Liste de vérification pour l'évaluation des prompts

Ce guide pratique évalue un prompt par ses résultats à travers de nombreuses exécutions, après qu'il a été exécuté, en fonction d'un ensemble fixe d'entrées et d'une grille d'évaluation écrite. C'est le complément post-résultat de la Liste de vérification pour la révision des prompts, qui vérifie un seul prompt avant que vous ne l'envoyiez. Utilisez ceci lorsqu'un prompt alimente un travail réel et que vous devez savoir si une modification l'a amélioré, détérioré ou simplement modifié.

Construire le harnais

- [ ] Fixer un ensemble d'entrées de 20 à 50 entrées réelles, tirées de la production et des échecs passés, qui ne changent pas entre les exécutions.
- [ ] Écrire la grille d'évaluation : chaque dimension (exactitude, fondé, format, ton, sécurité) a une définition et un seuil de réussite.
- [ ] Joindre une réponse de référence ou des critères de réussite clairs à chaque entrée pour qu'un évaluateur puisse distinguer réussite et échec.
- [ ] Fixer et enregistrer les conditions : la version du modèle, la température et le prompt système restent les mêmes à chaque exécution.

Exécuter et évaluer

- [ ] Évaluer chaque entrée 3 à 5 fois, pas une seule, pour que la variance entre les exécutions soit visible.
- [ ] Conserver le score pour chaque entrée, pas seulement la moyenne, pour qu'une correction qui brise un cas ne puisse pas se cacher.
- [ ] Évaluer l'ancien prompt et le nouveau prompt sur les mêmes entrées lors de la même exécution, et comparer la différence par cas.
- [ ] Lire vous-même les transcriptions à faible score, car une moyenne ne vous dit jamais pourquoi quelque chose a échoué.

Décider et protéger

- [ ] Bloquer tout changement où une entrée qui passait un contrôle de sécurité ou de refus échoue, même une seule fois.
- [ ] Enregistrer la décision (livrer, retenir ou revenir) avec le nombre et la date pour que la prochaine personne hérite des preuves.