

Corriger les invites ou le contexte plutôt que la sortie de l'IA

Il existe des invites utilisateur et des invites système. Lorsque la sortie ne fonctionne pas correctement, nous devrions ajouter la règle dans les invites utilisateur plutôt que de la corriger à chaque fois.

Comme mentionné dans un essai précédent, Les noms dans la traduction automatique de l'IA, il existe de nombreux noms chinois en double. Même les humains ne peuvent pas toujours traduire correctement les noms anglais en noms chinois, donc c'est encore plus difficile pour les machines.

Il n'est pas difficile de définir les règles pour la traduction des noms. Dans la vie réelle, le contexte est souvent nécessaire. Par exemple, dans une classe d'environ 30 personnes, si deux personnes ont le même nom en anglais (traduction en Pinyin), il devient difficile de les distinguer lors de la traduction. Des informations supplémentaires, comme leur apparence, sont alors nécessaires.

Cependant, dans certains contextes—comme le blog de quelqu'un, une plateforme de cours ou la liste de contacts d'un utilisateur—la liste des noms cibles possibles devient beaucoup plus réduite. Dans ces cas, l'IA peut traduire parfaitement.

Par conséquent, nous devrions définir ces règles dans nos appels d'API. Ne modifiez pas la sortie ; corrigez la cause racine au lieu d'ajuster le contenu généré temporairement. Concentrez-vous sur la résolution de la raison, pas sur le résultat.

Les éditeurs de code IA sont plus intelligents que les chatbots IA car ils disposent d'un contexte plus large. Cela leur permet de faire des inférences plus précises sur leur sortie.

Il en va de même pour les images, l'audio et la vidéo. Ces outils devraient offrir un contexte plus large. Par exemple, si des outils de création d'IA reçoivent une série de vidéos, de clips audio et de podcasts, ils peuvent générer de nouveaux contenus de manière beaucoup plus efficace.

Cela ressemble au RAG (Retrieval-Augmented Generation). Pour les outils de création d'IA, ils doivent trouver un équilibre — produire des résultats qui ne soient ni trop spécifiques ni trop génériques. Cependant, ils devraient offrir des outils ou des fonctionnalités pour rendre les résultats plus précis. ChatGPT, par exemple, dispose d'une fonctionnalité de projet qui vous permet de télécharger des fichiers et d'interagir avec eux. D'autres outils de création ont besoin de fonctionnalités similaires.