Liste de vérification de la préparation des agents

Ceci est une étape préalable au transfert, à effectuer une fois avant de laisser un agent accomplir une tâche de manière autonome, et non une révision du travail déjà effectué. Un agent est un modèle auquel on a donné des outils et un objectif, et qui est autorisé à boucler de manière autonome : lire, décider, agir, observer, répéter. La liste de vérification impose un oui ou un non à une question. Cette tâche spécifique est-elle sécuritaire à transférer, et si oui, avec quelles balises et à quel niveau d'autonomie.

Définir la portée de la tâche

- [ ] L'objectif est écrit, avec une condition d'arrêt claire que l'agent peut reconnaître comme « terminé ».
- [ ] La tâche est suffisamment précise pour être décrite en une phrase; ce n'est pas « gérer tout dans la boîte de réception ».
- [ ] Vous avez nommé la pire action unique que l'agent pourrait entreprendre, et exactement ce qu'elle toucherait.
- [ ] Vous savez à quel point le pire scénario est réversible et approximativement combien de temps un retour en arrière prendrait.

Balises de sécurité avant la première exécution

- [ ] Chaque action irréversible (supprimer, envoyer, facturer, déployer) est soit retirée, soit soumise à une approbation humaine.
- [ ] L'agent ne dispose que des outils et de la portée nécessaires à cette tâche, avec des identifiants limités, pas un jeton d'administration partagé.
- [ ] Vous savez quelles données entrent, où vont les sorties, et que rien ne franchit une limite qu'il ne devrait pas.
- [ ] Chaque point d'approbation humaine remet en question l'intention, la provenance des données et le rayon d'impact; il ne peut pas être approuvé automatiquement.
- [ ] Chaque appel d'outil, entrée et sortie est consigné dès la première exécution, dans un endroit qu'une personne peut lire plus tard.
- [ ] Une personne désignée peut annuler le travail de l'agent en utilisant un chemin que vous avez réellement testé.

Définir le niveau d'autonomie

- [ ] Vous avez choisi un niveau intentionnellement : suggérer seulement, agir avec approbation, ou agir et rapporter.
- [ ] Vous pouvez expliquer en une phrase pourquoi la tâche mérite ce niveau aujourd'hui, pas un jour.
- [ ] Vous connaissez les principales façons dont la tâche peut échouer (boucles, mauvais outil, injection de prompt) et qui est alerté.
- [ ] Vous avez un déclencheur pour élargir l'autonomie plus tard (par exemple, un nombre d'exécutions consignées sans erreur).