Une IA d'exécution des tests manuels : Quel impact pour le rôle de Testeur ?

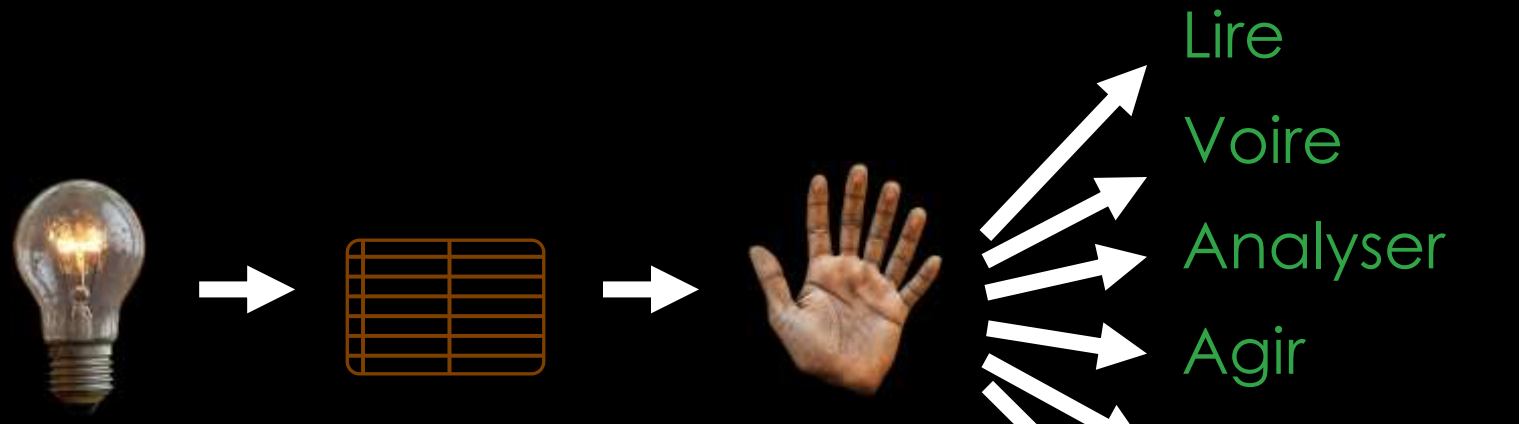
JFTL 2025

Eric Gavaldo

Xavier Blanc

TEST MANUEL

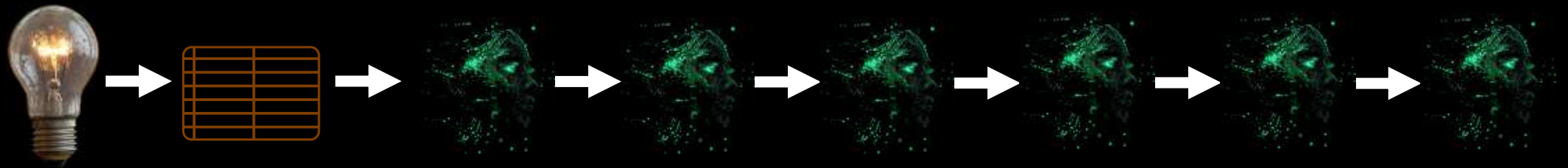
- Que fait un testeur lorsqu'il exécute manuellement des tests ?
  - **Lire** (les instructions de la procédure de test)
  - **Voir** (un écran, des images, des boutons, des champs, …)
  - **Analyser** (à quoi ils servent, significations des textes, tableaux, images, …)
  - **Agir** (cliquer, sélectionner, déplacer, taper du texte, …)
  - **Comparer** (un résultat observé par rapport au résultat attendu)
  - **Discuter** (avec les 2 autres amigos s'il y a des ambiguïtés, soumettre un bug, …)



L'IA DANS LE TEST MANUEL

- **Lire ➔** LLM, NLP
- **Voir ➔** Image2Text, ImageClassification, LLM
- **Analyser ➔** Classification, Detection, Estimation
- **Agir ➔** Sélectionner l'appel de commande, MCP
- **Comparer ➔** Detection
- **Discuter ➔** Text2Text, Answering

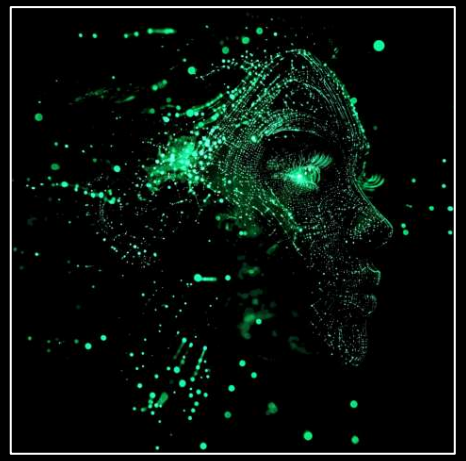
AI TestRunner (AGENT)



Lire Voir Analyser Agir Comparer Discuter

Testeur Concepteur

?

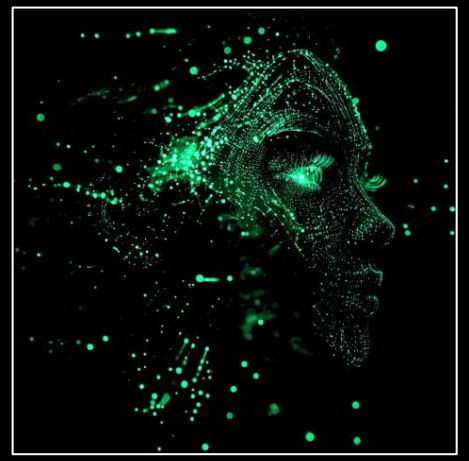


AI TestRunner

Agent IA d'exécution des tests manuels

Réalisation d'un AI TestRunner

- Un prototype pour évaluer :

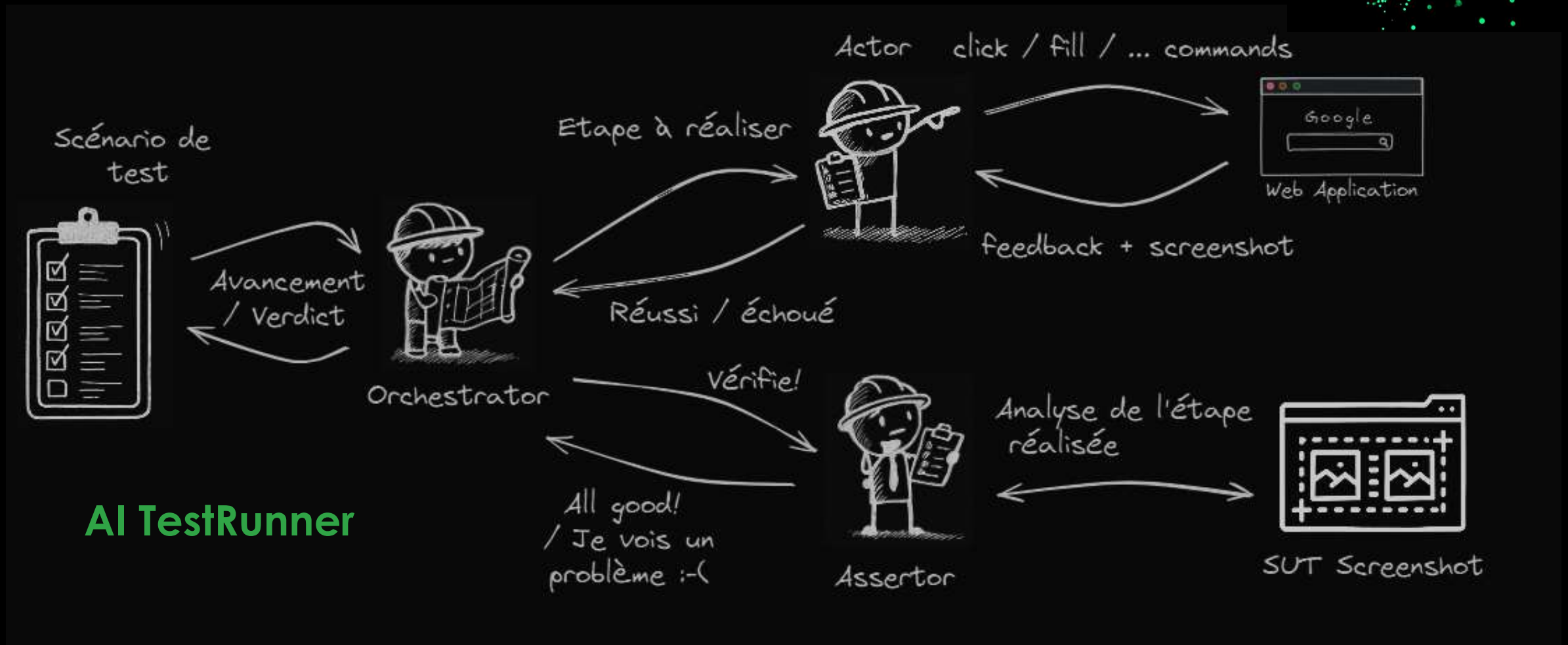


AI TestRunner

- **Architecture Agent IA :** Prototypé par le Labo IA de Smartesting
- **Intégration dans un outil de Test Management :** Les cas de test manuels et leur exécution par le co-testeur IA sont lancés depuis XQual

ARCHITECTURE - AGENT IA

Division des tâches avec trois agents : **Orchestrator, Actor, Assertor**



Validation par Benchmark Reproductible

AI TestRunner

Evaluer la performance d'un testeur exécuteur IA à exécuter un test :

- **3 applications** relativement représentatives Classified, Postmill et OneStopMarket (avec données et dockerisées)
- **~100 cas de tests manuels** rédigés par des testeurs professionnels (tests "passed" et "failed")

ISSTA 2025
co-located with FSE 2025

25-28 juin 2025

➔ Benchmark entièrement reproductible et accessible en open source

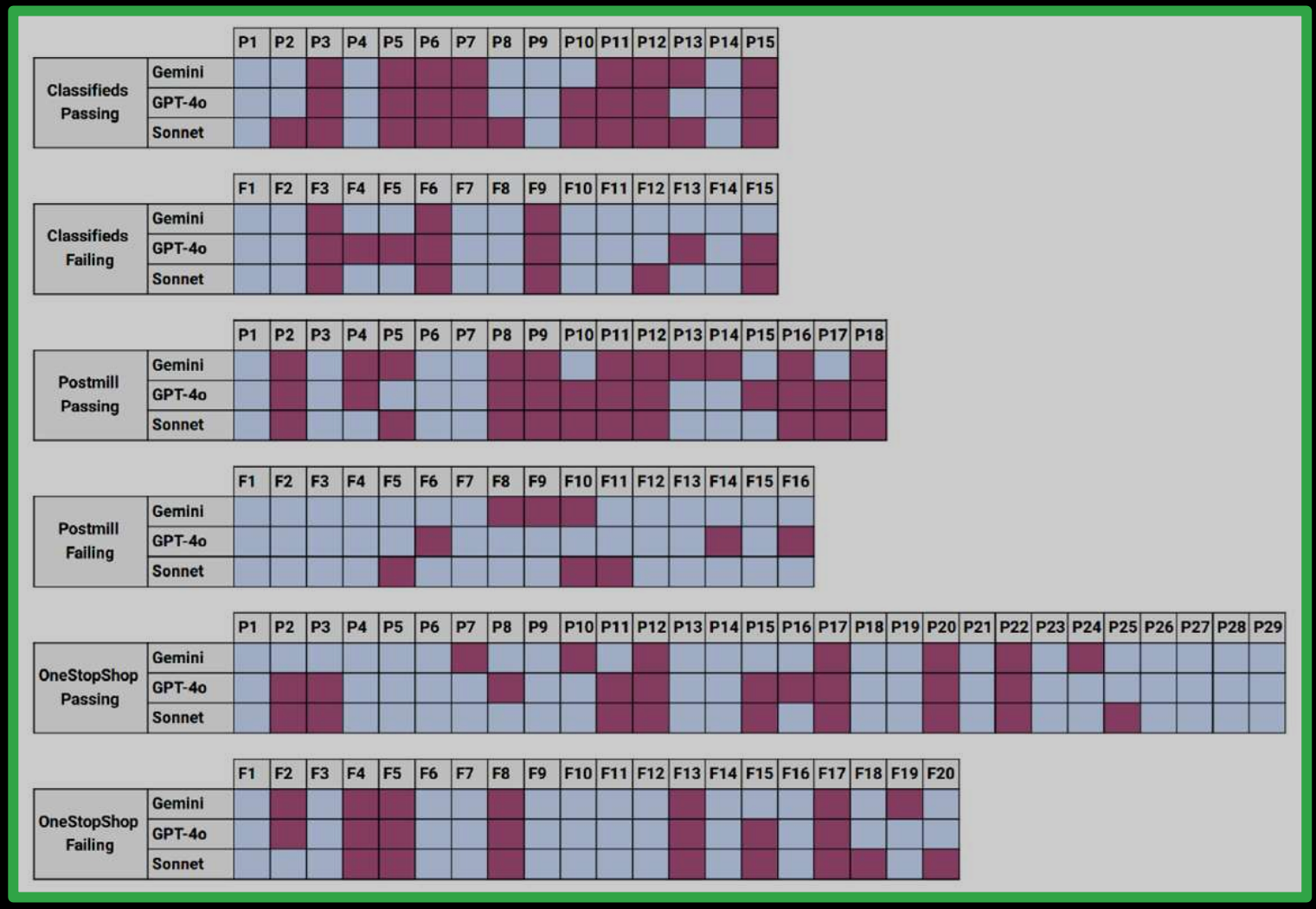
- https://zenodo.org/records/15198569 (sources)
- https://arxiv.org/abs/2504.01495 (publication scientifique)

RÉSULTATS DU BENCHMARK

Model	Application	Evaluation Metrics						
		Acc	Spec	Sens	AER	HER	SMER	TruAcc
GPT-4o	Classified	0.64	0.47	0.81	0.08	0.08	0.15	0.48
	Postmill	0.76	0.61	0.94	0.00	0.06	0.06	0.71
	OneStopShop	0.73	0.62	0.90	0.09	0.02	0.11	0.63
	Average	0.71	0.57	0.88	0.06	0.05	0.11	0.61
Sonnet	Classified	0.60	0.27	0.93	0.07	0.07	0.14	0.47
	Postmill	0.71	0.44	1.00	0.06	0.03	0.09	0.62
	OneStopShop	0.78	0.69	0.90	0.09	0.05	0.14	0.65
	Average	0.69	0.47	0.94	0.07	0.05	0.12	0.58
Gemini	Classified	0.70	0.47	0.93	0.07	0.04	0.11	0.60
	Postmill	0.62	0.44	0.81	0.04	0.04	0.07	0.56
	OneStopShop	0.78	0.76	0.80	0.10	0.03	0.13	0.71
	Average	0.70	0.56	0.85	0.07	0.03	0.10	0.62

Le prototype de AI TestRunner réussit **61%** des exécutions de tests manuels

RÉSULTATS DU BENCHMARK



26 tests ne sont pas exécutés par notre agent, quelque soit le LLM utilisé

RÉSULTATS DU BENCHMARK

Capacité : Besoin d’un meilleur cadre
Polyvalence : Besoin de stratégies d’action intelligentes
Observabilité : Besoin d’une meilleure analyse du UI
Vérifiabilité : Besoin d’un moteur de vérification dédié
Conformité : Besoin d’un meilleur orchestrateur

5 axes d’amélioration

Le AI TestRunner n'est pas un super héro

- L'exécution aboutit à des incertitudes
  - **Succès**
  - **Échec**
  - **Non-exécutable** (l'IA ne réussit pas à 'comprendre' ce qu'il faut faire)
  - **Incertain** (l'IA n'est pas sûre de son résultat)

- 2 questions se pose donc :
  - Comment avoir confiance envers les **Succès** remontés ?
    - Relecture/vérifications manuelle
    - Historique, bugs remontés, comparatifs avec exécutions manuelles, non-régressions
  - Comment gérer les **faux échecs**, **non-exécutables** et **incertains**
    - Améliorer les procédures (effet collatéral +)
    - Améliorer l'IHM (effet collatéral +)
    - Améliorer l'IA (alpha)

SUCCÈS

Analyse ➔ ‘vrai ’ positif

Dev

Test

AI TestRunner

Métier

SUCCÈS

Analyse ➔ ‘faux’ positif
➔ Correction de la procédure
➔ Faisabilité ?
➔ Dialogue/Enrichissement
➔ Correction du code
➔ Correction de l’exigence

Même en cas de succès, toujours **douter** et **vérifier**

Dev

Test

AI TestRunner

Métier

ECHEC

Analyse ➔ ‘vrai’ échec
➔ Correction du code
➔ Correction de l’exigence

Dev

Test

AI TestRunner

Métier

ECHEC

Analyse ➔ ‘faux’ échec

- ➔ Correction de la procédure
- ➔ Amélioration de l’IHM
- ➔ Faisabilité ?
- ➔ Dialogue/Enrichissement

Attention aux incompréhensions

Dev

Test

AI TestRunner

Métier

NON-EXECUTABLE

- Correction de la procédure
- Amélioration de l'IHM
- Faisabilité ?
- Dialogue/Enrichissement

L'agent ne comprend pas ou ne sait pas interpréter la procédure

Dev

Test

AI TestRunner

Métier

INCERTAIN

- ➔ Correction de la procédure
- ➔ Faisabilité ?
- ➔ Amélioration de l'IHM
- ➔ Dialogue/Enrichissement /Confirmation/Infirmation

} Faible taux de confiance

Dev

Test

AI TestRunner

Métier

TESTEUR CONCEPTEUR: DE NOUVELLES TÂCHES

Rédaction
Triage
Analyse
- Vérification
- Reporting

Rédaction
Triage
Analyse
+ Pilotage
+ Supervision
+ Arbitrage

TESTEUR CONCEPTEUR: DE NOUVELLES COMPÉTENCES

Adaptabilité
Curiosité
Compétence de communication
Esprit collaboratif
Sens de l'organisation
Rigueur
Créativité
- Patience
- Résilience
- Capacités analytiques
- Compétences techniques (automatisation, architecture)

Adaptabilité
Curiosité
Compétence de communication
Esprit collaboratif
Sens de l'organisation
Rigueur
+ Créativité profonde
+ Intelligence émotionnelle
+ Intuition

ET DEMAIN?

- Conception des procédures
- Priorisation
- ⊕ Arbitrage des 'non-conclusive'
- ⊕ Validation des procédures et bugs générés
- ⊕ Tests exploratoires (➔ Procédures)
- Analyse des bugs

- ⊕ Validation du code
- Débuggage assisté

Dev

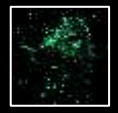
Test

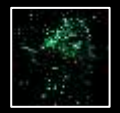
AI TestRunner

- Exécution des tests
- Rapport d'exécution
- ⊕ Génération des bugs

4ème Amigo

- ⊕ Génération de code





- ⊕ Génération des procédures

Métier

- Rédaction assistée
- ⊕ Revues/scoring
  - SMART
  - INVEST
- ⊕ Génération de Datasets
- ⊕ Validation des datasets

Avec une IA d'exécution à mes côtés, mon rôle se transforme : positif / négatif ?