Mesure de la qualité d'une extraction d'idée réalisée par un modèle de langue

Application aux consultations publiques avec questions ouvertes

Matthias MAZET, Garance MALNOË & Yannis PETIT
encadré.e.s par Cyril FRANÇOIS

Table des matières

1 Introduction **1**
1.1 Contexte . . . 1
1.2 Objectif des travaux . . . 2

2 Données et Méthodes **3**
2.1 Données . . . 3
2.2 Protocole d'extraction et d'évaluation . . . 4
2.3 Métriques . . . 5
2.4 Pipeline . . . 8
2.5 Impact du changement de LLM . . . 8

3 Résultats et discussion **8**
3.1 Extractions et notation humaine . . . 8
3.2 Métrique QualIT . . . 9
3.3 Métrique ROUGE . . . 11
3.4 Métrique NLI . . . 11
3.5 Pipeline . . . 11
3.6 Impact du changement de LLM . . . 11

4 Impact sociétal et environnemental **12**
4.1 Politique de la structure d'accueil : Data4Good . . . 12
4.2 Impact environnemental personnel . . . 12
4.3 Impact environnemental global . . . 12
4.4 Impact social global . . . 12

5 Conclusion **13**

Références **15**

6 Annexe **16**

Université Grenoble Alpes

Tracker de rédaction :

1. Intro
   (a) Contexte (G) (relu : M)
   (b) Obj des travauxs (G)
2. Données et méthodes
   (a) Données (G)
   (b) Protocole (M)
   (c) Métriques (Y)
   (d) Pipeline (M)
   (e) Comp. LLM (G)
3. Résultats et dicussion
   (a) Extractions et notation humaine (M)
   (b) QualIT (G)
   (c) ROUGE (Y)
   (d) NLI (Y)
   (e) Pipeline (M)
   (f) Comp. LLM (G)
4. RSE
   (a) Impact env perso (G)
   (b) Impact env global (G)
   (c) Impact social global (G)
   (d) Polotique de Data4Good (G)
5. Conclusion

Code couleur :

- A faire
- En cours
- A relire
- Fini (relu au moins une fois par qqun d'autre que le rédacteur)

Code auteur initial :

- (G) Garance
- (Y) Yannis
- (M) Matthias

1 Introduction

1.1 Contexte

Les consultations publiques représentent de puissants outils pour les élus et les organismes publics, mais également pour les citoyens. En effet, d'une part, elles permettent aux citoyens d'exprimer leurs attentes et leurs priorités et, d'autre part, elles permettent aux décideurs de mieux comprendre le point de vue des masses. Ainsi, ces derniers peuvent récolter de précieuses informations sur des problématiques spécifiques, accéder à leur réalité concrète et aux contraintes du terrain, et en conséquence réduire le décalage souvent perçu entre les politiques publiques implémentées et les besoins réels de la population.

Dans leur forme la plus simple, les consultations publiques s'appuient sur des questions fermées dont la réponse est un score (e.g. entre 1 et 10) ou un "Oui"/"Non". Ce format présente l'avantage d'être simple et très rapide à traiter et analyser. Cependant, les questions fermées présentent des limites évidentes. Prenons l'exemple de la question "Êtes-vous en accord avec la politique actuelle du gouvernement sur l'environnement ?". Avec une simple réponse binaire "Oui"/"Non", des individus aux opinions différentes pourraient donner la même réponse, alors que l'un pense que le gouvernement en fait trop sur la législation des pesticides, qu'un autre pense qu'il devrait en faire plus pour le contrôle des gaz à effet de serre, ou qu'un troisième pense qu'il devrait mieux aider lors des rénovations énergétiques. Ainsi, tous se diront en désaccord avec la politique du gouvernement en matière d'environnement, mais pour des raisons très différentes dont le détail est perdu à cause de l'utilisation d'une question fermée.

Face à ce problème surgit alors la solution des questions ouvertes. En permettant aux contributeurs de s'exprimer librement et de mieux détailler leur point de vue, elles offrent aux commanditaires un aperçu plus riche des opinions exprimées et de la réalité du terrain, et peuvent ainsi favoriser une prise de décision plus adéquate. Elles permettent également de mieux nuancer les réponses, ce qui peut contribuer à dépasser les idées préconçues ou qui bénéficient le plus de visibilité dans les médias. Enfin, lorsque les consultations publiques sont perçues comme sincères et suivies d'effets concrets, elles peuvent renforcer le sentiment que la participation individuelle a un impact réel sur les décisions publiques. Ce dernier point est d'autant plus crucial dans un contexte marqué par une abstention électorale élevée et une défiance croissante à l'égard des institutions [AB22, BCR26].

Deux initiatives de consultations publiques à grande échelle ont récemment été menées en France : le Grand Débat National et la Convention Citoyenne pour le Climat, organisées entre 2019 et 2020. Toutes deux avaient pour objectif, aux travers de questions ouvertes, de mieux capter l'opinion des citoyens français, notamment en réaction au mouvement des gilets jaunes de 2018 [BM19]. Ces deux consultations ont toutefois représenté un coût important : 12 millions d'euros et environ 2 millions de contributeurs pour la première, 5,4 millions d'euros et 150 participants pour la seconde [Cyr, Cou20]. Ces chiffres mettent en évidence une des limitations majeures de l'utilisation des questions ouvertes : elles nécessitent un investissement matériel et humain conséquent afin d'assurer la diffusion de l'information, le recueil

FIGURE 1 – Logo de la Convention Citoyenne pour le Climat

des contributions et leur analyse. La démocratie participative a donc un coût qui peut rapidement devenir élevé lorsqu'on souhaite organiser des dispositifs à grande échelle. Une autre problématique de ce type de consultations concerne leur temporalité. Pour qu'une consultation reste pertinente, le recueil et le traitement des avis doivent être réalisés dans un délai raisonnable. Si l'analyse est trop longue, il peut devenir difficile d'agir efficacement (e.g. la fin du mandat du commanditaire), les problématiques soulevées peuvent avoir évolué et des résultats tardifs peuvent fragiliser la confiance entre les participants et les institutions à l'origine de la démarche.

L'analyse manuelle de consultations publiques ouvertes fait donc face à plusieurs défis majeurs : elle est lente, coûteuse et reste sujette aux biais humains. Les outils numériques, et notamment les grands modèles de langues (*Large Language Models*, LLM), constituent une piste de réponse intéressante face à ces contraintes. En effet, ils peuvent automatiser certaines étapes de l'analyse, identifier les idées récurrentes au sein d'un grand nombre de contributions, produire des synthèses thématiques, favoriser une analyse plus transparente et reproductible, et réduire significativement le délai entre la collecte des contributions et la production du rapport final. Bien entendu, l'utilisation d'outils numériques dans le cadre des décisions publiques ne vise pas à remplacer la décision humaine ni à déléguer l'ensemble du processus décisionnel à une intelligence artificielle. Néanmoins, les processus démocratiques et la digitalisation ne sont pas incompatibles, et les associer pourraient contribuer à réduire les contraintes logistiques et administratives inhérentes à l'organisation de consultations à grande échelle. Plusieurs initiatives vont déjà dans ce sens, et nous pouvons notamment en citer deux :

- Le projet ouvert *Plurality*, qui promeut le développement de technologies collaboratives et décentralisées visant à renforcer la participation démocratique, la délibération collective et la protection des droits numériques [WTC26].
- L'outil open source *Talk To The City* d'Objectives AI , qui mobilise des méthodes de traitement automatique du langage pour analyser à grande échelle les contributions citoyennes et structurer les opinions exprimées lors de consultations publiques [Ins26].

1.2 Objectif des travaux

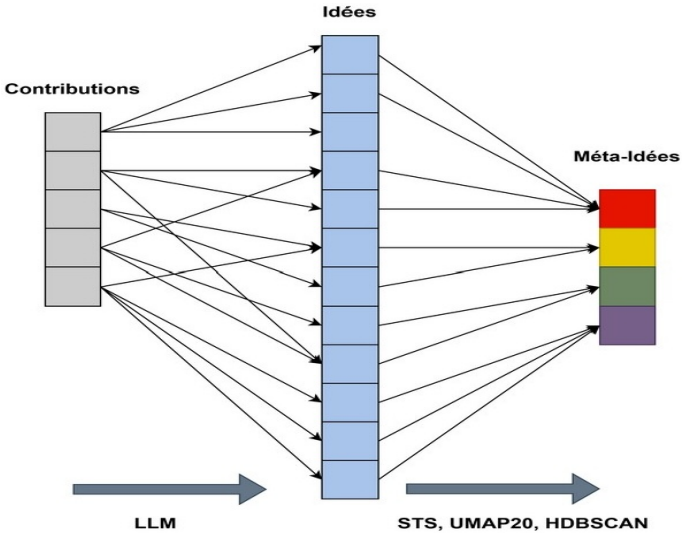

FIGURE 2 – Pipeline globale du projet Perspectiva

Perspectiva, le projet développé par Data4Good et dans lequel s'inscrit nos travaux, se trouve dans la continuité de cette démarche d'utilisation d'outils numériques afin d'analyser et comprendre ce qu'un large groupe de personnes pense. Il se décline actuellement en deux phases principales : i) l'extraction et l'analyse (type, syntaxe et sémantique), via un LLM, d'idées issues de contributions d'individus à une consultation, et ii) le regroupement de ces dites idées en méta-idées plus larges via l'utilisation d'algorithmes de clustering (STS, UMAP20, HDBSCAN, ...) (Figure 2).

En théorie, ce protocole permettrait de résumer de manière assez exhaustive les thèmes principaux exprimés dans un grand nombre de contributions. Bien que très puissants, les LLM ne sont cepen-

dant pas exempts de défauts. Nous observons parfois des *hallucinations* où la sortie du LLM contient des informations non présentent dans la contribution d'origine, voire fausses. À l'inverse, certaines idées exprimées sont parfois omises ou fragmentées en plusieurs morceaux sans aucun sens. Ce type d'erreurs, qui survient dans la première phase du projet, peut entraîner un certain nombre de problèmes par la suite. Des idées non exprimées pourraient se retrouver dans les méta-idées finales ou, à l'inverse, des idées intéressantes pourraient être perdues à cette étape. Cela conduirait alors à des interprétations erronées et à un bilan final de la consultation peu représentatif des opinions réelles de la population consultée.

Ainsi, une problématique importante pour le projet Perspectiva est de pouvoir mesurer la qualité des extractions réalisées avec le LLM dans cette première étape, l'idée est alors que les contributions avec une qualité d'extraction inférieure à un certain seuil soient exclues ou manuellement traitées. L'objectif de notre projet tutoré est donc de sélectionner ou d'élaborer une (plusieurs) métrique(s) et de la (les) valider sur des contributions issues d'une véritable consultation publique afin de construire une pipeline finale réalisant l'extraction, l'évaluation et la sélection sur les contributions.

Dans ce même contexte, nous souhaitons également étudier l'impact du choix du modèle de langage pour l'extraction : un "petit" modèle généraliste suffit-il pour réaliser une extraction correcte ou y a-t-il un véritable intérêt à usitiliser des modèles plus gros, ammenant supposément à des extractions de meilleure qualité, mais également plus gourmands ?

Ce rapport séparé en trois parties, vise à présenter notre démarche ainsi que nos résultats. La première partie présente les données utilisées, notre méthodologie pour la validation d'une métrique, les trois familles de métriques sélectionnées et notre méthodologie pour l'étude de l'impact du choix du LLM. La seconde partie présente les résultats de l'extraction réalisée par le LLM, la notation humaine de cette extraction, l'évaluation de la notation de chaque métrique considérée ainsi qu'une présentation de la pipeline finale et des résultats de l'impact du choix du LLM. Enfin, nous conclurons ce rapport par une réflexion sur l'impact environnemental et sociétal de ce projet à différentes échelles.

L'ensemble de notre travail peut-être retrouvé sur le GitHub du projet.

2 Données et Méthodes

Tous nos traitements et analyses ont été réalisés sur Visual Studio Code (v1.109.3) avec Python (v3.13.12). Nous avons utilisé les librairies numpy (v2.3.3), ollama (v0.6.0), pandas (v2.3.3), ploty (v6.3.1), requests (v2.32.5), rouge_score (v0.1.2), sentence-transformers (v5.2.2), tqdm (v4.67.1). Versions de hashlib, io, pathlib, re, tomllib et typing ? ? ?

Faire une sous-section organisation avec tableau kaban ?

2.1 Données

Afin de réaliser cette analyse, nous utilisons les contributions à la question n°163 du Grand Débat National. Présentation générale du contexte du Grand Débat National si pas fait en introduction, préciser qu'il s'agit d'une des sousquestions sur la fiscalité et les dépenses publiques. Parler du caractère accessible et ouvert des données par le gouvernement. Lien vers les données.

Statistiques descriptives du jeu de données. Aller plus loin que ce qui avait été mis dans la 1ere soutenance. Préciser que nous n'allons travailler que sur les 200 premières contributions parce que sinon manuellement c'est un enfer mais qu'elles sont plutôt représentatives du reste du jeu de données et qu'on a nettoyé le jeu de données.

Donner quelques exemples de contributions en citation.

> *Taxer davantage les très très riches....On m'a augmenté ma CSG pour la donner à Mr Bernard Arnaud ou à Mr Goshn, cet argent que l'on m'a volé, on l'a volé à mes enfants et petits enfants que je suis obligé d'aider à démarrer dans ma vie, on l'a volé à mes parents âgés car je sois les aider à payer toujours plus, un établissement de retraite décent!!*
>
> Contribution n°75

> *Surtout ne pas rétablir l'ISF,qui nous fait perdre 35 milliards (Monsieur Edouard Philippe le 09 octobre 2017)et qui fait de la France un des seuls pays au monde a posséder encore un tel impôt(IFI).Assez de démagogie*
>
> Contribution n°178

> *Qu'elle frappe en priorité les TRES RICHES : ISF, FLAT TAXE.....TAXER l'EVASION FISCALE ....Supprimer toutes les niches fiscales inefficaces et couteuses*
>
> Contribution n°163

FIGURE 3 – Exemples de contributions

2.2 Protocole d'extraction et d'évaluation

Afin de connaître la qualité des extractions obtenues et de pouvoir correctement analyser les métriques testées, nous avons noté chaque extraction à la main [Rei25]. Nous avons donc procédé comme suit :

1. Extraction des idées de chaque contribution via un LLM.
2. Calcul de la métrique sélectionnée (QualIT, ROUGE, NLI) sur chaque extraction.
3. Notation manuelle de l'extraction :
   a. Chacun des 3 membres du groupe a noté chaque extraction. Pour prévenir au maximum de biais éventuels dû à une notation humaine, nous avons fixé au préalable un ensemble de règles (Figure 4).
   b. Marquage de la présence d'hallucinations. Cette notion correspond aux contributions pour lesquelles le LLM a inventé certaines parties (Figure 5).

c. Marquage de la présence d'idées séparées. Cette notion correspond aux contributions pour lesquelles le LLM a découpé les phrases sans prendre en compte le sens (Figure 6).

4. Calcul des corrélations entre chaque annotateur.
5. Agrégation des notes humaines en un "score humain".
6. Calcul de la corrélation entre le score humain et la métrique sélectionnée.

1. La note doit être comprise entre 0 et 10. Plus la note est élevée, plus l'extraction est de qualité.
2. L'extraction doit contenir le même nombre d'idées que la contribution originelle. Aucune idée ne doit être ni ajoutée ni omise.
3. Les idées extraites doivent garder le même sens que l'idée originale (e.g. une idée négative doit le rester).
4. Les idées extraites doivent être distinctes. Il ne doit pas y avoir de redondance.
5. L'extraction doit être en français.

FIGURE 4 – Règles de notation

Nous avons ici utilisé le LLM "llama3 :8b-instruct-q4_K_M" avec une température de 0 et une top_p de 0.95. Le prompt utilisé (Annexe - Figure 10) est une adaptation du prompt initial du projet Perspectiva. Bien que nous ne nous y intéressions pas par la suite, nous avons conservé l'analyse de sentiments (type, syntaxe et sémantique) via le LLM afin que notre travail puisse s'intégrer directement dans le projet Perspectiva. il est important de noter que nous voulions ici obtenir des erreurs d'extraction, notre but premier étant d'analyser et de valider des métriques. Nous n'avons donc pas cherché à écrire un prompt parfaitement adapté ni à utiliser le meilleur LLM à notre disposition.

2.3 Métriques

L'extraction d'idées à l'aide d'un LLM permet de considérablement accélérer l'analyse des contributions de consultations publiques, en plus de pouvoir être neutre et automatisé. Toutefois, produire une extraction peut mener à certains problèmes. En effet, une extraction peut omettre certaines idées, en ajoutant ou bien en enlevant des idées de la contribution, modifier le sens initial d'une proposition ou encore mal découper les idées qui sont expriméees. Dans le cadre d'une consultation publique, destinée à des élue et où chaque voie compte, de telles erreurs peuvent avoir un impact négatif sur la fidélité du bilan final et nuire à la représentation des opinions exprimées.

Il devient donc nécessaire de disposer d'un outil permettant de quantifier objectivement les extractions issus d'un LLM en terme de qualité. On va donc utiliser des métriques pour évaluer celles-ci. Une métriqued'évaluation est une fonction qui associe un score à une production textuelle afin de mesurer une propriété donnée. Dans notre cas, nous cherchons à évaluer la qualité d'extraction d'un LLM. Il s'agit d'estimer dans quelle mesure l'extraction traduit fidèlement les idées continues dans la contribution d'origine.

Néanmoins, toutes les métriques ne mesurent pas les mêmes dimensions : Certaines reposent principalement sur le chevauchement lexical entre deux textes, tandis que d'autres cherchent à

capturer leur proximité sémantiques. Ainsi, une extraction peut obtenir un bon score score avec une métrique tout en présentant des défauts important du point de vue de la compréhension de la langue francaise et inversement.

L'objectif de cette section est donc multiple : Nous allons présenter les trois métriques testées dans ce projet et analyser leur capacité à corréler avec l'évaluation humaine.

Métrique QualIT

La première métrique que nous avons étudiée est issue du travail présenté dans le papier *Qualitative Insights Tool (QualIT) : LLM Enhanced Topic Modeling* [KGB+24]. Bien que QualIT soit initialement conçu comme un outil de topic modeling intégrant des modèles de langage, il introduit un mécanisme permettant de vérifier la cohérence des éléments extraits par un LLM.

Dans l'approche proposée par les auteurs, le LLM extrait d'abord des phrases clés à partir de chaque document. Afin de limiter les erreurs d'extraction, notamment les hallucinations, un score de cohérence est ensuite calculé pour chaque phrase extraite. Ce score mesure l'alignement sémantique entre le texte original et l'élément extrait.

$$C_i = \frac{1}{n}\sum_{j=1}^{n} \frac{V_{inp,ij} \cdot V_{key,ij}}{|V_{inp,ij}| \cdot |V_{key,ij}|}, \text{ avec } \begin{cases} C_i & \text{le score QualIT de la } i^e \text{ contribution.} \\ n & \text{la dimension de l'espace d'embedding.} \\ V_{inp,ij} & \text{la } j^e \text{ coordonnée du vecteur d'embedding de la } i^e \text{ contribution.} \\ V_{key,ij} & \text{la } j^e \text{ coordonnée du vecteur d'embedding des idées extraites de la } i^e \text{ contribution.} \\ |\cdot| & \text{une norme euclidienne.} \end{cases}$$

Métrique ROUGE

La seconde métrique que nous avons testée est ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*). ROUGE est un ensemble de métriques initialement conçu pour évaluer automatiquement la qualité de résumés, en comparant un texte généré à un texte de référence. Dans notre contexte, nous l'utilisons pour comparer l'extraction produite par le LLM (les idées extraites) au texte original de la contribution. L'idée est la suivante : si l'extraction est fidèle à la contribution, alors une partie du contenu lexical (mots, expressions, séquences) devrait se retrouver dans les deux textes.

ROUGE propose plusieurs variantes, correspondant à différentes manières de comparer deux textes :

- **ROUGE-N** : mesure le chevauchement de n-grammes (séquences de n mots) entre la référence et le texte généré. Dans notre cas, nous utilisons $n = 1$ (ROUGE-1), ce qui revient à comparer le recouvrement en unigrammes.
- **ROUGE-L** : se base sur la *plus longue sous-séquence commune* (*Longest Common Subsequence*, LCS), ce qui permet de capturer une similarité en séquence sans imposer que les mots soient strictement consécutifs.
- **ROUGE-S / ROUGE-SU** : repose sur des *skip-bigrammes*, c'est-à-dire des paires de mots apparaissant dans le même ordre, possiblement séparés par d'autres mots (avec ou sans uni-

grammes).

Les scores ROUGE sont compris entre 0 et 1. Un score proche de 1 indique une forte similarité entre les deux textes, tandis qu'un score proche de 0 suggère une faible similarité. Dans ce projet, notre objectif n'est pas d'évaluer une similarité *mot à mot* parfaite, car une bonne extraction peut reformuler le texte original. Néanmoins, ROUGE constitue une baseline intéressante car elle est simple, rapide à calculer et largement utilisée dans la littérature.

Nous avons calculé deux variantes : ROUGE-1 et ROUGE-L. L'implémentation repose sur la librairie `rouge_score` en Python . Pour chaque contribution, nous comparons le texte original (`contribution`) au texte d'extraction (`ideas_text`). Nous conservons la mesure F_1 (appelée *F-measure* dans la librairie) [Lin04], qui combine précision et rappel et permet de limiter certains effets liés aux différences de longueur entre les textes. En cas d'erreur de parsing (extraction vide, format non conforme), nous assignons un score nul afin de ne pas biaiser les résultats.

Métrique NLI

[MLJ23] (placer la citation) La troisième métrique que nous avons testé est la métrique NLI (Natural Language Inference). Elle est s'inspirée du cadre *Natural Language Inference* (NLI) dont le principe est de déterminer, à partir d'une prémisse p (c'est quoi ?) et d'une hypothèse h (c'est quoi ?), si l'hypothèse est entraînée par la prémisse, contradictoire ou neutre. Dans notre contexte, nous cherchons à vérifier si les idées extraites par le LLM sont bien supportées par la contribution originale, et à pénaliser les cas où le LLM inverse le sens (par exemple via une négation) ou introduit des éléments qui ne sont pas présents.

Dans l'idéal, cette étape pourrait être réalisée à l'aide d'un classifieur NLI pré-entraîné (par exemple via `transformers`). Néanmoins, une contrainte (quelle contrainte ?) nous ont conduits à une implémentation plus légère. D'une part, nous souhaitons conserver une métrique indépendante du LLM d'extraction, facilement réutilisable. D'autre part, notre environnement de développement ne permettait pas d'installer et d'exécuter correctement certaines bibliothèques de deep learning (`torch`, `transformers`) (c'est bizarre ça non ?) avec Python 3.13 (à voir avec les autres). Nous avons donc implémenté une approximation lexicale de type NLI, inspirée de l'idée suivante : une idée extraite est d'autant plus fiable qu'elle présente une forte similarité avec au moins une phrase de la contribution, et qu'elle ne contredit pas cette phrase.

Calcul du score de support. Nous découpons d'abord la contribution en phrases $\{p_j\}$ (ponctuation forte et retours à la ligne), et nous considérons les idées extraites $\{h_i\}$ obtenues par le LLM. Pour chaque idée h_i, nous calculons un score de similarité avec chacune des phrases p_j et nous conservons le meilleur appariement :

$$\text{support}(h_i) = \max_j \left(0.6 \cdot J(h_i, p_j) + 0.4 \cdot LCS(h_i, p_j)\right),$$

où $J(h_i, p_j)$ est un score de Jaccard calculé sur les tokens contenu (après retrait d'une mini liste de stopwords (détailler le score de Jaccard)), et $LCS(h_i, p_j)$ est une similarité de séquence calculée via un ratio de type longest common subsequence (détaillé, mettre en fr ?)(implémenté ici à l'aide de `SequenceMatcher`). Ce score est compris entre 0 et 1 : plus il est élevé, plus l'idée h_i est lexicalement proche d'au moins une phrase de la contribution.

Pénalisation des contradictions par négation. Une limite des métriques purement lexicales est qu'elles peuvent donner un score élevé à une phrase qui réutilise les mêmes mots, mais en inversant le sens (en ajoutant une négation par exemple). Afin de prendre en compte ce cas, nous détectons la présence d'une négation dans h_i et dans la phrase p_{j^*} (p_{j^*} pas défini précédemment) qui maximise le support. Si la négation est différente entre les deux, nous ajoutons une pénalité :

$$\text{contra}(h_i) = \mathbf{1}\left[\text{neg}(h_i) \neq \text{neg}(p_{j^*})\right] \cdot (1 - \text{support}(h_i)) \cdot \alpha,$$

où $\alpha \in]0, 1[$ est un facteur de pénalisation (fixé à $\alpha = 0.8$ dans nos tests (pourquoi ? besoin justification ou alors préciser arbitrairement.). L'idée est la suivante : si une idée est peu supportée par le texte et qu'elle inverse la polarité via une négation, il est probable qu'elle corresponde à une hallucination ou à une mauvaise reformulation.

Score final. Pour chaque contribution, nous agrégeons ensuite les scores sur l'ensemble des idées extraites. Nous calculons :

- NLI_{support} : la moyenne des $\text{support}(h_i)$;
- NLI_{contra} : la moyenne des $\text{contra}(h_i)$;
- $NLI_{\text{final}} = \max\left(0, \min\left(1, NLI_{\text{support}} - NLI_{\text{contra}}\right)\right)$.

Ce score final est compris entre 0 et 1. Un score élevé correspond à des idées globalement bien supportées par la contribution et ne présentant pas d'inversion évidente du sens via la négation.

2.4 Pipeline

Afin de permettre une réutilisation complète et rapide de nos travaux, nous avons construit un pipeline de traitement des contributions. À l'aide d'un LLM et d'un prompt donnés, il permet d'extraire les idées des contributions, de réaliser l'analyse de sentiments, puis de retourner un jeu de données concaténé par contribution. Ce pipeline permet aussi, à l'aide des différentes métriques, de filtrer le jeu de données final selon le rendu souhaité (résultats bruts ou résultats sans "mauvaise" extraction).

2.5 Impact du changement de LLM

A voir.

3 Résultats et discussion

3.1 Extractions et notation humaine

Comme attendu (et souhaité), un certain nombre d'extractions étaient mauvaises. Nous avons notamment identifié deux manières redondantes dont le LLM se trompait :

- Le LLM hallucine, c'est-à-dire qu'il invente des idées non présentes dans la contribution originale, mais dont le sens découle généralement d'autres idées bien présentes (Figure 5).
- Le LLM découpe les contributions originales de manière incohérente. Il lui arrive de couper une phrase en deux soit sans raison, soit à cause de la présence de pronom implicite (Figure 6).

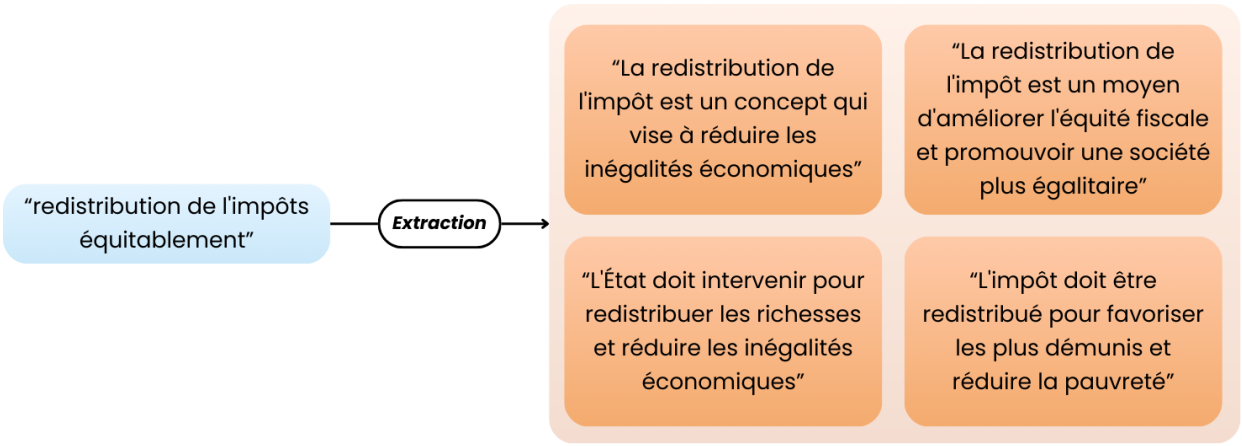

FIGURE 5 – Exemple d'hallucination du LLM

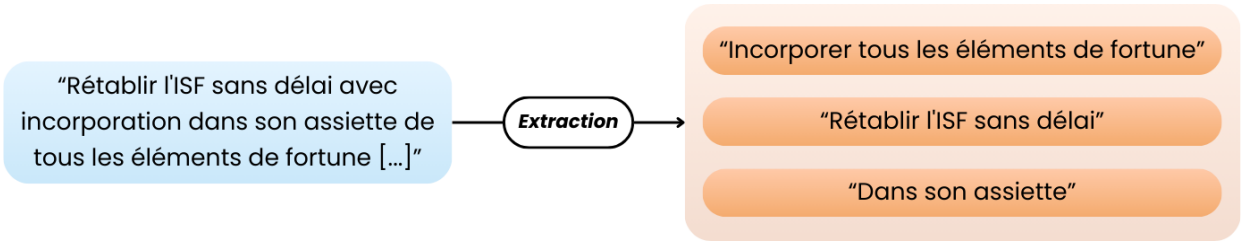

FIGURE 6 – Exemple d'idées séparées par le LLM

Ces deux cas de figure permettent de renforcer l'utilité d'une première notation à la main, afin de justement pouvoir les identifier.
Bien que nous ayions fixé des règles avant de réaliser la notation des contributions, il est probable que des biais personnels aient persisté, d'où l'intérêt d'avoir chacun noté les 200 extractions sélectionnées. Par un calcul de corrélation, nous avons donc vérifié notre accord, et nous étions généralement tous en accord (corrélation > 0.9 à chaque fois) (Table 1). Nous avons donc agrégé les trois notes via une moyenne standard.

	Garance	**Matthias**
Garance		0.93
Yannis	0.93	0.94

TABLE 1 – Corrélation des notes entre annotateurs du groupe

Une fois agrégé et normalisé dans $[0\,;1]$, nous obtenons un score moyen d'environ 0.54, avec une médiane à 0.6, un minimum à 0 et un maximum à 1 (Annexe - Figure 11).

3.2 Métrique QualIT

3.2.1 Résultats généraux

Correlation entre la note moyenne des annotateurs et le score QualIT avec figures. Description de ce que l'on observe.

3.2.2 Impact de la longueur du texte

3.2.3 Impact des hallucinations

3.2.4 Explications

Idées séparées :

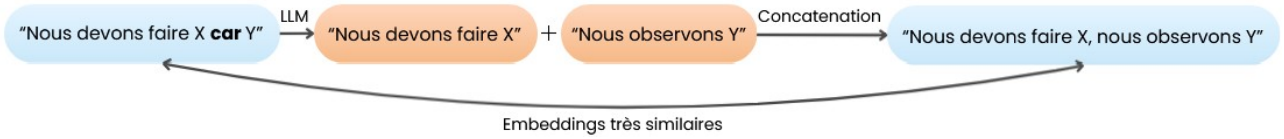

FIGURE 7 – Schéma d'un exemple d'échec de détection des idées séparées par QualIT à cause de la présence d'une conjonction de coordination

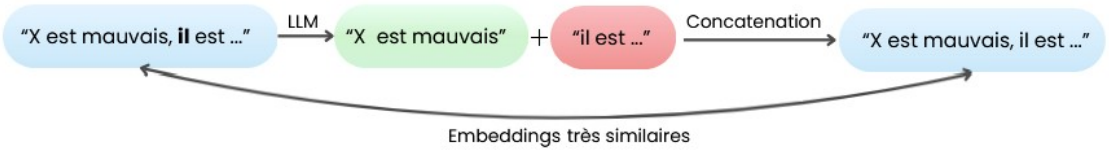

FIGURE 8 – Schéma d'un exemple d'échec de détection des idées séparées par QualIT à cause de la présence d'un pronom

Hallucinations :

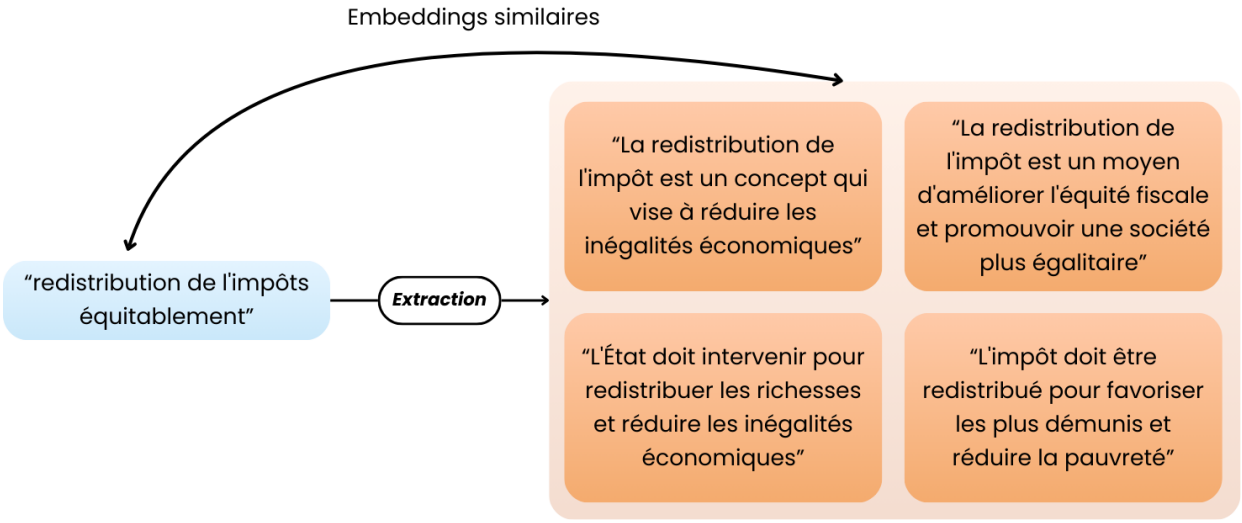

FIGURE 9 – Schéma d'un exemple d'échec de détection des hallucinations par QualIT

3.3 Métrique ROUGE

3.3.1 Résultats L-ROUGE

3.3.2 Résultats 1-gram ROUGE

Faire comparaison avec les résultats de QualIT.

3.4 Métrique NLI

Faire comparaison avec les résultats de QualIT et ROUGE.

3.5 Pipeline

Nous avons construit un pipeline avec pour objectif d'avoir un maximum de flexibilité sur sa réutilisation. Pour cela, il prend en entrée de nombreux arguments :

- `df` : Le jeu de données à analyser.
- `system_prompt` : Le prompt système pour le LLM.
- `user_template` : Le template utilisateur pour le LLM.
- `extract_model` : Le LLM pour l'extraction (avec par défaut celui que nous avions utilisé, "llama3 :8b-instruct-q4_K_M").
- `embed_model` : le modèle de sentence-transformers pour les embeddings (avec par défaut "sentence-transformers/all-MiniLM-L6-v2").
- `return_intermediate` : un argument binaire pour préciser si nous souhaitons aussi le jeu de données avec l'analyse de sentiments (défaut = False).
- `error_filter` : Un argument binaire pour filtrer ou non les erreurs de parsing (défaut = "Oui").
- `qualit_filter` : Le score QualIT minimal pour qu'une extraction soit conservée (défaut = 0 : pas de filtre).
- `rouge1_filter` : Le score ROUGE 1-gramme minimal pour qu'une extraction soit conservée (défaute = 0 : pas de filtre).
- `rougeL_filter` : Le score ROUGE L minimal pour qu'une extraction soit conservée (défaute = 0 : pas de filtre).
- `nli_filter` : Le score NLI minimal pour qu'une extraction soit conservée (défaut = 0 : pas de filtre).

Ainsi, le pipeline peut être utilisé de diverses manières :

- Tester différents LLMs et/ou prompts pour extraire les idées des contributions.
- Construire des jeux de données avec différents niveaux de qualité. Par exemple, en approfondissant au maximum l'analyse des métriques, il est possible d'obtenir un jeu de données ne contenant que de très bonnes extractions grâce aux bons seuils de métriques.
- Obtenir une première idée de la qualité de l'analyse de sentiments des contributions.

3.6 Impact du changement de LLM

A venir.

4 Impact sociétal et environnemental

Doit faire environ 1/5 du rapport.

4.1 Politique de la structure d'accueil : Data4Good

- Décrire le fonctionnement de Data4Good.
- Parler de l'aspect bénévole, vertueux autant sur l'impact social et environnemental.
- Pas de locaux, donc pas de dépenses énergétiques liées à ça, par contre du coup besoin d'outil en ligne pour la coordination. Est-ce que ça consomme plus qu'une entreprise normale ? Non parce qu'ils utiliseraient probablement un slack aussi s'ils étaient en présentiel dans des bureaux.
- Dimension "d'insertion", gain d'expérience sur des projets concrets pour des jeunes Data Scientists ?
- Parler du fait que vu que c'est associatif ?, il n'y a pas vraiment de politique à mener autre que celle des projets qui sont créés (pas de réduction des déchets, pas d'infrastructure).
- Parler peut-être d'autres projets passés réalisés par Data4Good comme exemple. Aspect concret de leur travail et engagement.

4.2 Impact environnemental personnel

Calcul de notre impact personnel pour le projet.

- Visios donc pas de déplacements. Voir si finalement pas plus de consommation que si on se déplaçait à la fac ?
- consommation des équipements personnels et à distance (GCP)
- consommation des équipements à distance (GitHub, Google Drive, GCP)
- Utilisation de l'IA pour le code et pour la rédaction.

4.3 Impact environnemental global

Parler du court, moyen et long terme.

- Utilisation d'un LLM -> coûteux en ressources pour l'entrainement du LLM et pour son utilisation. C'est le + gros postes de dépenses. Après faut le faire tourner qu'une fois, c'est peut-être pas si pire que ça. Intérêt de choisir un plus ou moins gros LLMs, est-ce que les petits sont suffisamment bons ou est-ce qu'on est tout de suite obligé de partir sur des gros modèles pour avoir des résultats qualitatif ? Est-ce que la différence est vraiment importante ou finalement relativement négligeable ?
- Nécessite ensuite quelques calculs sur un PC mais relativement léger pour l'utilisation de la pipeline.
- Projet qui incorpore des outils couteux, banalisation de l'utilisation de ces outils et donc participe à ce qu'ils soient potentiellement de + en plus utilisés. Problématique sur le long terme.

4.4 Impact social global

- Projet lié à la politique : court terme outil pour venir accompagner la prise de décision et "donner une idée" de ce que pense les gens mais long terme outil qui s'inscrit dans une démarche plus générale de démocratie participative, peut-être un futur où on a plus de dirigean, maiss

tout le monde donne son avis et décisions via IA ? (insisté sur le côté "idée", pas réaliste). C'est un outil qui ouvre la voie, un peu boite de Pandore potentiellement (positif ou négatif cela dit).
- Outil plutôt vertueux qui veut permettre une meilleure démocratie, permettre aux citoyens d'être mieux entendus/compris par les gens qui prennent les décisions publiques. Permet peut-être aussi de relancer le débatpublice, la parole qui est un peu lésée ces derniers temps (faire le lien avec le vote blanc et l'abstention grandissante) à condition que derrière les politiques écoutent ce que disent les gens, risque de frustration si pas écoutés ? Aussi, risque peut-être de ne pas entendre tout ce qui est dit dans les contributions si traitement des données qui ne fait ressortir que les idées les + populaires, risque peut-être de louper des choses.
- Problématique du choix du LLM et de la base de données sur laquelle il a été entrainé. Si la base d'entrainement est remplie d'un certain discours (raciste, sexiste, autre...) potentiellement le LLM pourrait plus facilement repérer et extraires ces idées. Potentielles déformations (plutôt qu'hallucinations) qui ne seraient pas détectées par les métriques et qui iraient dans le sens de ces discours.
- Niveau RGPD, anonymisé par le jeu de données déjà dispo.
- Question d'envoyer les avis des gens à un LLMs dans certains contextes notamment politique ? Utilité d'avoir un LLM en local plutôt que d'envoyer nos données à OpenAI notamment dans ces temps de crise politique et d'enjeux de connaissances / de données.
- il faut quand même des gens pour faire tourner les algorithmes même si pas d'analyse à la main. Et donc il faut que ce soit pas trop complexe sur le plan code/informatique/utilisation, un peu "clé en main" pour que des personnes pas forcément très informaticiennes puissent se servir de ces outils (employé de mairie, manager dans une entreprise, ...)
- Problématique de la fracture numérique potentielle : si les avis ne sont receuillis que par internet, les gens qui n'ont pas internet ou ne savent pas se servir d'un ordi ne peuvent pas participer -> exclusion du débat publique problématique notamment parce que concernera seulement certaines tranches de la pop (à vérifier, mettre lien vers étude sur le manque d'accès à internet selon ruralité, CSP et âge etc), besoin de faire attention à permettre le recueillement des avis de l'ensemble des citoyens.

5 Conclusion

Remettre bout d'introduction sur l'objectif du projet.

Résultats globaux sur les métriques, la pipeline et l'impact du choix du LLM. Qu'est-ce qui peut encore être amélioré ou mieux détecté, quelles sont les limites actuelles ? Faire lien peut-être rapidement avec enjeux environnementaux et sociaux.

Ouverture :
- Parler des problématiques des consultations publiques concernant leur organisation : nécessité de la représentativité des personnes participant à la consultation (reproche fait au GDN) [Bli19] [BGM20], nécessité de la transparence du processus (pour la démocratie, la confiance c'est bien. Lien avec le fait que les gens parlent de parcoursup comme une "boite noire", d'"algorithme") et derrière nécessité que ces consultations publiques soient réellement "utiles".
- En présence de sous-représentation d'une certaine frange de la population parmi les contributeurs, porter une attention toute particulière à ce que leurs idées soient bien conservées à l'étape des meta-idées.

- Choisir les seuils de la pipeline avec le "client" (est-ce qu'il veut quelque chose de nickel ou est-ce qu'il s'en fiche s'il reste des idées mal extraites et quelques hallucinations).
- Choisir les seuils en fonction de leur impact sur les meta-idées (deuxième partie du projet Perpectiva).

Références

[AB22] Elisabeth Algava and Kilian Bloch. Vingt ans de participation électorale : en 2022, les écarts selon l'âge et le diplôme continuent de se creuser. https ://www.insee.fr/fr/statistiques/6658143, 2022.

[BCR26] Damien Bol, Bruno Cautrès, and Luc Rouban. La défiance envers les politiques en france approche un point de non-retour, selon le baromètre cevipof. *Le Monde*, 2026.

[BGM20] Hamza Bennani, Pauline Gandré, and Benjamin Monnery. Les déterminants locaux de la participation numérique au grand débat national : une analyse économétrique. *Revue Economique*, 71 :715–737, 2020.

[Bli19] Simon Blin. Un public éloigné des traits sociologiques des gilets jaunes. *Libération*, 2019.

[BM19] Eric Buge and Camille Morio. Le grand débat national, apports et limites pour la participation citoyenne. *Revue du droit publique*, pages 1205–1238, 2019.

[Cou20] Dimitri Courant. La convention citoyenne pour le clmat. une représentation délibérative. *Revue Projet*, 378 :60–64, 2020.

[Cyr] François Cyril. Perspectiva - utiliser les outils llm pour analyser et forunir un retour plus rapode et fidèle aux contributeurs d'une consultation publique - enjeux sur la communication des résultats. `https://docs.google.com/presentation/d/1z22FzVRVmvoVikUffME6vtDrgkIA5p9acvcq-KzoHZ0/edit?slide=id.g35890f26d4e_0_48#slide=id.g35890f26d4e_0_48`.

[Ins26] AI Objectives Institute. Talk to the city. https ://ai.objectives.institute/talk-to-the-city, 2026.

[KGB+24] Satya Kapoor, Alex Gil, Sreyoshi Bhaduri, Anshul Mittal, and Rutu Mulkar. Qualitative insights tool (qualit) : Llm enhanced topic modeling. /, 2024.

[Lin04] Chin-Yew Lin. Rouge : A package for automatic evaluation of summaries. *In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, 2004.

[MLJ23] Potsawee Manakul, Adian Liusie, and Mark J.F.Gales. Selfcheckgpt : Zero-resource black-box hallucination detection for generative large language model. 2023.

[Rei25] Ehud Reiter. How to validate metric. `https://ehudreiter.com/2018/07/10/how-to-validate-metrics/`, 2025.

[WTC26] E.Glen Weyl, Audrey Tang, and Plurality Community. Plurality : The future of collaborative technology and democracy. https ://www.plurality.net/v/eng/, 2026.

6 Annexe

```
But: extraire les idées principales DISTINCTES d'un texte pour analyse.

Règles:
  1. N'utiliser QUE le contenu entre «<TEXT »>.
  2. Extraire la liste des idées DISTINCTES et PRINCIPALES.
      - Chaque idée = une phrase claire, autonome, reformulée si
        nécessaire.
  3. Pour CHAQUE idée, annoter:
      - type: "statement" (constat) OU "proposition"
        (suggestion/recommandation/objectif).
      - syntax: "negative" si la phrase contient une négation explicite
        (ex.: "ne", "n'", "ne pas", "ne plus", "non"), sinon "positive".
      - semantic: "positive", "negative" ou "neutral" (valence
        sémantique).
  4. Sortie STRICTEMENT en CSV avec entête EXACTE:
      - Délimiteur: virgule.
      - Chaque description entre guillemets doubles.
      - Chaque description entre guillemets doubles.
      - Échapper tout guillemet interne par duplication (ex.: ""chat"").
      - NE RIEN AJOUTER d'autre (pas de texte avant/après, pas de code
        fences).
      - Pas de lignes vides.

Exemple:
CSV:description,type,syntax,semantic
"Les chats retombent sur leurs pattes",statement,positive,neutral
"Les chats n'ont pas neuf vies",statement,negative,negative
"Il faut mieux prendre soin des chats pour prolonger leur
vie",proposition,positive,positive
```

FIGURE 10 – Prompt utilisé pour l'extraction d'idées

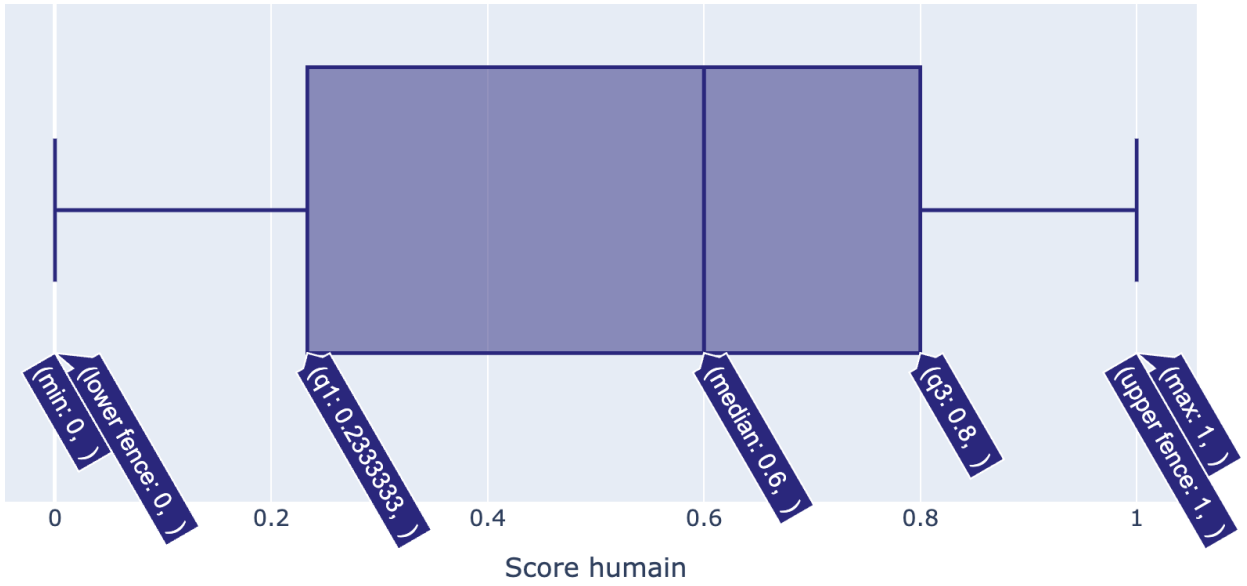

FIGURE 11 – Distribution du score humain