



HUB
FRANCE
IA

EVALUATION DE L'IA AGENTIQUE

Septembre 2026

EASYVISTA



A propos du Hub France IA

Le Hub France IA est une association loi 1901 d'intérêt général dont la mission est d'accompagner l'adoption d'une IA responsable, éthique et souveraine et de soutenir son déploiement en France et en Europe. Association indépendante, son rôle est d'éclairer et influencer le débat sur les grands enjeux de l'IA.

Le Hub France IA rassemble Grands groupes, PME & ETI, startups, ESN et éditeurs, collectivités, institutions, services de l'état, écoles, fonds d'investissement, cabinets de conseils et juridiques, tous mobilisés pour produire ensemble les communs de l'IA nécessaires à son accélération.

Guidé par la rigueur scientifique et technique, le Hub France IA produit de nombreuses publications annuelles pour accompagner les entreprises, les citoyens et les pouvoirs publics vers une meilleure compréhension de l'IA, diffuser des bonnes pratiques et sensibiliser à l'impact de l'IA.



Table des matières

Introduction.....	7
Pourquoi évaluer les agents IA ?	7
L'efficacité : un critère déterminant	8
La complexité de l'évaluation : multiplicité des composants et non-déterminisme.....	9
Un champ d'étude en pleine structuration.....	9
Exemples illustratifs	10
Audience	12
1 Concepts	14
1.1 Agentique.....	14
1.1.1 Boucle agentique et appels d'outils.....	14
1.2 Observabilité des LLM et notion de traces	17
1.2.1 Définition des traces et rôle des plateformes d'observabilité.....	18
1.2.2 Les traces, des données indispensables à collecter en continu	19
1.3 Raisonnement et planification	20
1.4 Trajectoire d'un agent IA	22
1.4.1 Trajectoires basiques pour des IA conversationnels interagissant avec l'utilisateur ...	23
1.4.2 Trajectoires avec appels d'outils	23
1.4.3 Trajectoires problématiques : coût, sécurité	24
1.4.4 Trajectoires complexes, systèmes multi-agents	26
1.5 Validation des agents IA en tant que systèmes logiciels.....	26
1.6 Juge LLM – « LLM-as-a-judge »	28
1.7 Désalignement.....	29
2 Métriques	33
2.1 Métriques d'évaluation des agents IA.....	33
2.2 Évaluation globale de la performance	35
2.3 Qualité de la recherche documentaire	36
2.3.1 Précision	36
2.3.2 Rappel.....	37



2.3.3	F-score (F1 score).....	37
2.3.4	Précision@k / Rappel@k.....	37
2.3.5	Qualité des citations.....	37
2.4	Efficienc e de l'agent IA.....	37
2.4.1	Coûts.....	37
2.4.2	Latence.....	38
2.4.3	Débit.....	38
2.4.4	Efficacité des étapes.....	38
2.5	Choix des outils.....	38
2.5.1	Exactitude de sélection des outils.....	38
2.5.2	Validité des arguments.....	38
2.6	Résilience et robustesse.....	38
2.6.1	Robustesse face aux outils distracteurs.....	39
2.6.2	Résilience aux pannes.....	39
2.6.3	Taux de récupération.....	39
2.6.4	Score de cohérence.....	39
3	Méthodologies d'évaluation	41
3.1	Principe directeur.....	41
3.2	Axes de décision.....	41
3.2.1	Axe 1 : structure de la tâche.....	41
3.2.2	Axe 2 : profil de risque du déploiement.....	41
3.2.3	Axe 3 : priorité des parties prenantes.....	42
3.3	Quelques méthodes d'évaluation	42
3.3.1	Évaluation par benchmarks statiques.....	42
3.3.2	Évaluation des trajectoires : une discipline en soi.....	42
3.3.3	Comparaison à une trajectoire de référence.....	44
3.3.4	Simulation et <i>Sandbox</i>	45
3.3.5	Évaluation adverse.....	45
3.3.6	Évaluation en production.....	46
3.3.7	<i>OpenAI</i> - Évaluation macroscopique par partitionnement des traces.....	46



4	Outils	49
4.1	Outils de Gouvernance	50
4.1.1	Control Planes et Control Towers	50
4.1.2	Fonctionnalités de gouvernance	51
4.1.3	Transformer les principes de gouvernance en contrôles	53
4.1.4	Articulation avec les outils de monitoring et d'évaluation	53
4.2	Outils d'observabilité, de monitoring et d'évaluation des agents IA	54
4.2.1	<i>OpenTelemetry</i> et <i>OpenLLMetry</i> : un standard pour les données d'observabilité des IA génératives et agentiques	54
4.2.2	<i>LangSmith</i> : la plateforme de monitoring, d'évaluation et de déploiement de l'écosystème <i>LangChain</i>	55
4.2.3	<i>Langfuse</i> : plateforme libre pour le monitoring des applications génératives et agentiques	56
4.2.4	<i>Mastra</i> : un <i>framework</i> et un studio de développement <i>open source</i> pour l'IA agentique en <i>TypeScript</i>	57
4.2.5	Inspect (UK AI Safety Institute)	58
4.2.6	Agent Trajectory Explorer	58
4.2.7	Outils pour l'évaluation des modèles de langage	59
4.2.8	<i>Datadog</i> : extension du monitoring des infrastructures et des applications aux applications agentiques	59
4.2.9	<i>Giskard</i> : un <i>control plane</i> pour l'évaluation des agents IA	61
4.2.10	Konverso / Easyvista	62
4.2.11	Prisme	63
4.2.12	Mankinds	63
4.2.13	<i>Idun Platform</i> – <i>control plane</i> pour la mise en production des agents IA	64
4.2.14	<i>Ripple Tide</i> – Évaluation à chaud via <i>Graphes</i> en production	64
5	Benchmarks	67
5.1	Benchmarks pour agents IA généralistes	68
5.1.1	GAIA	68
5.2	Benchmarks pour agents IA à base de <i>tool use</i>	69
5.2.1	StableToolBench	69



5.3	<i>Benchmarks</i> pour agents IA de navigation web	71
5.3.1	WebArena	71
5.3.2	MIND2WEB.....	73
5.4	<i>Benchmarks</i> pour agents IA bureautiques	74
5.4.1	SWE-bench.....	74
5.4.2	SWE-Bench Pro.....	76
5.4.3	DELEGATE 52.....	78
5.5	Aller plus loin.....	78
6	Ouverture	82
6.1	Créer un agent IA : une tâche devenue accessible, mais pas anodine.....	82
6.2	Vers l'alignement : au-delà de l'évaluation	85
	Conclusion	87
	Références	90
	Glossaire	93
	Remerciements	99