Perspectiva - How to assess the quality of the extraction of ideas performed by an LLM?

Application to public consultations with open questions

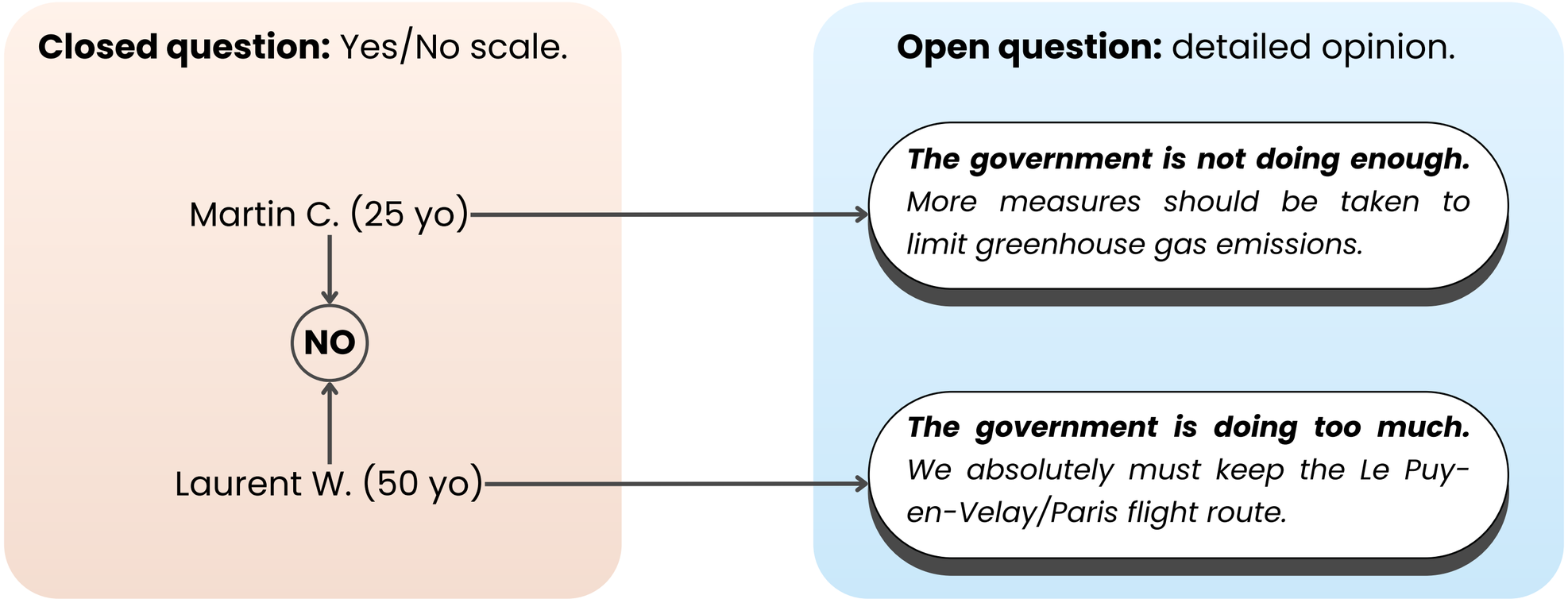
Matthias Mazet, Garance Malnoë & Yannis Petit
with Cyril François

Université Grenoble Alpes

I. Contextualisation

- **Public consultations** = good way to know what citizens think about a particular issue.
  - Particularly useful for elected officials and public services.

- Example of question: ***"Do you agree with the government's current policy on environment?"***



I. Contextualisation

- Examples of public consultations in France: [1]

Fig. 1 - CCPLC logo

- 150 contributors.
- €5 million.

Fig. 2 - GDN logo

- 2 million contributors.
- €12 million.

- **Challenges:** “reasonable” cost and fast processing time.

- **Problems with manual processing:** slow, costly, potentially biased.
  - **Solution** = use of **digital tools**.

⚠ **Not** replacing democracy with algorithms (e.g. Talk To The City, Audrey Tang in Taiwan, etc.).
  - → Having new **complementary** tools to process information.

- **Our project:** suggest and implement improvements for the already existing tools addressing this issue.

1. *Dimitri Courant, « La Convention citoyenne pour le climat. Une représentation délibérative »*

I. Contextualisation

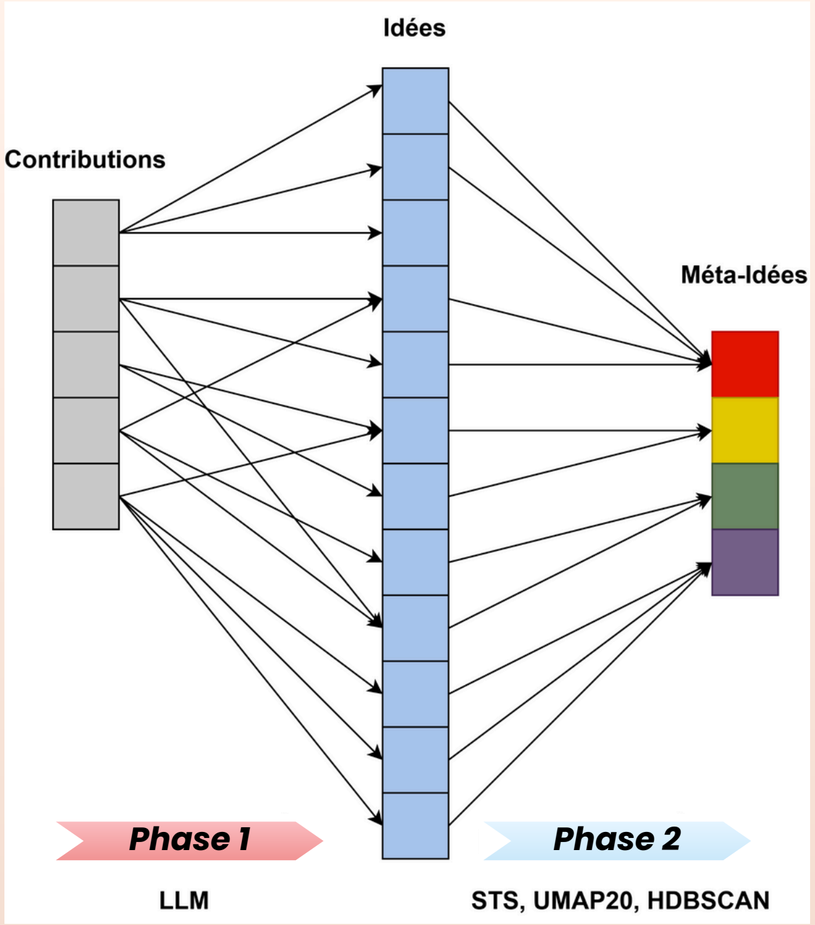


Fig. 3 - Treatment diagram [2]

- **Perpectiva project:**
  - Extraction of ideas from contributions with **Large Language Models** (LLMs).
  - Grouping ideas into meta-ideas using **clustering algorithms.**

- **Issues with LLMs:** hallucinations, loss of ideas and/or sense, “broken” ideas.
  - → **First step of the project: phase 1.**
    - finding and validating a metric to measure the quality of the extractions obtained.

2. *Cyril François - Perspectiva presentation*

II. Data and Methods

- **Objective:** obtain a metric that correlates closely with the score a human might give regarding the quality of LLM's extraction.

- **Data:** related to Question n°163 of the "Grand débat national":
  - ***"Que faudrait-il faire pour rendre la fiscalité plus juste et plus efficace ?"***
    - *(What should be done to make taxation fairer and more efficient?)*

  - 154,000 out of the 186,711 people who participated answered this question.
  - **Contributions' statistics:**
    - Between 1 character and 80,412 characters.
    - Mean of 358 characters and median of 193 characters.
  - **Examples** of contributions:

Remettre L'ISF, taxer les paradis fiscaux

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 !!

II. Data and Methods

- **Protocol :** [3]
  1. **Clean the dataset** and limit it to the first 200 contributions.
  2. **Extract ideas** for each contribution **using an LLM** and a specifically designed prompt.
  3. **Compute the chosen metric** on the extractions.
  4. The three members of the group **manually assign a grade** (0 to 10) to the extraction and **flag the presence of hallucinations and broken ideas** based on common rules.
  5. **Aggregate** human grades into a **human score**.
  6. Compute the correlations: between **annotators**, between the **human score** and the **chosen metric**.

- **Chosen metric:** Qualitative Insight Tool - **QualIT** [4]

$$C_i = \frac{1}{n} \sum_{j=1}^{n} \frac{\left(V_{\text{input},ij} \cdot V_{\text{keyphrases},ij}\right)}{\left|V_{\text{input},ij}\right| \cdot \left|V_{\text{keyphrases},ij}\right|}$$

with

C_i the coherence score of the i[th] contribution.
n the dimension of embedding space.
$V_{input,ij}$ the j[th] coordinates of embedding vector of contribution i.
$V_{keyphrases,ij}$ the j[th] coordinates of embedding vector of the extracted ideas from contribution i.
$|v|$ Euclidian norm.
$(v.u)$ Dot product.

3. *How to Validate Metrics - Ehud Reiter*
4. *Kapoor S. et al. "Qualitative Insights Tool (QualIT): LLM Enhanced Topic Modeling" (2024), Amazon*

- **Prompt's features** (same version as the one used in the original project):
  - Use only the contribution's content.
  - **Extract the list of the principal and distinct ideas.**
  - Annotate for the *type*, the *syntax* and the *semantic*.
  - Output in **CSV format.**
  - Three **examples** of extraction.

- **Rules** for the **extractions' human grading:**
  - **Grade** = integer between **0 and 10**.
  - **Same number of ideas** in the extraction and the contribution.
    - No idea should be neither omitted nor added.
  - Ideas transcribed in the **same sense**.
  - Only **distinct ideas** (no redundancy).
  - Output in **French**.

- **Code:**
  - **Python** environment (~3.12.4).
  - LLM call manager : **Ollama** for the LLM call.
  - LLM: **Meta-Llama-3.1-8B**-Instruct-AWQ-INT4.



- **Project management:**
  - **GitHub** for code and *small* data files.
  - **Google Drive** for large data files, reports and shared notes.


III. Results

	Garance	Matthias
Garance		0.93
Yannis	0.93	0.94

Fig. 5 - Correlation between group members' grading

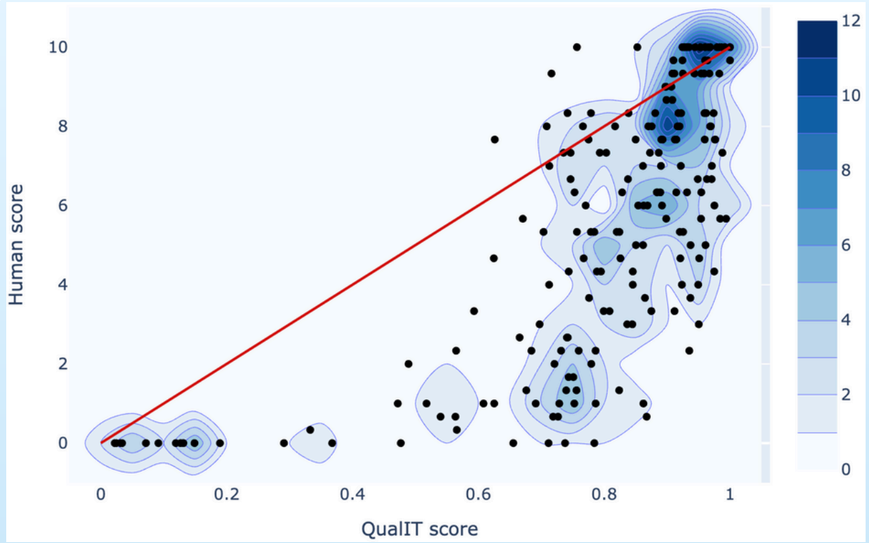


Fig. 6 - Human score vs QualIT score

- Pearson's correlation between annotators: all in **agreement**.
- Pearson's correlation between the **human score** and the **QualIT score**: 0.71.
  - Rather in agreement, but **some disagreement**.

- Most of the disagreement = contributions with a QualIT score between 0.4 and 0.8.
- Numerous **good extractions** (intense blue around (1, 10) coordinates).
- QualIT metric fails to grade extractions with a low human score → why ?

III. Results

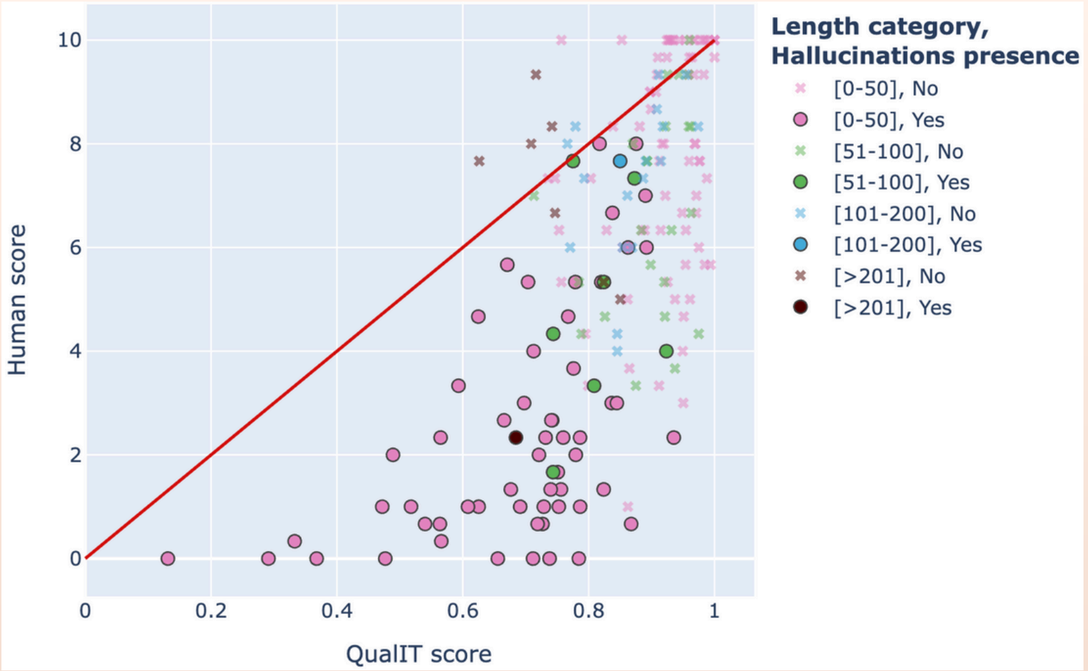


Fig. 7 - Human score vs QualIT score, with hallucinations

⚠ ***Less points than before***: *failed extractions not on the graph.*

- **Hallucinations** occurrences:
  - Mostly on **short contributions** and/or contributions with a **low human score**.
- Same graph with **broken ideas**: less relevant.

III. Results

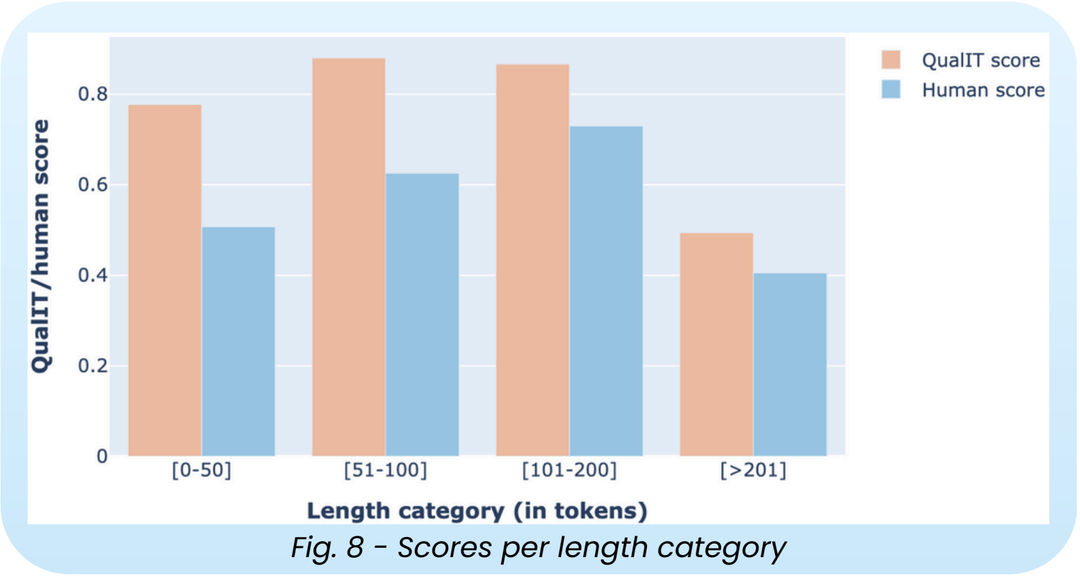


Fig. 8 - Scores per length category

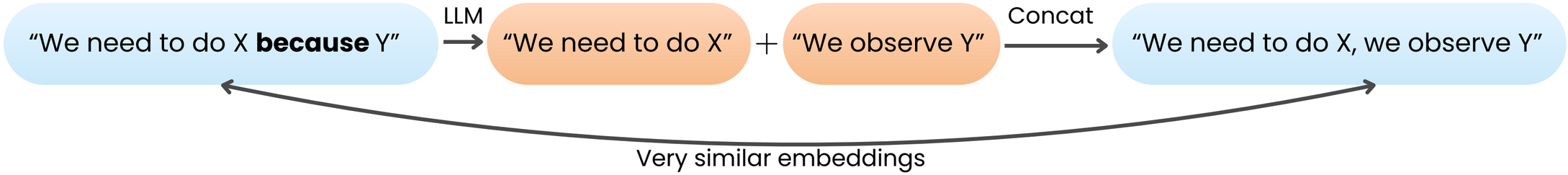
- **QualIT:** overestimate the extractions' quality.
  - Even more on short contributions.
- **Extraction:** more efficient on mid-length contributions.

III. Results

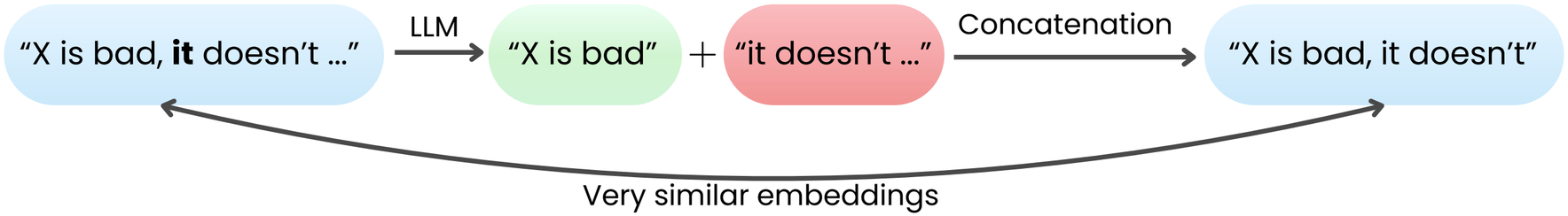
- What is happening with separated ideas and hallucinations ?

- **Separated ideas :**

1. Presence of **conjunctions** : “so”, “therefore”, “but”, “because”

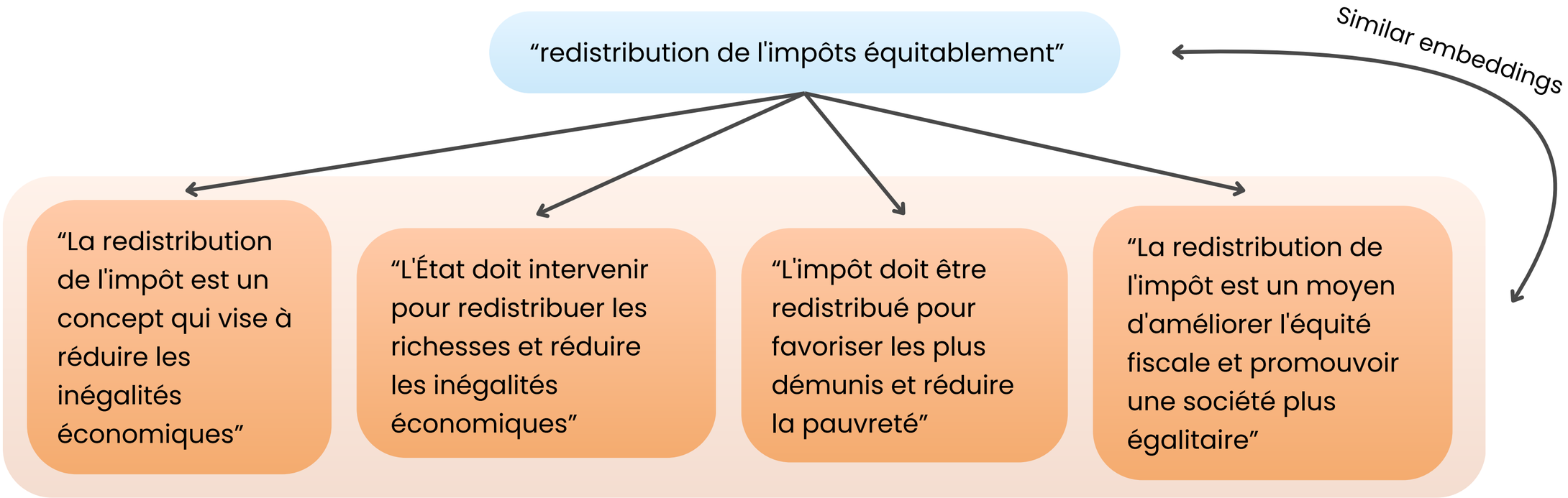


2. **Use of a pronoun** in a new idea



III. Results

- **Hallucinations :** adding **similar ideas.**
  - Example with contribution n°7 :



Metric score : 0.73

Human score : 0.06

III. Results

- QualIT metric → **ROUGE metric.**
  - **Reduce** hallucinations and parsing failures.
  - **L ROUGE** metric: focuses on longest matching sequences.
  - **1-gram ROUGE** metric: measures how much of the original text is preserved.

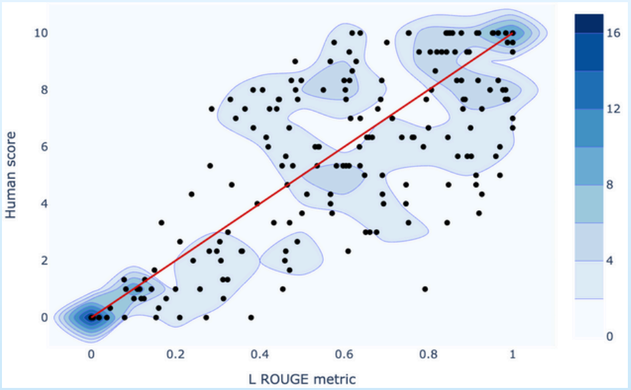


Fig. 9 - Human score vs L ROUGE score

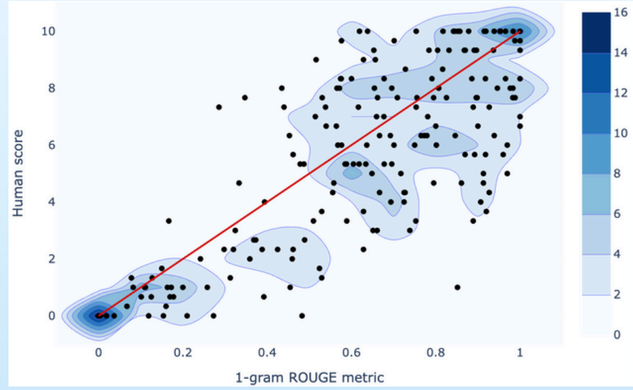


Fig. 10 - Human score vs 1-gram ROUGE score

Correlation with human score:

- 1-gram ROUGE: 0.79.
- L ROUGE: 0.78.

- Both Rouge-based metrics match human scores.
- **Good detection** of correct extractions **and** errors.

- → Ideas to **adjust** the QualIT metric.

IV. Next steps

Short term

Long term

1. **Adjustment** of the QualIT metric:
   - Failed extraction at 0.
   - Penalizing hallucinations and invalid ideas.
   - Penalizing length difference (in tokens) between the contribution and its extraction.

2. Prompt **improvements** (to prevent failed extraction).

3. **Creation** of a processing pipeline to keep only **good extractions:**
   a. **Extraction** via LLM.
   b. Filter for **failed extractions** (removing them).
   c. Filter for **hallucinations** and **invalid ideas** (rules to establish).
   d. Filter for the **adjusted QualIT metric** (rules to establish).

4. **Creation** of meta-ideas map and Personas (relations between meta-ideas).

References

1. Dimitri Courant, « La Convention citoyenne pour le climat. Une représentation délibérative ». In Revue Projet 378 (2020), p. 60-64 doi : 10.3917/pro.378.0060.
2. Cyril François (Data4Good), Perspectiva presentation.
3. Ehud Reiter, « How to Validate Metric » at https://ehudreiter.com/2018/07/10/how-to-validate-metrics/.
4. Kapoor S. et al. at Amazon “Qualitative Insights Tool (QualIT): LLM Enhanced Topic Modeling” (2024) doi : 10.48550/arXiv.2409.15626.

Back-up - Prompt I

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:
   CSV:description,type,syntax,semantic
   - Délimiteur: virgule.
   - 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.

Back-up - Prompt II

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