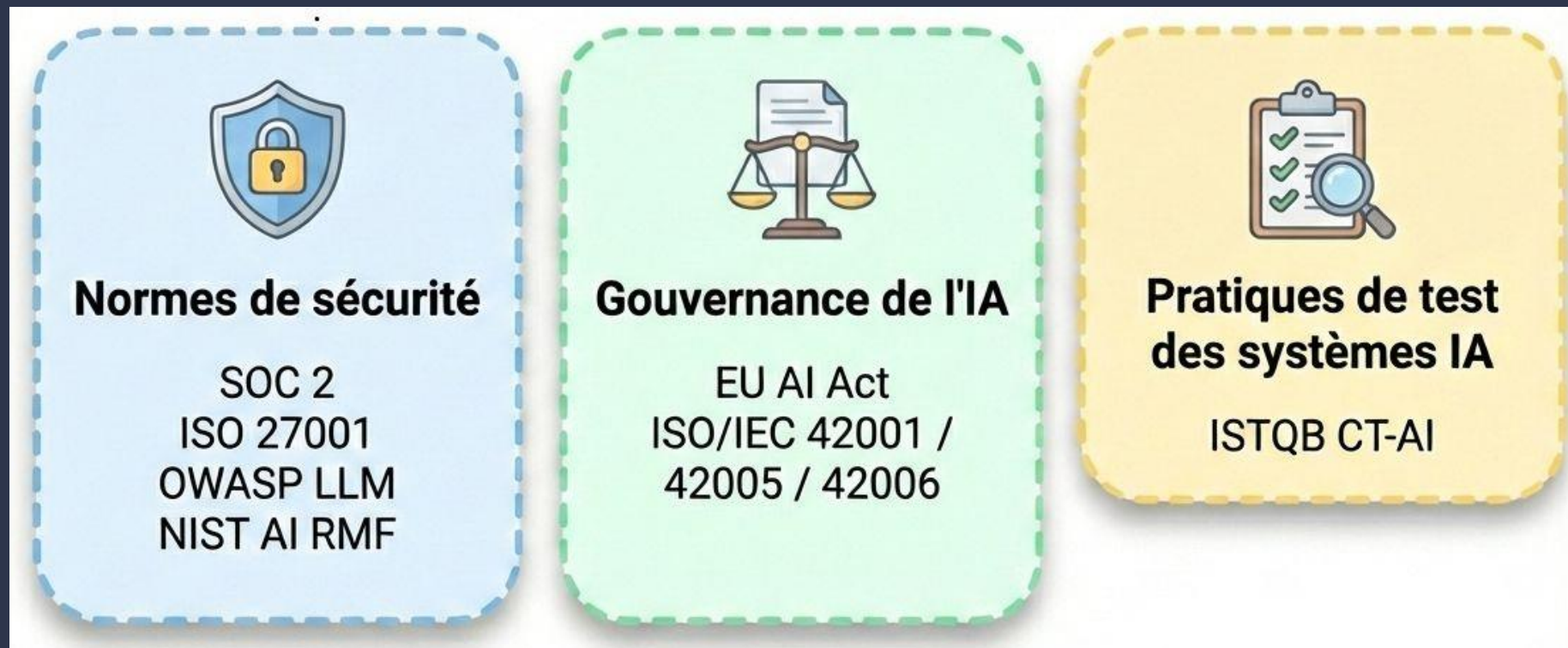
Consentement humain requis pour les actions critiques.

EU AI Act
Supervision humaine obligatoire si risques avec possibilité de passer outre à l'IA (opt-out)

Tester la fiabilité de la supervision humaine fait partie des objectifs de test

Au-delà des tests : assurer la conformité des agents IA

Les agents IA font partie du système d'information et doivent respecter les normes et réglementations adaptées à l'IA.



ISO/IEC 42001:2023 — système de management de l'IA (AIMS)

Première norme internationale dédiée à la gestion de l'IA en entreprise.

Elle définit un cadre (politiques, objectifs, processus, cycle Plan-Do-Check-Act) pour gérer de façon responsable les risques et enjeux propres à l'IA (éthique, transparence, apprentissage continu), sans entrer dans le détail technique des systèmes. Publiée le 18 décembre 2023.

ISO/IEC 42005:2025 — évaluation d'impact des systèmes d'IA

Norme apportant un cadre méthodologique pour réaliser des évaluations d'impact des systèmes d'IA (AI system impact assessment) — c'est l'équivalent, pour l'IA, d'une étude d'impact.

ISO/IEC 42006:2025 — exigences pour les organismes de certification

Complète l'ISO/IEC 17021-1 avec des exigences spécifiques à l'IA pour les organismes qui audient et certifient les systèmes de management de l'IA (AIMS) définis par l'ISO/IEC 42001. Publiée en juillet 2025

Obligations qui en découlent pour les agents IA

Déjà applicable : interdiction de manipulation nuisible / exploitation de vulnérabilités.

Depuis le 2 août 2026 : obligation de transparence si l'agent interagit avec des personnes (informer qu'il s'agit d'une IA)

2027-2028 : pour les agents qualifiés "haut risque" (ex. recrutement, scoring crédit, services publics), exigences supplémentaires — supervision humaine, traçabilité, documentation
Pour les modèles généralistes sous-jacents à "risque systémique", obligations de gestion des risques incluant les capacités autonomes

Répartition des obligations par rôle

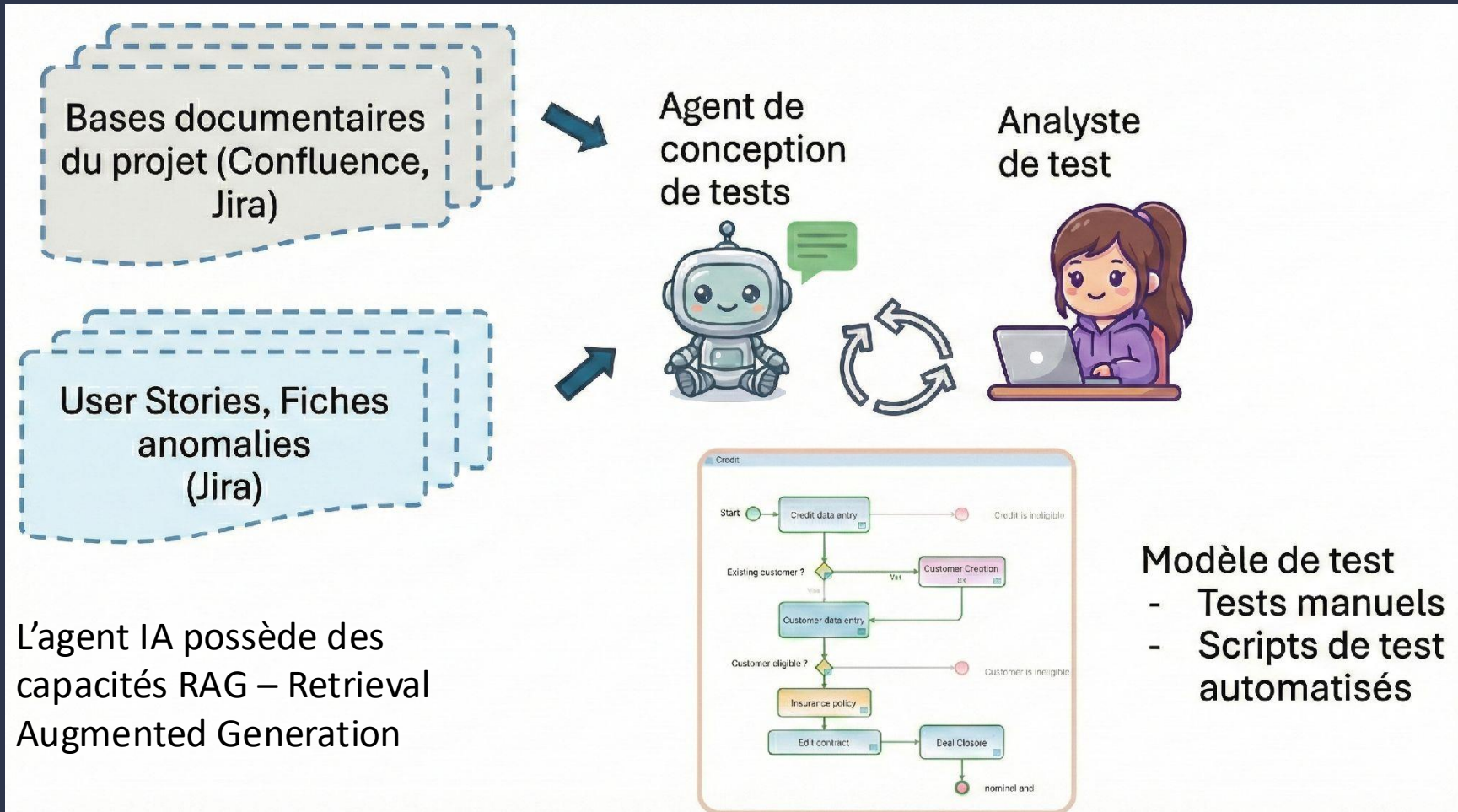
Fournisseur (celui qui crée/développe l'agent) : documentation des capacités et limites, gestion des risques sur le cycle de vie, journalisation des actions si haut risque

Déploieur (celui qui utilise l'agent en entreprise) : qualifier le niveau de risque selon l'usage réel (pas seulement selon le degré d'autonomie), mettre en place une supervision humaine, vérifier la conformité des agents tiers avant déploiement

Sommaire de la présentation

- L'essor des agents IA
- Tester un agent IA : un défi pour les équipes de test
- Stratégies et techniques adaptées aux agents IA
- • Deux retours d'expériences
 - Un agent de conception de tests avec RAG
 - Un agent d'exécution des tests fonctionnels
- Résumé et ressources

Agent de conception de tests avec RAG



Exemple de requête : « *Met à jour le parcours de test avec les nouvelles US #26 et #32.* »

Caractéristiques et tests de l'agent

Caractéristiques spécifiques

- Extraction des connaissances par RAG
- Génération de modèles graphiques de test et de conditions de tests dans des tables
- Tâches diversifiées de conception de test
- Supervision humaine synchrone de l'agent

Méthodes de test utilisées

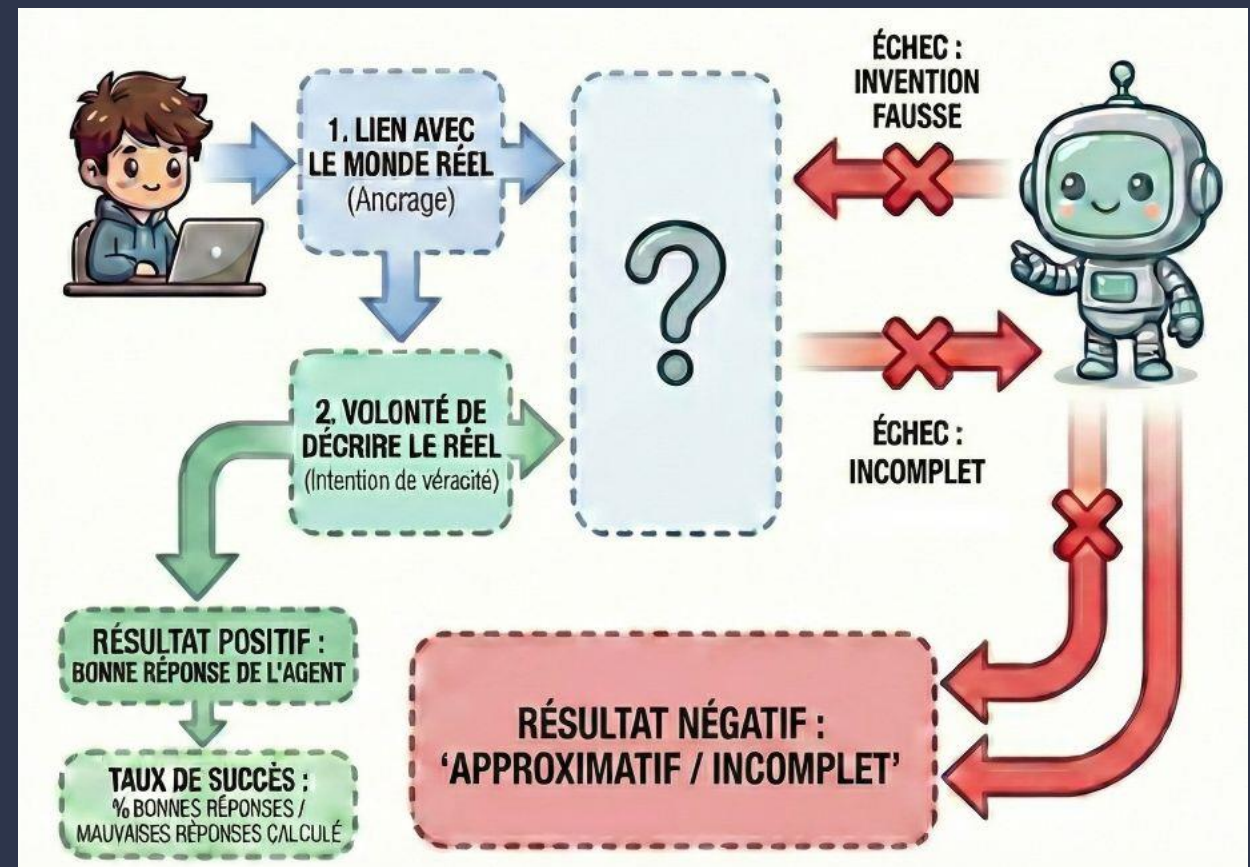
- Évaluation humaine par « *jeu de vérité* » pour le RAG
- Constitution d'un « *golden set* » de tâches / réponses représentatives
- Données manuelles et données générées
- Ateliers réalisés avec des utilisateurs clés

Tests du RAG par la technique du jeu de vérité

Popularisée par Murray Shanahan dans « Talking about LLMs » - 2022

Le testeur est un expert du domaine métier.

Les tests sont des questions réponses à l'agent RAG avec évaluation des *résultats corrects* et des *résultats approximatifs / erronés / incomplets*.



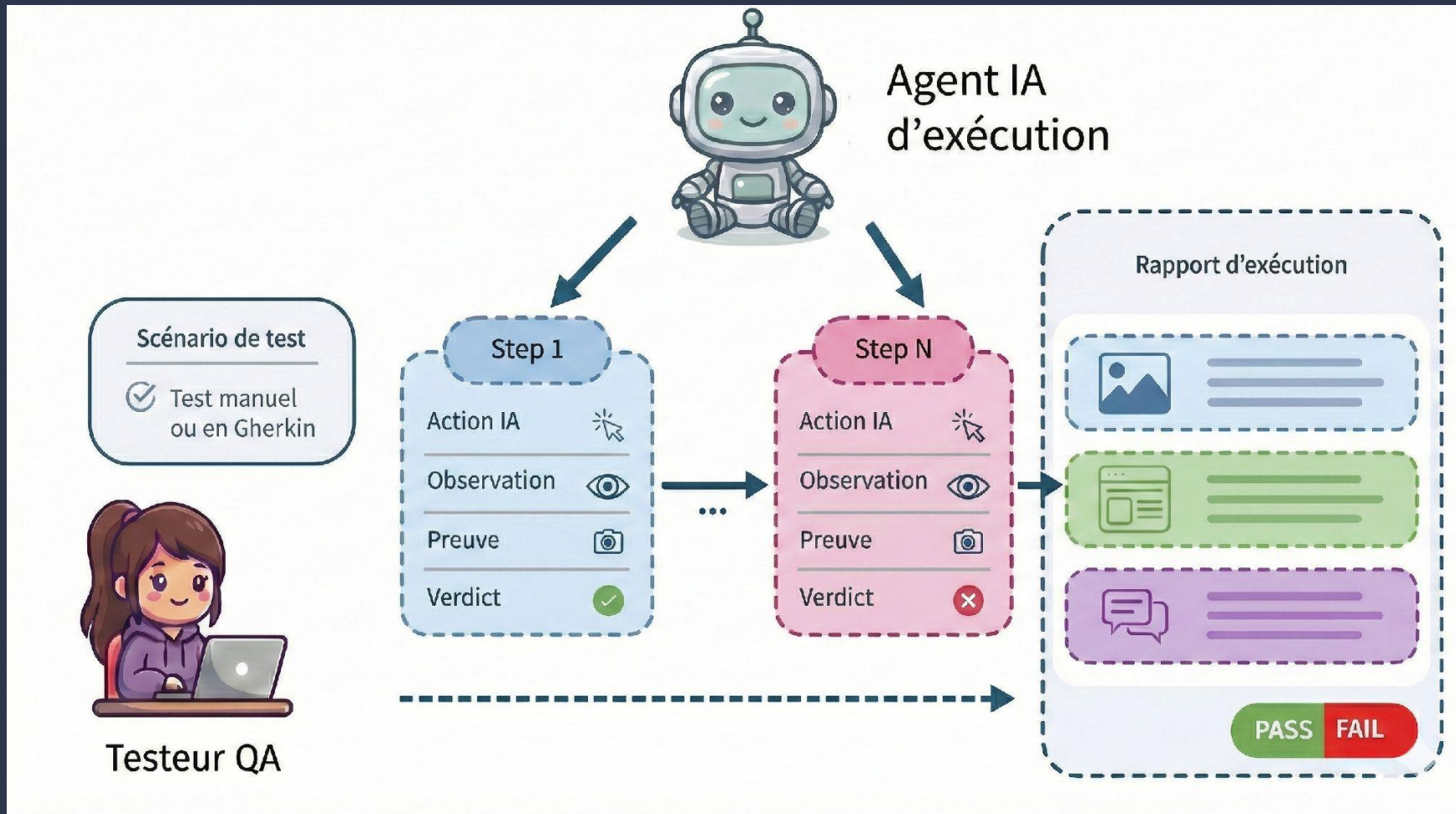
Données quantitatives et Bilan / Bonnes pratiques

- Test du RAG de l'agent : 350 items dans le jeu de données de test, dont 50 humains et 300 générés
- Test de la création de parcours de test : 5 projets exemples
- 4 ateliers avec des utilisateurs clés

Bilan / Bonnes pratiques

- La génération de jeux de données par IA est très pertinente et a permis une forte progression de l'agent
- Procéder par itération courte : évaluation / adaptation / réévaluation

Agent d'exécution des tests fonctionnels



Le **testeur QA** lance les exécutions et valide le rapport.

L'**agent IA** exécute les étapes de test sur l'IHM et produit les rapports à chaque étape et un verdict de test.

Caractéristiques et tests de l'agent

Caractéristiques spécifiques

- Agent qui agit sur l'interface (IHM) et qui vérifie le résultat attendu sur les captures d'écran
- L'agent commente chaque étape de test (détails des actions réalisées, analyse détaillée pour le verdict par étape du test)
- Supervision humaine asynchrone : revue des preuves d'exécution par l'utilisateurs

Méthodes de test utilisées

- Jeux de test représentatifs : des scénarii de test construits sur des applications diversifiées.
- Métriques automatisées : % tests exécutés correctement, % variabilité, ...
- Infrastructure pour monitorer les exécutions et évaluations de tests
- Évaluation humaine des résultats d'exécution
- par agent IA

Une infrastructure de test ad-hoc pour l'agent

Testeur de
l'agent IA



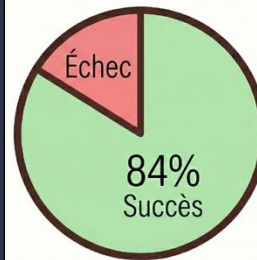
Agent IA
d'exécution des
tests
(l'agent testé)



Pilote, harnais de test et IHM de l'infrastructure de test de l'agent IA

Jeux d'essai
(ensemble de
scénarios de test
à exécuter)

Agent Dashboard Métriques



- ✓ Taux de succès : 84%
- ✗ Erreurs : 16%
- 🕒 Temps moyen : 1.2s
- 📉 Variabilité : 3%
- 📄 Scénarios : 500

**Traces
d'exécution de
l'agent IA
(logs)**

Infrastructure
de test de
l'agent IA

Données quantitatives et Bilan / Bonnes pratiques

- Jeux d'essai : plus de **700 scénarios de test** représentatifs sur plus de **50 applications diversifiées**
- Plus de **100 campagnes d'évaluation de l'agent** lancées en 18 mois (variation des architectures, des versions de l'agent, ...) avec l'infrastructure de test
- 12 sessions suivies avec des **utilisateurs pilotes**

Bilan / Bonnes pratiques - Construire une **infrastructure de test dédiée** pour :

- Enregistrer les traces d'exécutions de l'agent (logs de test)
- Mesurer et calculer les métriques (fournir un dashboard)
- Gérer et maintenir les jeux d'essai représentatifs
- Gérer les campagnes d'évaluation (benchmarking) des architectures et des versions de l'agent
- Mesurer la progression de l'agent

Sommaire de la présentation

- L'essor des agents IA
- Tester un agent IA : un défi pour les équipes de test
- Stratégies et techniques adaptées aux agents IA
- Deux retours d'expériences
- • Résumé et ressources

Résumé – Ce qu'il faut retenir

Nouveau paradigme QA

Les agents IA nécessitent une évaluation sémantique et de comportement, et non des assertions déterministes.

Niveaux de test

Tests de la tâche complète (bout-en-bout), par composant et sur la trajectoire de l'agent.

Établir la confiance

Au cœur de la conception, du développement et des tests des agents IA pour la fiabilité et la sécurité.

Supervision humaine

Supervision synchrone, asynchrone ou hybride, et qui est aussi à tester.

Techniques de test

Agent as a Judge, évaluation du résultats attendus gradués, test de la variabilité.

Conformité

Les agents sont un composant du SI qui doivent respecter les normes de conformité (SOC 2, ISO 27001, EU AI Act, ...).

Ressources

Benchmarks et évaluation

- **GAIA Benchmark.** Mialon et al. (2023). "GAIA: A Benchmark for General AI Assistants." arXiv:2311.12983. Meta-FAIR & HuggingFace.
- **AgentBench.** Liu et al. (2023). "AgentBench: Evaluating LLMs as Agents." arXiv:2308.03688. Tsinghua University.
- **SWE-bench.** Jimenez et al. (2024). Princeton University. swe-bench.github.io
- **WebArena.** Zhou et al. (2024). Carnegie Mellon University. webarena.dev

Sécurité

- **OWASP.** "Top 10 for LLM Applications 2025." genai.owasp.org
- **PromptGuard Framework.** (2025). "Injection Resilient Language Models." Nature Scientific Reports.
- **Microsoft.** (2025). "Defending Against Indirect Prompt Injection." Microsoft Security Blog.

Standards et Frameworks

- **NIST.** "AI Risk Management Framework 2.0" (2024) & "Gen AI Profile" AI 600-1. nist.gov
- **ISO/IEC 42001:2023.** "AI Management Systems." International Organization for Standardization.
- **EU AI Act.** (2024). European Parliament Regulation on AI. eur-lex.europa.eu
- **ISTQB.** "Certified Tester AI Testing (CT-AI)" Syllabus. istqb.org

Rapports techniques

- **Anthropic.** (2025). "Writing Effective Tools for AI Agents." anthropic.com/engineering
- **Confident AI.** "LLM Agent Evaluation Guide." DeepEval Documentation. deepeval.com
- **Google.** "Agent Quality." kaggle.com/whitepaper-agent-quality

Vous souhaitez mettre en place ces bonnes pratiques ?

Besoin d'un coach ou d'une formation certifiante sur le sujet ?

contact@all4test.com

Bonne Soirée 😊