Mail

Mail de Marine du 18/10/24

Hello Cyril,

Merci bcp pour ces réponses !

J’ai justement demandé à bluenove hier de chiffer si on passe par eux pour collecter données. Je te reviens dès que j’ai les résultats mais d’ici là explorons de notre côté.

Pas de pb pour le format de la BD.

Pour le formulaire 2023, je viens de faire la demande côté bluenove pour accès à ce que nous avions mais d’ici là tu peux regarder ici tout le texte (intro, questions, etc.) et les images utilisées. À noter que nous ne voulons pas autant de questions ouvertes ! J’aimerais 4 à 6 gros max si OK avec l’UNESCO – ce qui semble être le cas.

Aussi, il y a la question des serveurs sécurisés pour l’hébergement.

Dernier point, je me demandais si tu pensais judicieux de faire un essai du point 1 de l’analyse (voir ci-dessous) avec la BD de Youth Talks Edition 1. On a en effet des réponses aux questions ouvertes et des réponses à un questionnaire empathy dont réponses sont un score. Je te mets en p.j. un échantillon de la BD (1000 lignes) pour que tu vois. Si tu penses que cela serait utile alors je peux te fournir la BD complète. Je ne me rends pas compte du temps que cela prendrait de le faire pour une seule question du questionnaire empathy. Par exemple pour la Q1 du questionnaire empathy, les jeunes ont évoqué ci et ça. Je ne sais pas si cela peut bien prédire ce que le LLM pourra faire pour YT Edition 2. Je ne sais pas si cela serait rassurant pour toi ou pour nous. Je te laisse me revenir avec tes réflexions là-dessus mais j’ai pensé que j’allais te le mentionner.

FYI : Pour Edition 2, j’aimerais peut-être que les questions soient toutes obligatoires… à valider avec l’équipe mais cela nous ferait des lignes plus complètes. Je te joints également d’autres documents qui décrivent la BD etc.

Qu’en penses-tu ?

Merci Cyril - tout cela commence à bien prendre forme et je pense que nous allons vers une aventure intéressante.

Bon we !

Marine

1. Identifying and Correlating Key Topics with MHQ Scores\*:

The open-ended responses will be analyzed alongside the MHQ scores to identify the most frequently cited topics related to each of the six major areas of the MHQ. Using large language models (LLMs), we will uncover the key themes that resonate most with young people based on their mental health status, as measured by the MHQ. This will allow us to highlight which issues are most prevalent among youth with different levels of mental health, giving insight into how these issues correspond to the varying strengths and challenges within each of the six major mental health domains.

\*This process will be carried out in three major steps:

• Associating the cited topics with positive or negative MHQ scores.

• Determining the frequency of these topics to identify the most cited ones.

• Refining the analysis by associating the most cited topics with specific MHQ score ranges (-100, -50, 0, 50, 100, 150, 200).

Réponse

Bonjour Marine

Faire un site Web avec 4 à 6 questions sur le modèle de l’édition 1 est vraiment facile et basique.

Je travaille depuis de nombreuses années avec des personnes qui savent faire cela très bien.

Je vais les contacter la semaine prochaine pour sonder leurs disponibilités et le coût.

Je te confirmerai des tarifs quand tu me donneras ton go pour des rencontres commerciales avec les agences web.

Si Bluenove est dans la même gamme de prix il est plus sécurisant de passer par eux.

J’ai beaucoup travaillé dans le monde médical et j’ai l’habitude d’avoir des sévères audits de sécurités sur mes serveurs sécurisés

Moi et mes partenaires savons créer des serveurs haute disponibilités, redondants et sécurisés. J’utilise des serveurs Web AWS. AWS est le leader du marché et le meilleur pour la sécurité, en revanche ce n’est pas le plus économique.

Le serveur web serait dupliqué, stateless c’est-à-dire qu’il ne stocke rien. Ce serveur sécurisé peut être détruit en cas de problème de sécurité et reconstruit à un autre endroit en < 1 minute automatiquement. Les échanges avec le serveur sont cryptés.

La base de données ne sera pas accessible par le réseau extérieur. Les données sensibles sont cryptées et la bd est physiquement stockée à 2 endroit différents, 1 à Paris et l’autre ailleurs en France. Les 2 bases de données sont sauvegardées toutes les 6h. Le tout est accessible uniquement par une MFA (Multi-Factor Authentication) c’est-à-dire un mot de passe + une application téléphonique d’authentification avec clés tournantes de 30 secondes.

J’utilise des bases de données SQL qui sont plus sérieuses et plus fiables que les autres comme pour le projet de Rodolphe

Concernant ton 2ème point

Il s’agirait de faire une pré-étude sur la bd de l’édition 1 et ainsi de corréler les réponses de la question 1 et de ses sous-questions avec les scores d’empathie des colonnes empathy\_EC-2 à empathy\_PT-28 pour trouver les main topics.

Ça ressemble beaucoup à ce qu’il sera fait dans l’édition 2.

C’est une très bonne idée.

Il est tout à fait possible de faire une telle étude en mode Proof of Concept avec des modèles de LLM libres, sans serveur de calcul ni de base de données SQL, afin de diminuer les coûts. Ça sera moins performant mais je pense qu’on pourrait dégager un résultat avec un budget très raisonnable.

Je te propose un atelier d’1h pour affiner cela et établir un budget, par exemple la semaine prochaine.