The GDN-CC Dataset: Automatic Corpus Clarification for AI-enhanced Democratic Citizen Consultations

Anonymous ACL submission

Abstract

LLMs are ubiquitous in modern NLP, and while their applicability extends to texts produced for democratic activities such as online deliberations or large-scale citizen consultations, ethical questions have been raised for their usage as analysis tools. We continue this line of research with two main goals: (a) to develop resources that can help standardize citizen contributions in public forums at the **pragmatic level**, and make them easier to use in topic modeling and political analysis; (b) to study how well this standardization can reliably be performed by small, open-weights LLMs, *i.e.* models that can be run locally and transparently with limited resources. Accordingly, we introduce **Corpus Clarification** as a preprocessing framework for large-scale consultation data that transforms noisy, multi-topic contributions into structured, self-contained argumentative units ready for downstream analysis. We present **GDN-CC**, a manually-curated dataset of 1,231 contributions to the French *Grand Débat National*, comprising 2,285 argumentative units annotated for argumentative structure and manually clarified. We then show that finetuned Small Language Models match or outperform LLMs on reproducing these annotations, and measure their usability for an opinion clustering task. We finally release **GDN-CC-large**, an automatically annotated corpus of 240k contributions, the largest annotated democratic consultation dataset to date.

1 Introduction

The huge improvement of Natural Language Processing (NLP) systems in the last decade, owing to the rapid adoption of the transformer architecture (Vaswani et al., 2017), has made Large Language Models (LLMs) a key component for processing texts in fields such as healthcare (Nazi and Peng, 2024), education (Yan et al., 2024), or business and industry decision-making (Chkirbene et al., 2024). Applications of LLMs in the context of politics and democratic activities have also quickly been considered (Aoki, 2024; Coeckelbergh, 2025; Summerfield et al., 2024). However, using such powerful tools in end-to-end systems for political analysis and decision-making raises serious ethical questions (Galariotis, 2024; Revel and Pénigaud, 2025). This is due to the opacity of the decision-making process and the difficulty to steer it through human intervention. Explanations delivered by Chain-of-Thought methods hardly mitigate these defects, as discussed, *inter alia*, by Barez et al. (2025). LLMs outputs also tend to be heavily biased along multiple dimensions such as gender, age, social background; culturally, they mostly reflect the values of developed western countries (Zhao et al., 2024; Arzaghi et al., 2024; Ranjan et al., 2024) in their decisions and generations, a problem yet to be fully tackled. Moreover, the most popular models are proprietary and their behavior is poorly documented, which could create a dangerous dependence on private companies if they were used to assist democratic processes (Feldstein, 2023). Democratic decision-making requires a fully intelligible and transparent process that humans can supervise, judge, and modify. Each step must be criticizable with a clear entity to hold accountable.

Still, LLMs bring new opportunities for monitoring democratic activities. In particular for participatory democracy and citizen consultations, such as Taiwan's *vTaiwan* platform[1] (2014), the European Citizens' Consultations[2] (2018), or France's *Grand Débat National*[3] (Great National Debate, 2019), LLMs can help decipher the large amount of contributions through topic-modeling, summarization, and community detection (Small et al., 2023; Galariotis, 2024; Guembour et al., 2025).

[1] info.vtaiwan.tw/
[2] oecd-opsi.org/innovations/the-european-citizens-consultations-eccs/
[3] https://granddebat.fr/

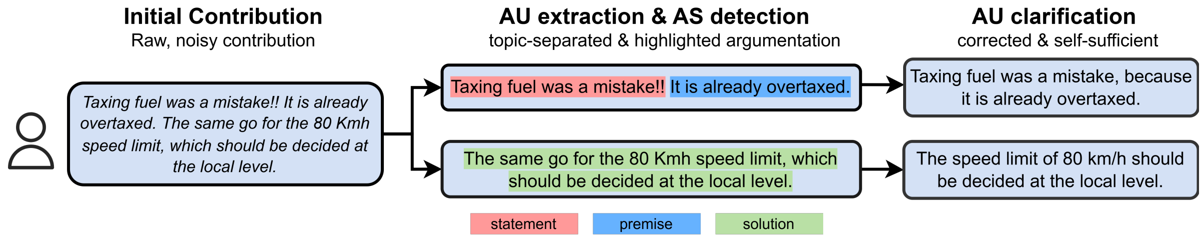


Figure 1: The Corpus Clarification task applied to one contribution. The original contribution (left) is segmented in topics, then in argumentative units (middle), which are automatically reformulated to generate clearer versions of the expressed opinions (right). The contribution was translated to English and simplified for illustrative purposes.

Previous works have explored the usage of LLMs in democratic deliberations analysis (Tessler et al., 2024; Fish et al., 2025). However, these studies focus on relatively small assemblies of up to a hundred participants, with clear on-topic statements. Small et al. (2023) explore larger consultations, but rely on LLM-based[4] clustering and analyses.

Large-scale consultations are notably difficult to process due to the diversity and sometimes noisiness of real-world citizen contributions (Guembour et al., 2025), which can simultaneously address multiple topics and policies, mixing-up facts, vindications, criticisms, proposals, and demands. The consequences are twofold: the synthesis or aggregation of a set of diverse contributions is difficult to explain or inspect; it requires extremely powerful NLP tools, capable of seamlessly handling such textual variability – at the expense, however, of transparency and with increased dependence on very large, closed-weights LLMs.

To avoid this problem, we propose the novel task of **Corpus Clarification**, illustrated in Figure 1. It aims at refining the initial corpus into a set of clear mono-topic opinions that are easier to process, analyze and inspect, even with small models. We first define the multiple steps of the task in Section 3. We then manually annotate French contributions to the *Grand Débat National* for this task. Our AI-human collaborative annotation process, detailed in Section 4, ensures high-quality annotations. It also allows us to evaluate the capacities of multiple LLMs. We explore the resulting corpus **GDN-CC** in Section 5. We then evaluate the capacities of locally-runnable models for Corpus Clarification in Section 6 and show that, when finetuned, they can perform on par or better than large API-based LLMs. We use these models to create **GDN-CC-large**[5] by automatically annotating 240k contributions of the *Grand Débat National* as a resource for both political scientists and NLP researchers. Finally, we showcase the importance of *Corpus Clarification* on an opinion clustering task, obtaining better clusters after clarification than with the original contributions.

2 Related Work

2.1 LLMs and Opinions

LLMs and citizen consultations Recent works have explored LLMs for enhanced deliberation, from both experimental and ethical standpoints. Small et al. (2023) conducted an extensive exploration of the usage of LLMs for deliberation as a complement to the Polis platform (Small et al., 2021),[6] focusing on topic modeling, summarization, prediction of votes and moderation. They highlighted the importance of human supervision in these tasks, for example by validating summaries to ensure accuracy and fairness. Lazar and Manuali (2024) argue that LLMs should be used to enhance *non-instrumental* values of democracy, values that are intrinsically worthwhile such as deliberating or informing (Anderson, 2009). However, as highlighted by Revel and Pénigaud (2025), LLM-based systems for democratic deliberation exist, but they focus on small-scale processes. Tessler et al. (2024) identify common ground in groups of five citizens using an iterative human feedback loop grounded in Jürgen Habermas's theory of deliberative democracy (Habermas, 1985). Fish et al. (2025) use LLMs to aggregate opinions for groups of up to 100 participants within a framework based on Social Choice theory (Sen, 1986). Although built on strong theoretical backbones,

[4]Using Anthropic's Claude model (Bai et al., 2022).

[5]The annotation platform and datasets will be shared in the camera-ready version

[6]`pol.is/`

these works explore AI for deliberation in controlled and strongly constrained scenarios, which are difficult to scale to real-life large-scale deliberations.

Existing work on democratic discourse either (i) focuses on small, manually curated datasets (Tessler et al., 2024; Fish et al., 2025), or (ii) relies on end-to-end LLM-based aggregation and analysis (Small et al., 2023). In contrast, our work focuses on scalable processing that enables traditional, more transparent NLP methods on large corpora, without involving LLMs in the decision loop. This work brings together political science and philosophical recommendations on the use of AI for democracy (Revel and Pénigaud, 2025; Galariotis, 2024; Lazar and Manuali, 2024) with the technological constraints of analyzing large-scale consultation corpora.

Argument mining Many recent works explore LLMs for opinion argument mining (Chen et al., 2024; Guida et al., 2025). This is for instance the case of Ding et al. (2023); Favero et al. (2025) who applied it to the evaluation of opinion essays in education. Our work is also specifically related to argument mining in text written by citizens rather than argumentation professionals. Many works study this settings in tweets (Schaefer and Stede, 2020; Dutta et al., 2020; Iskender et al., 2021) or in citizen consultations (Liebeck et al., 2016; Fierro et al., 2017; Romberg and Conrad, 2021). We contribute to this literature with a manually-annotated and a large automatically-annotated corpora in French, using a three-class span classification designed to be actionable for downstream tasks.

2.2 Citizen Consultations Datasets

In the last few years, multiple works have published citizen consultations: Polis (Small et al., 2021) openly shared twenty of their consultations.[7] Similarly, CoFE (Barriere et al., 2022b) contains 4.2k proposals on the *Conference on the Future of Europe* consultation[8] and the Debating Europe dataset (Barriere et al., 2022a) contains 2.6k opinions from 18 debates related to the European Green Deal.[9] Romberg and Conrad (2021) published five public consultations in German related to mobility transitions. More broadly, political tweets corpora on specific subjects have been released (Li et al., 2021; Iskender et al., 2021; Bilal et al., 2022; Fourati et al., 2024), which are relatively close to citizen consultations.

Our work brings a new large-scale opinion corpus, the first processed specifically for democratically-acceptable downstream tasks. We share both a high-quality manually annotated dataset of 1,231 French contributions, as well as an additional set of 240k automatically annotated contributions from the *Grand Débat National*, respectively **GDN-CC** and **GDN-CC-large**. The resulting corpus is, to our knowledge, the largest annotated democratic citizen consultation corpus shared to the community, and the only one in French.

3 The Corpus Clarification task

The analysis of citizen consultations' corpora strongly relies on aggregating similar opinions together (Small et al., 2023; Galariotis, 2024; Guembour et al., 2025). However, ethics-oriented works do not take into account the noisy reality of such free-form corpora (Galariotis, 2024), and practical works rely heavily on proprietary LLMs to group (Small et al., 2023) or aggregate (Tessler et al., 2024; Fish et al., 2025) opinions. We designed the task of Corpus Clarification to turn a large noisy set of opinions into a clear standardized dataset that can be processed with more traditional and explainable NLP approaches such as embedding-based clustering. We achieve this transformation through a structured, three-step process, illustrated in Figure 1, and detailed below.

The first step is the **Argumentative Units (AU) extraction**. This step aims to identify the various themes or points covered within a contribution. This process breaks down the source material into smaller, distinct Argumentative Units, each focused on a single coherent topic. AUs are not necessarily contiguous, and some segments of the text might not be part of any AU.

The second step is the **Argumentative Structure (AS) detection** inside argumentative units. This task identifies the specific discourse types present within each Argumentative Unit, which provides fine-grained information for argument mining analysis. While the argument mining literature mostly uses a claim-premise classification (Chen et al., 2024), we adopt a three-type classi-

[7] github.com/compdemocracy/openData/

[8] commission.europa.eu/strategy-and-policy/priorities-2019-2024/new-push-european-democracy/conference-future-europe_en

[9] commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en

fication to better align with the requirements of downstream tasks that focus mostly on either policy proposals or citizens' feelings:

- **Statement**: The expression of a sentiment or an observation;
- **Solution**: A proposal for action or policy change;
- **Premise**: A segment providing justification, evidence, or an example to support a statement or solution.

While *Solutions* function similarly to *claims* in traditional argument mining, we adopt this specific terminology to eliminate the ambiguity between *statements* and *claims* that often arises in citizens' contributions.

The final step, **Argumentative Unit Clarification**, aims at both forming context-independent argumentative units and normalizing the style to remove linguistic markers of the writers. Before clarification, AUs may lack self-sufficiency if they rely on context provided outside their boundaries (such as in the preceding text). The clarification process adds this necessary context to ensure that every AU is a fully comprehensible text, independent from its original contribution. Moreover, this step standardizes texts to a shared style and quality, similar to the task of style transfer (Toshevska and Gievska, 2025). This ensures that all citizens are handled similarly in downstream tasks. This is crucial since writing clarity was shown to be related to socioeconomic status (Dölek and Hamzadayi, 2018).

4 Annotation Process

4.1 The *Grand Débat National* corpus

The *Grand Débat National* (GDN) is a nationwide citizen consultation held in France in 2019. The consultation prompted citizens to express their views across four main themes: *Taxation and public spending*, *Organization of the state and public services*, *Democracy and citizenship*, and *Ecological transition*. A significant portion of this consultation involved online questionnaires, each concluding with a critical open-ended prompt: "Do you have anything to add about [theme]?". Our starting point are the responses to these open-ended questions. These specific contributions are valuable as they allow citizens to freely express their opinions. The complete anonymized corpus is publicly available for download, use and sharing via the French *Grand Débat* platform.[10]

Data Preparation The original dataset comprises 355k individual contributions. To ensure data quality and relevance, we applied several filtering steps. After removing duplicate entries, we excluded contributions shorter than 30 characters or longer than 600 characters to focus on responses with meaningful content that were not excessively long for annotation, resulting in 240k contributions. Finally, aiming for diversity in response complexity, we ensured a uniform distribution across the four themes and across lengths of contributions. We used the French sentencizer provided by the `spaCy` library (Vasiliev, 2020) to estimate the number of sentences in each text, as sentences were used as an approximation of opinions in a previous analysis (Guembour et al., 2025).

4.2 Annotation guidelines and process

Five native French speakers specializing in political science were hired to annotate 300 contributions each. The annotation was performed in two main steps: segmentation and clarification.

Contributions segmentations The tasks of *Argumentative unit segmentation* and *Argumentative structure detection* were performed jointly. Annotators were instructed to parse the raw text and identify three primary segment types: Premises, Statements, and Solutions. These segments were grouped into topically coherent Argumentative Units. This combined segmentation and detection process was designed to accelerate the annotation process and to ensure that each AU contains only salient, argument-relevant information.

Argumentative units clarification Once the initial segmentation step was complete, the argumentative units were processed for clarification by a randomly selected LLM in a pool of four. The resulting clarifications were then displayed to annotators, who were asked to revise them if necessary. This hybrid human-LLM process allowed us to speed-up the annotation. It also yielded useful data to compare the ability of the various LLMs systems to clarify argumentative units.

Annotation was performed in two phases. In the first phase, annotators could regenerate clarifications using different LLMs to maximize annotation quality. During the second phase, which took place

[10] https://granddebat.fr/pages/donnees-ouvertes

during the annotation of the last quarter of the contributions, annotators were asked to modify the first generated clarification, even when its quality was poor. This allowed us to characterize the types and frequency of errors made by LLM systems, which we report in Section 5.4.

The process totaled 1,553 annotations of 1,231 individual contributions, the remaining allowing for the computation of inter-annotator agreement. We describe the annotators' training and monitoring in Appendix A.2.

4.3 Annotation tool

As there was, to our knowledge, no existing tool suitable for our two-step annotation process, we implemented our own. This interface enables the annotators to perform segmentation and rewriting in a single step, so that the clarification can benefit from the understanding of the contribution gained during segmentation. To generate the initial automatic clarifications, we relied on four families and sizes of models: `GPT-4.1` (Achiam et al., 2023), `Qwen-3-32b` (Yang et al., 2025), `Llama-3.1-8b` and `Llama-3.3-70b` (Grattafiori et al., 2024). More information regarding the annotation tool, its implementation and the annotation process is in Appendix A.1.

5 GDN-CC: A French Dataset for the task of Corpus Clarification

5.1 Statistics of the dataset

GDN-CC consists of 1,231 contribution that have been manually-annotated for the Corpus Clarification task. Table 1 reports basic statistics regarding this dataset. Overall, GDN-CC contains significantly more solutions (57.9%) than statements (21.2%) and premises (20.9%). This result is expected, as citizens who are actively going to the consultation platform expect to be heard and to influence the decision-making process. Interestingly, the *Democracy and Citizenship* theme contains fewer argumentative units, and in proportion less solutions (54% versus 58-60%) and more statements (24% versus 19-21%) than the other themes. Participants may struggle to express solutions for these less-discussed topics, resulting in more descriptive statements. The number of argumentative units per contribution spans from one to eleven, and the number of argumentative segments from one to twelve. This highlights the complex and variable nature of the corpus. More statistics and figures are given in Appendix B.1.

5.2 Discrepancies and agreements between annotators

We evaluated the inter-annotator agreement separately for the two tasks of argumentative unit (AU) extraction and argumentative structure (AS) detection.

AU extraction For AU extraction, we rely on a span-limitation based metric, WindowDiff (Pevzner and Hearst, 2002), used in previous studies for inter-annotator agreements of span annotations (Javorský et al., 2025). It uses a sliding window of size k ($k = 15$ tokens in our case) and evaluates, at each step, how many boundaries the two annotations have within the window. These are aggregated in a probability of boundary disagreement within a sliding window, with lower values indicating greater consistency of the segmentation. Following recent works on segmentation (Ding et al., 2023; Favero et al., 2025), we also compute a token-overlap metric: Given two spans S_1 and S_2 annotated by two annotators, we evaluate $s(S_1, S_2) = \min(\frac{|S_1| \cap |S_2|}{|S_1|}, \frac{|S_1| \cap |S_2|}{|S_2|})$ for each pair of spans in the annotation. To ensure optimal matching, we solve the span-pair assignment problem the `SciPy` package[11] (Virtanen et al., 2020). A match is defined if $s(S_1, S_2)$ is above a set threshold λ.

We observed a strong level of agreement between annotators with both metrics. The mean and median WindowDiff values over all annotations are 0.09 and 0.08 respectively. Similarly, the token overlap metric yields micro- and macro-F1 of 0.72 and 0.77 for a 50% overlap threshold ($\lambda = 0.5$, a value used in previous studies), and 0.42 and 0.44 for perfect alignment ($\lambda = 1$). Table 3 in Appendix A.3 also reports precision and recall.

AS detection For AS detection, we computed the inter-annotator agreement independently of the extracted argumentative units. We considered the task as a token tagging problem. For each token, annotators classified it as *statement*, *premise*, *solution*, or did not classify it (*none* class). We computed the agreement ratio between two opinions as the proportion of tokens that were classified similarly by the two annotators. We found a mean agreement ratio of 0.65 (median: 0.67). However, the agreement varies significantly by category, reflecting the

[11] using the `linear_sum_assignment` method

Theme	contribs.	AUs	statements	solutions	premises
Taxation and Public Spending	312	594	175 (18.7%)	556(59.6%)	202 (21.7%)
Ecological Transition	305	590	187 (20.5%)	543 (59.5%)	183 (20.0%)
Organization of the State	308	577	197 (21.4%)	532 (57.9%)	190 (20.7%)
Democracy and Citizenship	306	524	212 (24.2%)	474 (54.2%)	187 (21.4%)
Total	**1231**	**2285**	**771** (21.2%)	**2105** (57.9%)	**762** (20.9%)

Table 1: Number of contributions, argumentative units, statements, solutions, and premises per theme and in total, percentages are calculated row-wise.

nature of citizen speech: we found the strongest agreement for *solutions*, reaching 79%. This category is the most critical for downstream policy analysis, and its high reliability demonstrates that annotators can consistently identify actionable citizen proposals. Agreement rates for *premises* and *statements* were lower (58.7% and 50.3% respectively). This discrepancy stems from a structural confusion between the two classes: in non-expert discourse, a sentiment (*statement*) frequently acts as a justification (*premise*) for a solution. Such confusions were similar to those observed in previous argument mining studies (Ding et al., 2023). By merging statements and premises into a single class, we obtained a 74.3% agreement between annotators for the joint class. Although we kept the three-class granularity in the rest of this study, future work could consider merging them into a single class. Examples of confusions between statements and premises, and the confusion matrix for the three classes is in Appendix A.3.

5.3 Evaluating the capacities of models for Clarification

During the first annotation phase, we considered four models: `GPT-4.1`, `Qwen-32b`, `Llama-70b`, and `Llama-8b` for AU clarification. Annotators could iterate through multiple AI-generated clarifications from randomly-chosen LLMs before accepting the one they edit. Observing this process allows us to derive a comparative evaluation of these models using a probabilistic model of annotators' behavior. In summary, we treat the unknown quality of each model l as a parametric distribution Q_l. A user accepts the LLM clarification if Q_l exceeds a threshold τ_k, which varies with the number of clarification attempts k. If an output is accepted, its quality q_l is observed via the ROUGE-L score against the final human-validated text. By jointly optimizing the parameters of Q_l and τ_k to maximize the likelihood of user actions, we found that `GPT-4.1` performed best ($\mu = 0.94$), closely followed by `Llama-70B` ($\mu = 0.93$) and `Qwen-32B` ($\mu = 0.92$. which proved to be close competitors for zero-shot clarification. `Llama-8B` ($\mu = 0.90$) yielded good performances despite its limited size. This finding provides a first hint at the abilities of smaller models for the task, that we further observe in Section 6. A detailed explanation of the model and its estimation is in Appendix D.

5.4 Qualitative exploration of the clarifications corrections

Although the first phase of the annotation delivered high-quality clarifications and an evaluation of the LLMs systems, it made it difficult to perform a qualitative analysis of clarification errors. This is because annotators were permitted to reject some low-quality LLM outputs in their revision process. The second annotation phase required annotators to accept (and edit when necessary) the first automatically generated clarification, irrespective of its quality. We manually examined 100 argumentative units annotated during this phase (25 for each LLM) for which the clarification was edited by the annotator. We identified four main types of errors:

- **Over-analysis**: LLM added analysis or a conclusion absent from the original text;
- **Miscomprehension**: LLM did not fully regenerate the opinion, by being wrong or by omitting important information;
- **Over-specificity**: LLM added information present in the text but removed by the annotator, such as unimportant details or content from other AUs;
- **Misformulation**: LLM output contained the right information, but its phrasing was unnatural or contained an extraneous introductory phrase.[12]

Table 8 in Appendix B presents an example of each type of error, while Table 7 reports the propor-

[12] Such as "here is the argument: **{{argument}}**".

tion of each error for each model. Over-specificity, miscomprehension and misformulation are seldom found (respectively 3%, 12% and 13% of all errors) but all four models consistently struggle with over-analysis (72% of all errors), even when specifically prompted not suppress any information not expressed in the text. This behavior is highly detrimental in the context of democratic processes, especially when models add justifications that were not present in the author's contribution. Complementing this finding, further analyses presented in Appendix B.2 suggest that a significant part of the human annotation effort of clarifications was to remove information added by the LLMs.

6 Experiments

Focusing on having locally-runnable systems and not relying on large proprietary LLMs, we explore the use of different systems to automate the tasks of AU extraction and AS detection. In particular, we evaluate the capacities of Small Language Models (SLMs, $\leq 10B$ parameters), including their instruct and task-specific finetuned variants. For all three tasks, we thus evaluated `llama-3.1-8B` (Grattafiori et al., 2024), `Mistral-7B-v0.3` (Jiang et al., 2023), `Qwen2.5-7B` (Team et al., 2024b) and `gemma-2-9b` (Team et al., 2024a) families of models. We used 70% of the data for training and 15% each for validation and testing. We used `GPT-4.1` as a comparison baseline for all tasks. Results are reported in Table 2. For finetuned models, we report the results of the best-performing learning rate. See Appendix C for finetuning details and figures.

6.1 Argumentative Units segmentation

Given a written contribution to the citizen consultation, the system has to extract the spans of texts corresponding to different argumentative units. More specifically, we prompt the models to output argumentative units as a list. Table 2's first row displays the Macro-F1 and Micro-F1 using the token overlap metric used for inter-annotator agreement (Section 5.2). While the base models mostly failed this task, instruct models perform better, with supervised finetuning further improving the results. In particular, we found that `Qwen2.5-7B` outperformed all models on Macro-F1 (0.81) and performed on par with `GPT-4.1` on Micro-F1 (0.75).

6.2 Argumentative Structure detection

Given an argumentative unit, the system has to extract and classify the spans as *statement*, a *premise* or a *solution*. We provide the LLMs both the initial contribution and the argumentative unit from which spans need to be extracted. The second row in Table 2 shows the Macro-F1 and Micro-F1 using a label-constrained version of the token overlap metric used for AU segmentation:

$$s_{cons}(S1, S2) = \delta_{label(S1),label(S2)} \times s(S1, S2),$$

where $\delta_{i,j} = 1$ if $i = j$ else 0. As for argumentative units segmentation, non-instruct SLMs are unable to perform this task, while instruct models achieve much better results. Regarding finetuned models, `Gemma-2-9B` (Macro-F1: 0.79, Micro-F1: 0.73) outperforms other SLMs and performs on par with `gpt-4.1` (Macro-F1: 0.76, Micro-F1: 0.76). As this task can be seen as a simple span extraction and tagging problem, we also experimented with a small encoder-based model in Appendix C.2 which proved effective for resource-constrained scenarios (Macro-F1: 0.76, Micro-F1: 0.70).

6.3 Argumentative Unit clarification

Given an argumentative unit, the system has to rewrite it as a clear, self-sufficient argument by extracting only the important information from the text. We provided the models with the initial contribution for context and the argumentative unit to clarify. We report BERTScore[13] (Zhang et al., 2019) and Rouge-L[14] (Lin, 2004) for all models in Table 2's last row. While all finetuned SLMs achieve comparable performance in terms of BERTScore, slightly exceeding `GPT-4.1`, more pronounced disparities are observed for the ROUGE-L metric. We hypothesize that informational content remains consistent across models but those with higher ROUGE-L scores produce a semantic structure that more closely aligns with the style of human annotators, explaining lower scores of non-finetuned models. We find `Gemma-2-9B` (BERTScore: 0.86, ROUGE-L: 0.60) to outperform other systems including `GPT-4.1`.

Using `GPT-4.1-nano` as a judge, we compared finetuned clarifications against initial LLM outputs. Finetuned `Gemma-2-9b` was preferred 66% of the time, initial outputs 26% of the time. Qualitative review confirms that finetuning significantly mitigated the "over-analysis" issue. See Appendix E for the evaluation prompt and examples of fixed over-analyses.

[13]Using `bert-base-multilingual-cased` as backbone.
[14]using the `rouge-score` 0.1.2 package

Task	Metric	GPT-4.1	Llama-8b			Mistral-7b			Qwen-7b			Gemma-9b		
		Base	Base	Inst.	Ft.	Base	Inst.	Ft.	Base	Inst.	Ft.	Base	Inst.	Ft.
AU Extraction	Micro-F1	**0.75**	0.09	0.54	0.61	0.04	0.62	0.71	0.46	0.66	**0.75**	0.03	0.67	0.66
	Macro-F1	0.73	0.10	0.59	0.72	0.03	0.58	0.78	0.40	0.66	**0.81**	0.04	0.69	0.76
AS Detection	Micro-F1	**0.76**	0.12	0.31	0.65	0.02	0.17	0.70	0.34	0.56	0.67	0.00	0.49	0.73
	Macro-F1	0.76	0.12	0.36	0.71	0.01	0.15	0.78	0.37	0.64	0.75	0.00	0.57	**0.79**
AU Clarification	BERTScore	0.81	0.64	0.79	0.85	0.26	0.77	**0.86**	0.65	0.82	0.85	0.59	0.80	**0.86**
	Rouge-L	0.45	0.19	0.39	0.56	0.05	0.33	0.56	0.18	0.52	0.54	0.21	0.39	**0.60**

Table 2: Performance metrics across models and tasks with the best values for each metric in **bold** and the second best values underlined. Low performance of Base LLMs are due to their inability to consistently generate the expected output format.

6.4 Implication for downstream tasks

Clustering similar opinions is a key step in many downstream tasks such as topic modeling or summarization (Small et al., 2023).

To assess the potential improvements for democratic downstream tasks, we evaluate the effect on clustering quality of (i) AU segmentation and (ii) AU clarification. We use BERTopic (Grootendorst, 2022) for clustering and topic modeling and for the three types of texts – namely, the initial contributions, the extracted AUs, and the associated clarifications. We evaluate clustering quality using `GPT-4.1-nano` as a judge in a pairwise comparison setup. From each clustering, we sample pairs of texts coming from the *same* cluster. The model is then presented with pairs from two clusterings; the task is to identify the most coherent pair based on thematic specificity and consistency. We tested three settings with 100 sampled pairs per theme:

1. Initial contributions vs. Argumentative units;
2. Argumentative units vs. Clarifications;
3. Argumentative units vs. Argumentative units, using the clusters based on clarifications.

The latter setting serves as a control to ensure the judge's preference is driven by thematic coherence rather than the improved syntax of clarifications.

The LLM judge showed a near-total preference for the more granular units in all settings. In Setting 1, argumentative units were prefered over raw contributions in 84% of cases. Setting 2 showed a 91% preference for clarifications. Setting 3 confirmed this trend (90% preference), showing that the coherence gain persists even when evaluating the same text type (AU) rearranged into clusters from clarifications. All results are statistically significant ($p < 0.001$ in all settings). Prompts and statistical analyses are in Appendix F.

6.5 Annotating GDN-CC-large

Having shown that finetuned SLMs were reliable for the task of Corpus Clarification, and that this task was actually helping clustering and potentially other downstream tasks, we applied our pipeline to the full corpus consisting of 240k contributions, yiedling a dataset of 300k argumentative units, split into 155k statements, 282k solutions and 62k premises. Appendix G reports detailed statistics regarding this corpus.

7 Conclusion

In this work, we formalized **Corpus Clarification** as a first step for the automated analysis of large-scale democratic consultations. By defining and evaluating a three-step pipeline, we have developed a structured methodology to transform noisy, multi-topic citizen contributions into standardized, self-sufficient opinions. Our results, based on the manual annotation of **GDN-CC** demonstrate that finetuned SLMs match or outperform large proprietary models. Specifically, derivatives from `Qwen2.5-7b` and `Gemma-2-9b` proved amply sufficient to obtain high quality annotations. Furthermore, we provided empirical evidence that this clarification process significantly enhances the performance of embedding-based clustering, a critical prerequisite for reliable topic modeling and debate facilitation at scale. By releasing the 240k processed contributions **GDN-CC-large**, we provide the research community with the largest democratic consultation corpus annotated for downstream tasks.

Limitations

Though we bring a new approach to LLMs for citizen consultation analysis, we acknowledge some limitations to our work. While the Corpus Clarification task was thought of as language and topicagnostic, our work only focus on French contributions for the *Grand Débat National* and do not explore other languages or other types of consultations, with potential unforeseen language- or consultation-specific limitations to this framework. Moreover, we did not perform any extensive exploration of LLM's prompts, which could further improve their capacities for the Corpus Clarification task. We also focus on a restricted set of LLMs systems (both for the manual annotation and the automatic annotation). Considering a larger set could bring deeper insights about the capacities of various systems for this task. Finally, some evaluations rely on LLM-as-a-Judge which, while giving statistically significant results, is known to be sensitive to prompting.

Ethical Considerations

The deployment of AI in democratic processes necessitates a rigorous examination of the trade-offs between computational efficiency and representational fidelity. The argumentative unit clarification step of our pipeline prioritizes legibility to facilitate large-scale analysis. However, standardizing raw expression carries the inherent risk of discarding sentiment, nuance, or the unique "voice" of the contributor. While we demonstrate that this process significantly improves downstream clustering, an ideal democratic analysis should not completely forget the original form of expression. Moreover, while finetuned LLMs reduce reliance on opaque systems for analysis, the models we used are still developed by private entities. This does not fully resolve the problem of private dependence in public democratic infrastructure.

Concerning the data from the *Grand Débat National* that we use, all participants were aware that their answers would be shared publicly. While anonymity and the removal of hate speech was ensured for the manual annotation, we did not further check if there was potential identifying information or hate speech in the automatically-annotated data. However, the original data source from the French open data platform affirms that the data has been anonymized.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. *CoRR*, abs/1907.10902.

Elizabeth Anderson. 2009. Democracy: instrumental vs. non-instrumental value. *Contemporary debates in political philosophy*, pages 213–227.

Wissam Antoun, Francis Kulumba, Rian Touchent, Éric de la Clergerie, Benoît Sagot, and Djamé Seddah. 2024. CamemBERT 2.0: A smarter French language model aged to perfection. *Preprint*, arXiv:2411.08868.

Wissam Antoun, Benoît Sagot, and Djamé Seddah. 2025. Modernbert or debertav3? examining architecture and data influence on transformer encoder models performance. *Preprint*, arXiv:2504.08716.

Goshi Aoki. 2024. Large language models in politics and democracy: A comprehensive survey. *arXiv preprint arXiv:2412.04498*.

Mina Arzaghi, Florian Carichon, and Golnoosh Farnadi. 2024. Understanding intrinsic socioeconomic biases in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 49–60.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *Preprint*, arXiv:2204.05862.

Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clément Neo, Isabelle Lee, Alasdair Paren, Adel Bibi, Robert Trager, Damiano Fornasiere, John Yan, Yanai Elazar, and Yoshua Bengio. 2025. Chain-of-thought is not explainability. *Preprint, alphaXiv:2025.02v1*.

Valentin Barriere, Alexandra Balahur, and Brian Ravenet. 2022a. Debating Europe: A multilingual multi-target stance classification dataset of online debates. In *Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences*, pages 16–21, Marseille, France. European Language Resources Association.

Valentin Barriere, Guillaume Guillaume Jacquet, and Leo Hemamou. 2022b. CoFE: A new dataset of intra-multilingual multi-target stance classification from an online European participatory democracy platform. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*

(Volume 2: Short Papers), pages 418–422, Online only. Association for Computational Linguistics.

Iman Munire Bilal, Bo Wang, Adam Tsakalidis, Dong Nguyen, Rob Procter, and Maria Liakata. 2022. Template-based abstractive microblog opinion summarization. *Transactions of the Association for Computational Linguistics*, 10:1229–1248.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. 2024. Exploring the potential of large language models in computational argumentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2309–2330, Bangkok, Thailand. Association for Computational Linguistics.

Zina Chkirbene, Ridha Hamila, Ala Gouissem, and Unal Devrim. 2024. Large language models (llm) in industry: A survey of applications, challenges, and trends. In *2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)*, pages 229–234. IEEE.

Mark Coeckelbergh. 2025. LLMs, truth, and democracy: an overview of risks. *Science and Engineering Ethics*, 31(1):4.

Yuning Ding, Marie Bexte, and Andrea Horbach. 2023. Score it all together: A multi-task learning study on automatic scoring of argumentative essays. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13052–13063, Toronto, Canada. Association for Computational Linguistics.

Onur Dölek and Ergün Hamzadayi. 2018. Comparison of writing skills of students of different socioeconomic status. *International Journal of Progressive Education*, 14(6):117–131.

Subhabrata Dutta, Dipankar Das, and Tanmoy Chakraborty. 2020. Changing views: Persuasion modeling and argument extraction from online discussions. *Information Processing & Management*, 57(2):102085.

Lucile Favero, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. 2025. Leveraging Small LLMs for Argument Mining in Education: Argument Component Identification, Classification, and Assessment. *Preprint*, arXiv:2502.14389.

Steven Feldstein. 2023. The consequences of generative ai for democracy, governance and war. In *Survival: October–November 2023*, pages 117–142. Routledge.

Constanza Fierro, Claudio Fuentes, Jorge Pérez, and Mauricio Quezada. 2017. 200K+ crowdsourced political arguments for a new Chilean constitution. In *Proceedings of the 4th Workshop on Argument Mining*, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics.

Sara Fish, Paul Gölz, David C. Parkes, Ariel D. Procaccia, Gili Rusak, Itai Shapira, and Manuel Wüthrich. 2025. Generative Social Choice. *Preprint*, arXiv:2309.01291.

Chayma Fourati, Roua Hammami, Chiraz Latiri, and Hatem Haddad. 2024. PoliTun: Tunisian political dataset for detecting public opinions and categories orientation. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 178–185, Trento. Association for Computational Linguistics.

Ioannis Galariotis. 2024. *Is Artificial Intelligence Threatening Democracy?* European University Institute.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Miguel Grinberg. 2018. *Flask web development*. " O'Reilly Media, Inc.".

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Sami Guembour, Catherine Dominguès, and Sabine Ploux. 2025. Semantic analysis experiments for French citizens' contribution : Combinations of language models and community detection algorithms. In *Proceedings of the 16th International Conference on Computational Semantics*, pages 231–241, Düsseldorf, Germany. Association for Computational Linguistics.

Matteo Guida, Yulia Otmakhova, Eduard Hovy, and Lea Frermann. 2025. LLMs for argument mining: Detection, extraction, and relationship classification of pre-defined arguments in online comments. In *Proceedings of the 23rd Annual Workshop of the Australasian Language Technology Association*, pages 176–191, Sydney, Australia. Association for Computational Linguistics.

Jürgen Habermas. 1985. *The theory of communicative action: Volume 2: Lifeword and system: A critique of functionalist reason*, volume 2. Beacon press.

Neslihan Iskender, Robin Schaefer, Tim Polzehl, and Sebastian Möller. 2021. Argument mining in tweets: comparing crowd and expert annotations for automated claim and evidence detection. In *International Conference on Applications of Natural Language to Information Systems*, pages 275–288. Springer.

Dávid Javorský, Ondřej Bojar, and François Yvon. 2025. MockConf: A student interpretation dataset: Analysis, word- and span-level alignment and baselines. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16339–16356, Vienna, Austria. Association for Computational Linguistics.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Seth Lazar and Lorenzo Manuali. 2024. Can LLMs advance democratic values? *arXiv preprint arXiv:2410.08418*.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365, Online. Association for Computational Linguistics.

Matthias Liebeck, Katharina Esau, and Stefan Conrad. 2016. What to do with an airport? mining arguments in the German online participation project tempelhofer feld. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 144–153, Berlin, Germany. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Rajesh Ranjan, Shailja Gupta, and Surya Narayan Singh. 2024. A comprehensive survey of bias in llms: Current landscape and future directions. *arXiv preprint arXiv:2409.16430*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Manon Revel and Théophile Pénigaud. 2025. AI-Facilitated Collective Judgements. *Preprint*, arXiv:2503.05830.

Julia Romberg and Stefan Conrad. 2021. Citizen involvement in urban planning - how can municipalities be supported in evaluating public participation processes for mobility transitions? In *Proceedings of the 8th Workshop on Argument Mining*, pages 89–99, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tim Sainburg, Leland McInnes, and Timothy Q Gentner. 2021. Parametric umap embeddings for representation and semisupervised learning. *Neural Computation*, 33(11):2881–2907.

Robin Schaefer and Manfred Stede. 2020. Annotation and detection of arguments in tweets. In *Proceedings of the 7th Workshop on Argument Mining*, pages 53–58, Online. Association for Computational Linguistics.

Amartya Sen. 1986. Social choice theory. *Handbook of mathematical economics*, 3:1073–1181.

Christopher Small, Michael Bjorkegren, Timo Erkkilä, Lynette Shaw, and Colin Megill. 2021. Polis: Scaling deliberation by mapping high dimensional opinion spaces. *Recerca: revista de pensament i anàlisi*, 26(2).

Christopher T Small, Ivan Vendrov, Esin Durmus, Hadjar Homaei, Elizabeth Barry, Julien Cornebise, Ted Suzman, Deep Ganguli, and Colin Megill. 2023. Opportunities and risks of LLMs for scalable deliberation with polis. *arXiv preprint arXiv:2306.11932*.

Christopher Summerfield, Lisa Argyle, Michiel Bakker, Teddy Collins, Esin Durmus, Tyna Eloundou, Iason Gabriel, Deep Ganguli, Kobi Hackenburg, Gillian Hadfield, and 1 others. 2024. How will advanced ai systems impact democracy? *arXiv preprint arXiv:2409.06729*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024a. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team and 1 others. 2024b. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Michael Henry Tessler, Michiel A Bakker, Daniel Jarrett, Hannah Sheahan, Martin J Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C Parkes, and 1 others. 2024. Ai can help humans find common ground in democratic deliberation. *Science*, 386(6719):eadq2852.

Martina Toshevska and Sonja Gievska. 2025. Llm-based text style transfer: Have we taken a step forward? *IEEE Access*.

Yuli Vasiliev. 2020. *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, and 1 others. 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. Trl: Transformer reinforcement learning. https://github.com/huggingface/trl.

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Jinman Zhao, Yitian Ding, Chen Jia, Yining Wang, and Zifan Qian. 2024. Gender bias in large language models across multiple languages. *arXiv preprint arXiv:2403.00277*.

A Annotation process and platform

A.1 Annotation Platform

The annotation platform was developed in javascript, using the React.js[15] and Next.js frameworks[16]. The back-end server to handle contributions distributions, annotator accounts, and saving the annotations was developed in Python, using the Flask framework (Grinberg, 2018). For the automatic clarification, we use the Open AI's API[17] for `GPT-4.1` and Groq inference provider[18] for the three other models.

The platform contains a welcome/tutorial page accessible to all (Figure 3), as well as an example page to see some annotations already done (Figure 4). Annotators are given a unique token. They must validate three examples for which the expected annotation is given and explained before having access to the annotation interface shown in Figure 2. Annotators also have access to an account page (Figure 5), from which they can see the number of annotations they have done so far, as well as all the examples they annotated. They are allowed to go back to previous annotations to re-do them if needed. The interface also implements an administration page for the allowed tokens to see all done annotations (Figure 6). During the annotation, annotators were allowed to report and skip contributions to annotate if they were not understandable, contained hate speech, were too long, or if they contained personal information.

We share our code for the platform to the community[19]. While all written parts are in French, They can easily be modified to fit other annotations needs in other languages.

A.2 Annotators Training and Monitoring

Annotators are graduate students in political science who were paid 25 euros an hour for the annotation (300 euros each in total), above the minimum wage in France. All annotators participated in a live presentation of the task before having access to the platform. The platform, tasks, and expectations were explained to them. Examples of gold annotations were provided to further explain the task. After the meeting, annotators had to annotate three controlled examples before starting the real annotation process.

After a third of the annotation was done, a meeting was held with four of the five annotators. The fifth annotator could not attend, but was sent a report of the meeting afterwards. The objective of this meeting was to discuss annotation difficulties and methods and to discuss some contributions for which two annotations by two different annotators were significantly different. This meeting greatly helped with improving the quality of annotations: The mean WindowDiff agreement went from 0.10 before the meeting to 0.08 (median 0.09 to 0.06). The token-overlap metric went from 0.74 and 0.69 macro-F1 and micro-F1 to 0.83 and 0.80 respectively. The mean and median accuracy of span types (*statement/premise/solution*) went from 0.64 and 0.63 to 0.66 and 0.77.

[15] react.dev/
[16] nextjs.org/
[17] platform.openai.com/
[18] groq.com/
[19] In the camera-ready version of the paper

Figure 2: Annotation interface. On the left is the segmented contribution, and on the right the clarification for each argumentative unit.

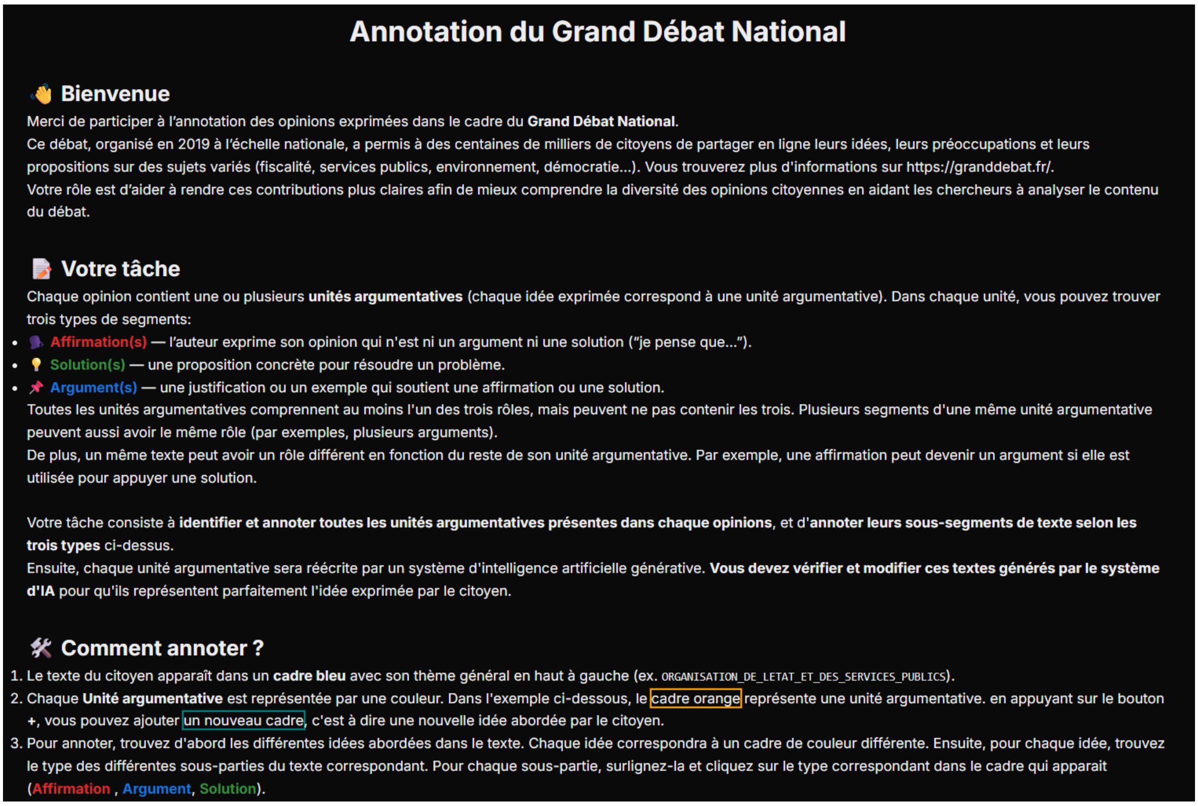


Figure 3: Welcome page. See Figure 7 for translation of the task explanation.

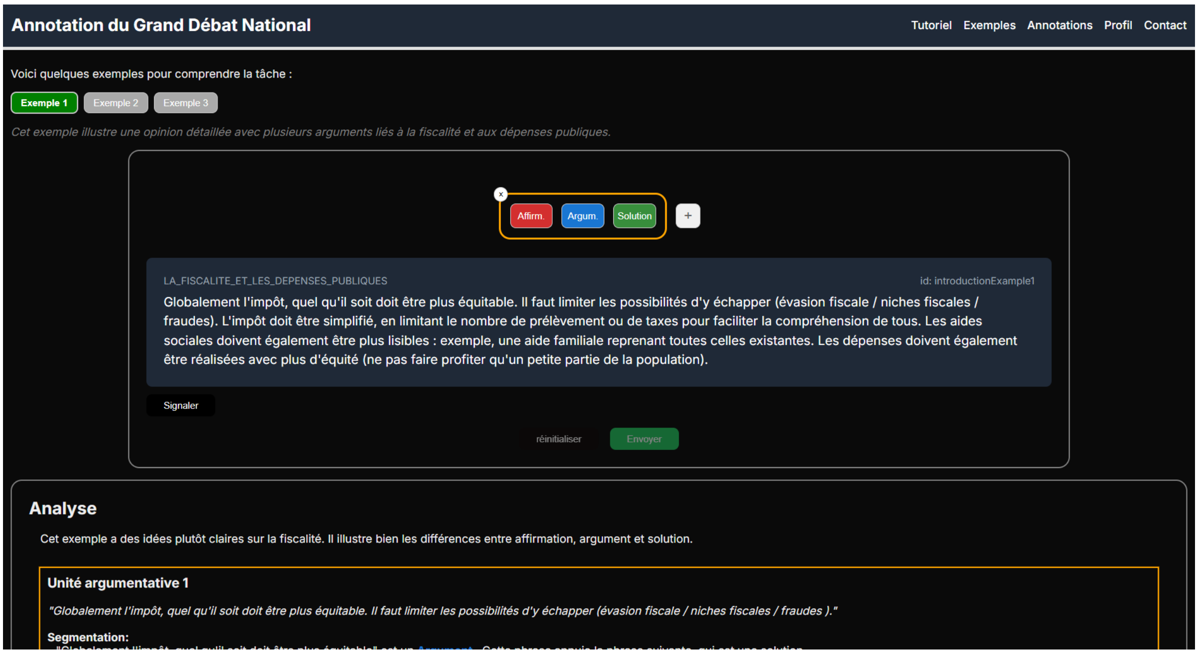


Figure 4: Example page.

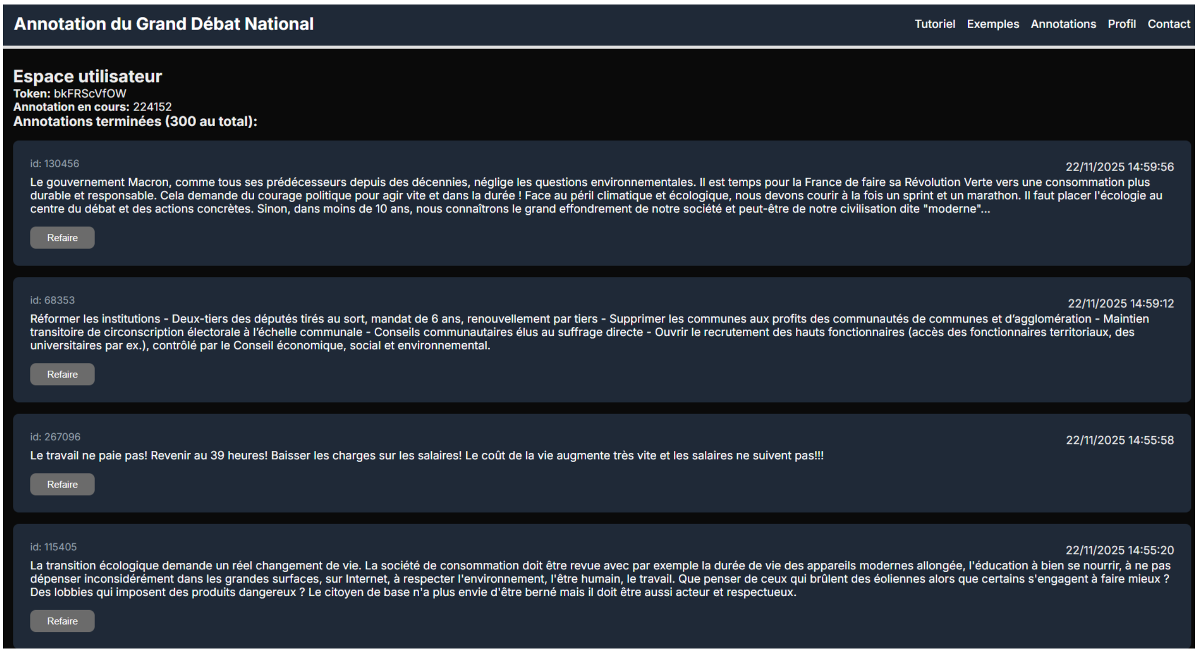


Figure 5: Account page.

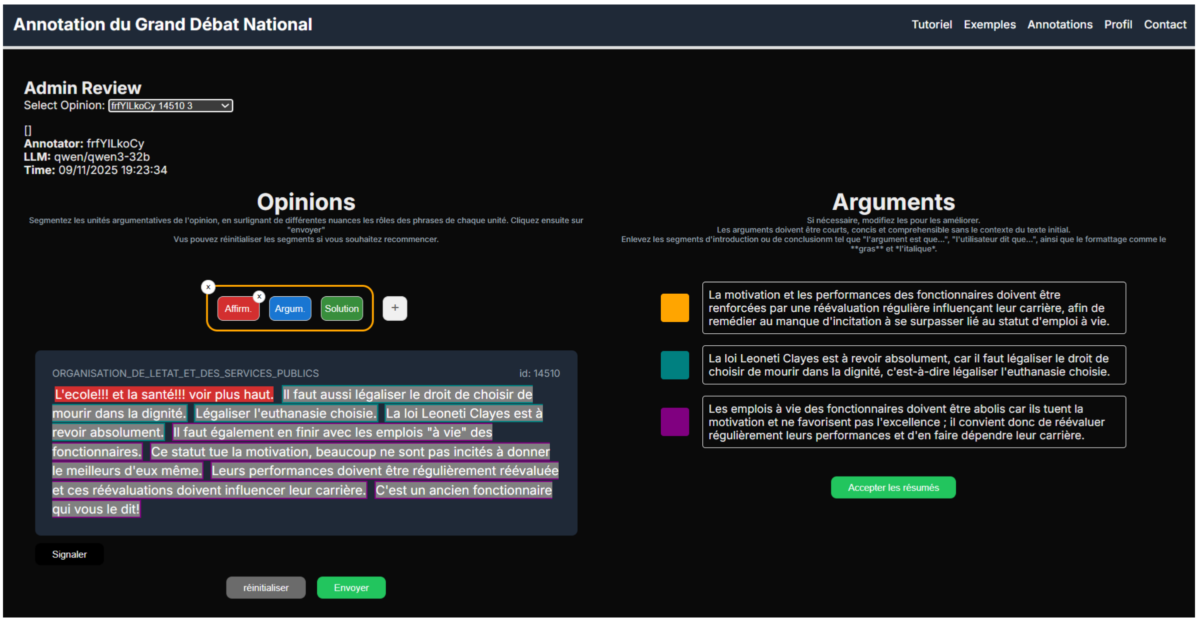


Figure 6: Administration page.

A.3 Inter-Annotator Agreement

Table 3 provides micro and macro precision, recall, and F1 for the task of Argumentative Units segmentation. Figure 8 shows the proportions of each annotated type pairs for tokens' argumentative type. We find that 43% of the text tokens are annotated twice as *solutions*, showing both the importance of this type and the strong agreement between annotators. This figure also displays the rather strong confusion between *statements* and *premises*.

		$\lambda = 0.5$	$\lambda = 1$
precision	micro	0.78	0.45
	macro	0.80	0.42
recall	micro	0.65	0.38
	macro	0.76	0.41
F1	micro	0.71	0.41
	macro	0.76	0.42

Table 3: Argumentative Units match between annotators.

We provide some examples of confusions between *statements* and *premises*. All texts are originally French and translated to English. In many cases, the confusion is made when the citizen uses its own belief or experience as justification:

- "*Every country has its own history and customs.*" in "*Every country has its own history and customs. When integrating into a country, you must be able to respect and accept those of the host country.*";
- "*Over 3 km near my house: 70-90-70-90-80-70-50*" in "*You have to stop changing speed limits too often. Over 3 km near my house: 70-90-70-90-80-70-50*";
- *"There have been more significant climate variations throughout history."* in "*Let's stop focusing on humanity's ability to influence "climate change." There have been more significant climate variations throughout history.*".

B Complementary Analysis of the Manually-Annotated Corpus

B.1 Statistics of the Corpus

In this section, we provide more statistics about the manually-annotated corpus. Figures 9 and 10 respectively displays the distributions of number of argumentative units and of argumentative spans types per contribution.

B.2 Characterization of Annotators Corrections

In this section, we explore the corrections made by the annotators in details by computing the differences between the Argumentative Units, the clarification generated by AI systems, and the final clarifications accepted by the annotators using different metrics and systems.

Levenshtein edit distance To better understand the modifications introduced by human annotators, we computed the edit distance between: (i) the

```
Each opinion contains one or more argumentative units (each idea expressed
corresponds to one argumentative unit). In each unit, you can find
three types of segments:

- Statement(s): the author expresses their opinion, which is neither an argument nor
a solution (“I think that...”).
- Solution(s): a concrete proposal to solve a problem.
- Argument(s): a justification or example that supports a statement or solution.

All argumentative units include at least one of the three roles, but may not contain
all three. Several segments of the same argumentative unit may also have the
same role (for example, several arguments). In addition, the same text may have a
different role depending on the rest of its argumentative unit.
For example, a statement can become an argument if it is used to support a solution.


Your task is to identify and annotate all argumentative units present in each
opinion, and to annotate their text sub-segments according to the three types
above. Each argumentative unit will then be rewritten by a generative
artificial intelligence system. You must check and modify these texts generated by
the AI system so that they perfectly represent the idea expressed by the citizen.
```

Figure 7: System Prompt for the LLM-as-a-Judge evaluation of clustering (translated from French to English).

original argumentative unit and the LLM clarification, (ii) the original argumentative unit and the final clarification, and (iii) the LLM clarification and the final clarification.

The edit distance, also known as Levenshtein distance, measures the similarity between two strings by counting the minimum number of character-level insertions, deletions, and substitutions required to transform one string into the other. The results are presented in Table 4.

On average, the edit distance between the original text and the LLM clarification is higher than between the original text and the final clarification. This suggests that LLMs tend to introduce additional information that is later removed by human annotators. This interpretation is consistent with the smaller edit distance observed between the LLM clarification and the final clarification.

Furthermore, 40% of the final clarifications (331 out of 828) are entirely contained within the corresponding LLM clarification, ignoring punctuation. This indicates that, in many cases, the annotator's intervention consists mainly in deleting superfluous content rather than rewriting the clarification. We further compare these results with the edit distances computed for argumentative units whose clarifications were not modified by annotators, shown in the last row of Table 4. The edit distances in this case are relatively high and comparable to those observed between the original units and the LLM clarifications in Table 4. This suggests that LLMs systematically attempt to reformulate the original text, even when no modification is ultimately necessary. When such changes are appropriate, the clarification is kept; otherwise, it is corrected. This observation is supported by length statistics. Argumentative units with modified clarifications are shorter on average (143 characters) than those with unmodified clarifications (203 characters). However, the LLM clarifications have a similar average length in both cases (178 characters). The final clarifications, after human correction, are significantly shorter, with an average length of 120 characters, reinforcing the hypothesis that human corrections primarily involve removing unnecessary content.

ROUGE scores ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores measure textual similarity by comparing overlapping units such as n-grams and longest common subsequences, typically between an automatically generated summary and a human reference summary.

	GPT-4.1	qwen3-32b	llama3.3-70b	llama3.1-8b	Mean
AU → LLM	125	135	134	119	128
AU → Clarif.	81	88	93	85	87
LLM → Clarif.	78	81	71	74	76
AU → LLM=Clarif.	134	166	139	141	145

Table 4: edit distance between the original argumentative unit (UA), the LLM clarification (LLM), and the final clarification (Clarif.). The last row corresponds to the cases for which the LLM output was accepted as-is by the annotator.

ROUGE scores range from 0 to 1, with higher values indicating greater similarity.

We report ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), and ROUGE-L (longest common subsequence) scores. The results are shown in Table 5. Once again, the final clarifications appear more similar to the original argumentative units than the LLM clarifications.

For argumentative units with unmodified clarifications, ROUGE scores between the clarification and the original unit (last row of Table 5) are close to the LLM-original scores observed above, which is consistent with our previous findings.

Perplexity Perplexity measures the uncertainty of a language model when predicting the next token in a sequence. We computed the perplexity of four LLMs (`gemma-2-9b-it`, `Meta-Llama-3.1-8B-Instruct`, `Mistral-7B-Instruct-v0.3`, and `Qwen2.5-7B-Instruct`) on the original argumentative units, the LLM clarifications, and the final clarifications. Results are shown in Table 6.

Across all models, the lowest perplexity is observed for the LLM clarifications, which is expected since these texts are generated by LLMs.

B.3 Clarification Errors

Table 8 provides an example of LLM output and annotator correction for each error type, in French and with their English translations. The information removed by the annotator is displayed in red while the information added by the annotator is shown in green. Table 7 displays the proportion of types of errors for each model and in total found in the manually overviewed corrections.

C Experiments

In this section, we give details and complementary results of the experiments described in Section 6.

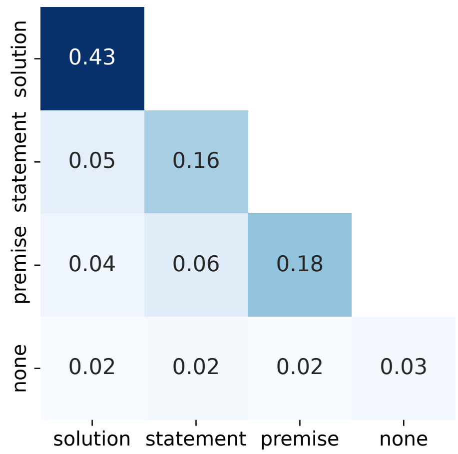


Figure 8: Proportions of tokens labels pairs in all contributions annotated by two different annotators.

C.1 Finetuning details

All models are trained for up to 5 epochs with a batch size of 2 and 8 steps of gradient accumulation. We report the best results for all model trying learning rates in $[1e-5, 3e-5, 5e-5, 1e-4]$. We set the learning rate warm-up ratio at 5%. All models have an early stopping callback based on the evaluation set loss, and most stopped training after 2 to 3 epochs for all tasks. We use the TRL package (von Werra et al., 2020) for the supervised finetuning. All models were finetuned on one H100 GPU and evaluated using one A100 GPU, and the full finetuning and evaluation pipeline took a couple of hours per model, which totals around 100 hours of GPU runtime for all experiments. We use OpenAI's `GPT-4.1` as a comparison point, prompted in a one-shot manner. We report all the prompts used for all tasks in appendix H. Figures 11, 12 and 13 display the results of base, instruct and finetuned SLMs on the three tasks of AU extraction, AS detection and AU clarification compared to the `GPT-4.1` baseline.

The selected models for each task are `Qwen-7B` ($lr = 5e-5$) for AU extraction, `Gemma-9B` ($lr =$

	ROUGE-1	ROUGE-2	ROUGE-L
LLM ↔ Clarif.	0.68	0.63	0.67
LLM ↔ AU	0.53	0.40	0.46
Clarif. ↔ AU	0.62	0.48	0.56
LLM=Clarif. ↔ AU	0.57	0.42	0.47

Table 5: ROUGE scores between the LLM clarification (LLM), the final clarification (Clarif.), and the original argumentative unit (AU). The last row corresponds to the cases for which the LLM output was accepted as-is by the annotator.

	gemma-9b	Llama-8B	Mistral-7B	Qwen2.5-7B
AU	96.5	10.0	7.5	10.3
LLM clarification	41.9	4.6	4.0	4.9
final clarification	30.0	5.9	4.7	6.1

Table 6: Perplexity of the original argumentative unit, the LLM clarification, and the final clarification. We use the Instruct variant for all models.

	GPT-4.1	Qwen-32b	Llama-70b	Llama-8b	Total
Over-analysis	76%	72%	80%	60%	72%
Miscomprehension	8%	16%	0%	24%	12%
Over-specificity	4%	4%	4%	0%	3%
Misformulation	12%	8%	16%	16%	13%

Table 7: Proportions of types of error per theme for each AI system during the second annotation phase.

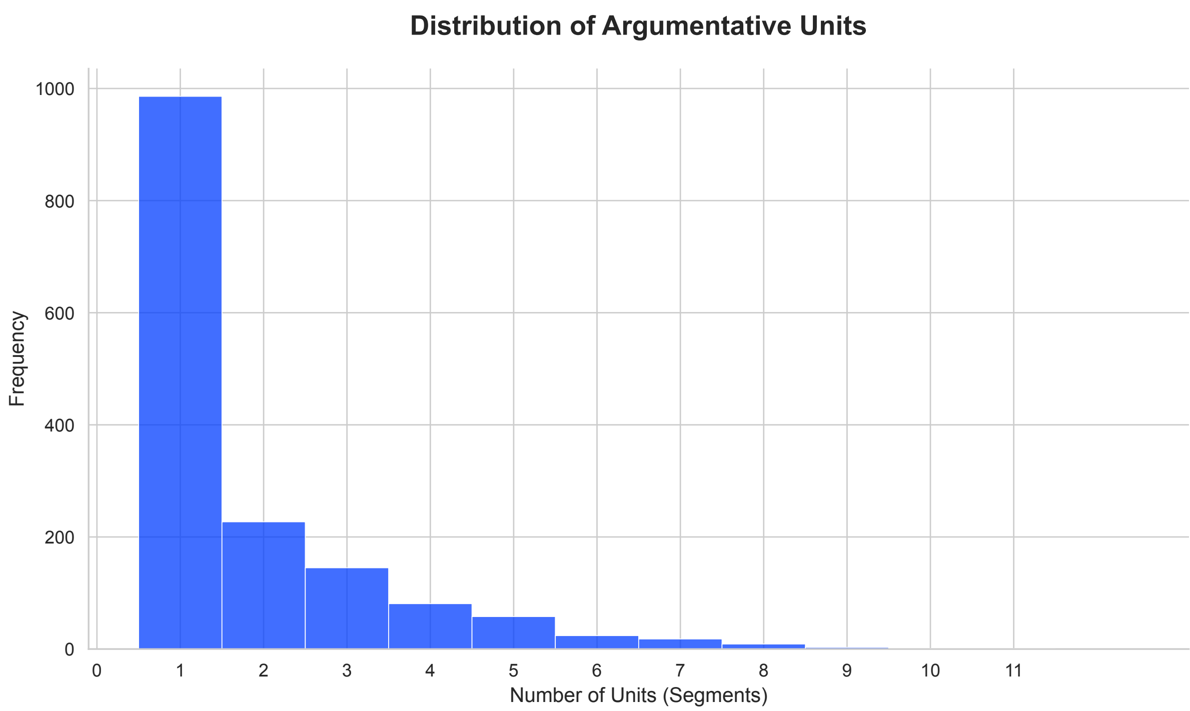


Figure 9: Distribution of numbers of Argumentative Units in the manually-annotated corpus.

Error type	Output	Correction
Over-analysis	Réduire les frais de fonctionnement, les avantages et le nombre d'élus permettrait de faire des économies significatives dans les dépenses publiques. *Reducing operating costs, benefits, and the number of elected officials would lead to significant savings in public spending.*	Il faut réduire les frais de fonctionnement, les avantages et le nombre d'élus. *Operating costs, benefits, and the number of elected officials must be reduced.*
Miscomprehension	Développer un navigateur, un moteur de recherche et une messagerie open source pour concurrencer les géants de l'internet actuels. *Develop an open source browser, search engine, and email service to compete with the current internet giants.*	Développer un navigateur, un moteur de recherche et une messagerie open source pour concurrencer les géants de l'internet actuels, en mobilisant toutes les universités européennes sur ce projet. *Develop an open source browser, search engine, and email service to compete with the current internet giants, by mobilizing all European universities on this project.*
Over-specificity	La solution est de respecter le vote des citoyens et le référendum, sans plan B, plutôt que de supprimer une ville entière pour des zadistes. *The solution is to respect the citizens' vote and the referendum, without a plan B, rather than removing an entire city for the sake of protesters.*	La solution est de respecter le vote des citoyens et le référendum, sans plan B. *The solution is to respect the citizens' vote and the referendum, without a plan B.*
Misformulation	L'argument clair et auto-suffisant sous-jacent est : **Les plus pollueurs devraient payer plus, selon un système proportionnel à leur degré de pollution.** *The clear and self-sufficient underlying argument is: **The biggest polluters should pay more, according to a system proportional to their degree of pollution.***	Les plus pollueurs devraient payer plus, selon un système proportionnel à leur degré de pollution. *The biggest polluters should pay more, according to a system proportional to their degree of pollution.*

Table 8: Example of AI system output and correction by the annotators for each type of error. In over-analysis, the removed text was not present in the initial contribution. In Over-specificity, it was present in the contribution but not in the argumentative unit.

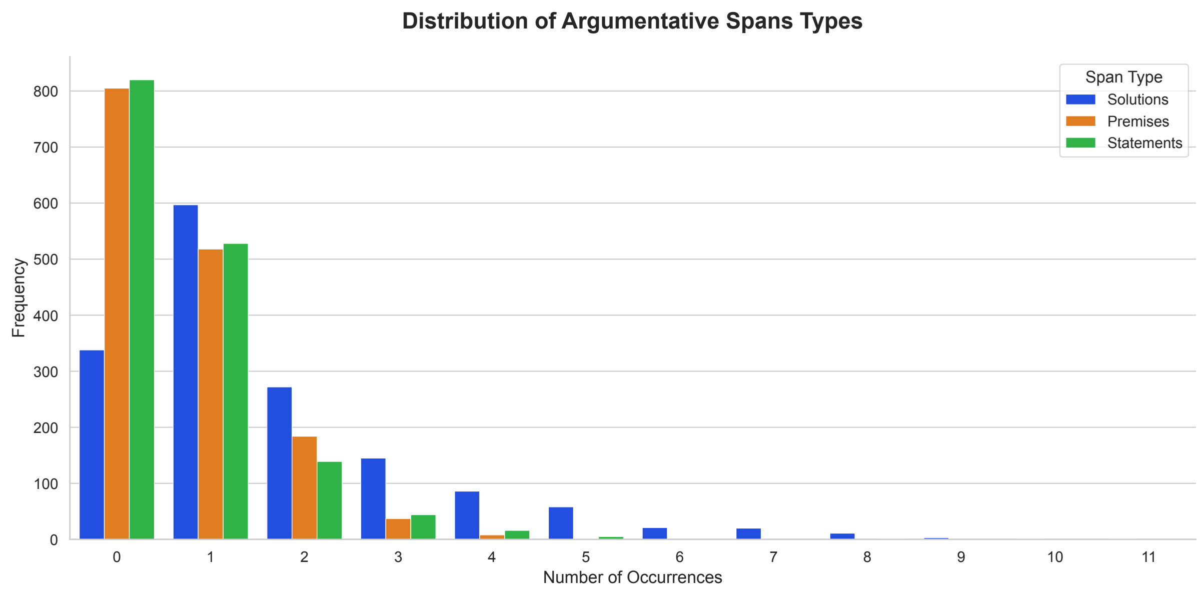


Figure 10: Distribution of numbers of argumentative spans types in the manually-annotated corpus.

$5e-5$) for AS Detection and `Gemma-9B` ($lr = 3e-5$) for AU Clarification.

C.2 Encoder-based Argument Structure Detection

As the *Argumentative Structure Detection* task can be seen as a span extraction and tagging task, we explore the capacities of a simple encoder-based model for this task, which is faster and more memory-efficient to train and use. This type of models could be useful in resource-constrained scenarios, and prove to be reliable for this task.

Model Architecture The segmentation model is built on top of a pretrained transformer encoder. Given an input sequence of tokens $(t_1, \dots, t_n)$, the encoder produces contextualized representations $(\mathbf{h}_1, \dots, \mathbf{h}_n)$.

For each candidate segment (i, j) with $1 \leq i \leq j \leq n$, we construct a segment representation by concatenating:

- the representation of the start token $\mathbf{h}_i$,
- the representation of the end token $\mathbf{h}_j$,
- the average representation of tokens within the span $\frac{1}{j-i+1} \sum_{k=i}^{j} \mathbf{h}_k$,
- the representation of the special `[CLS]` token $\mathbf{h}_0$.

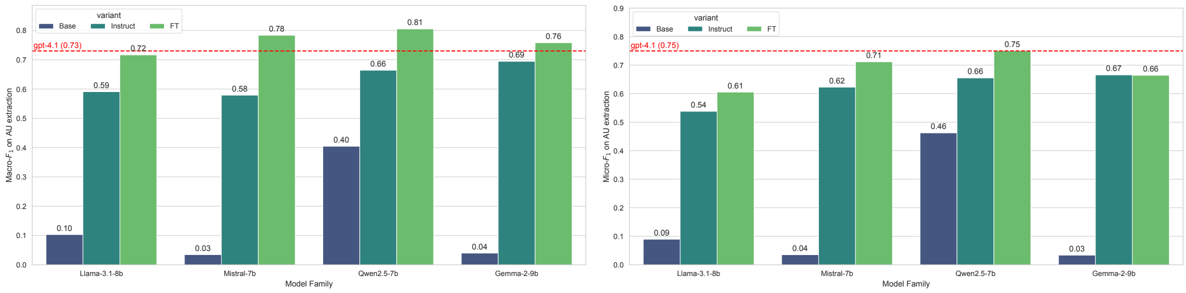
This results in a fixed-size representation for each candidate span, independently of its length.

Training Objective The model is trained to jointly predict the end position and the type of each gold span. For each gold segment (s_k, e_k, τ_k) in the training set, the start token is fixed to its gold value s_k (the beginning of the gold span), and we enumerate all candidate ends, thereby considering all possible spans up to a maximum span length, compute their representations and their scores using a feedforward scoring head. We use a cross-entropy loss over candidate ends with the gold end token e_k as reference. Similarly, a cross-entropy loss is performed over types using the gold type τ_k. The total loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{end}} + \lambda \cdot \mathcal{L}_{\text{type}}$$

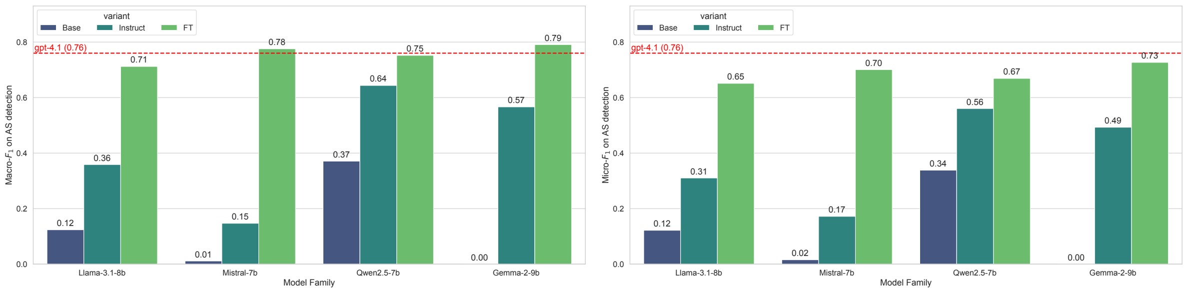
where $\mathcal{L}_{\text{end}}$ is the average loss for end prediction, $\mathcal{L}_{\text{type}}$ is the average loss for type classification, and λ is a tunable weight (`type_loss_weight`, set via hyperparameter optimization).

Inference Procedure Since each Argumentative Unit (AU) is assumed to contain only relevant argumentative material (as opposed to full-text detection), we adopt a greedy decoding strategy that sequentially consumes the entire sequence without leaving gaps. This design reflects the annotation convention of the corpus, in which AUs are extracted first and every token inside an AU belongs to exactly one argumentative span. Accordingly, we perform inference using a sequential greedy strategy to extract text spans and assign each one an argument type. Let $\mathbf{x} =$ `[CLS]`$, x_1, x_2, \dots, x_T,$ `[SEP]` be the tokenized input sequence, and $\mathbf{H} = [\mathbf{h}_0, \mathbf{h}_1, \dots, \mathbf{h}_T]$ be the



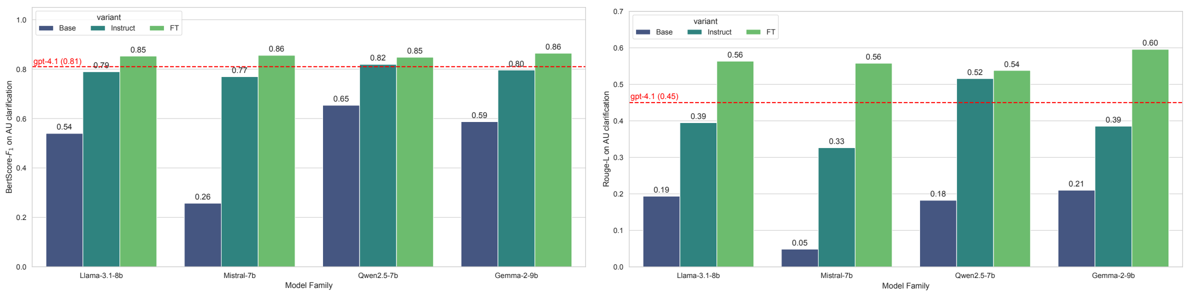
(a) Macro-F1 (b) Micro-F1

Figure 11: Macro and Micro F1 of different model families on the task of Argumentative Units extraction



(a) Macro-F1 (b) Micro-F1

Figure 12: Macro and Micro F1 of different model families on the task of Argumentative Structure detection



(a) BERTScore (b) ROUGE-L

Figure 13: BERTScore and ROUGE-L of different model families on the task of Argumentative Unit clarification

corresponding contextual representations obtained from the encoder, where $\mathbf{h}_0$ corresponds to the `[CLS]` token. Given a start token index s, we enumerate all possible spans (s, e) up to the maximum possible length (or end of the whole sequence) and compute their scores for span and type as:

$$(\text{score}_{s,e}, \tau_{s,e}) = f_{\text{span}}(\mathbf{h}_s, \mathbf{h}_e, \mathbf{h}_{\text{mean}}, \mathbf{h}_0)$$

where f_{span} is a span scoring function and $\mathbf{h}_{\text{mean}}$ is the mean-pooled representation of all token hidden states from index s to e. The score $\text{score}_{s,e}$ assesses the likelihood of the span being argumentative, and $\tau_{s,e}$ denotes its associated type. We select the best end point based on the span score:

$$(e^*, \tau_{s,e^*}) = \arg\max_{e} \ \text{score}_{s,e}$$

Once an end token is predicted, we take the next token to be the start of the next span and repeat the procedure until the whole input sequence is consumed. This design choice was made bearing in mind that spans in each Argumentative Unit (AU) form a contiguous, non-overlapping segmentation that exhaustively covers the text. Consequently, during training the model is only supervised on end positions and types given a gold start, and during inference start positions are determined deterministically by the greedy segmentation process.

Hyperparameter Tuning We optimize all hyperparameters with Optuna (Akiba et al., 2019). The search space includes the choice of encoder (`camembertav2-base` (Antoun et al., 2024) or `moderncamembert-base` (Antoun et al., 2025)), learning rate (log-uniform), dropout rate, maximum span length, and the weighting of the type-classification loss. We use Optuna's Median Pruner to discard low-performing configurations early based on development micro-F1. For each trial, models are trained with a batch size of 4 examples and using early stopping, for up to a maximum of 10 epochs. The final selected configuration yielded a micro-F1 score of 0.704 over the development set after 4 epochs, and is shown in Table 9. Its results are displayed in Table 10.

Encoder Performances Micro- and macro-metrics similar to those presented for decoder-based approaches are reported in Table 10.

D Statistical Model for LLMs clarification qualities

In this appendix, we elaborate the statistical model introduced in Section 5.

Hyperparameter	Value
Model	`almanach/camembertav2-base`
Learning rate	5.379×10^{-5}
Dropout	0.202
Type loss weight	0.878
Max span length	261

Table 9: Best hyperparameter configuration selected via Optuna.

Metric	Precision	Recall	F1
Macro	0.769	0.779	0.765
Micro	0.697	0.711	0.704

Table 10: Macro and Micro F1-scores on the Test Set for Encoder-based Argument Structure Detection ($\lambda_{\text{overlap}} = 0.5$)

We formalize the AI-human hybrid annotation process as a decision model to evaluate the intrinsic quality of the four clarification systems. Following the argumentative segmentation, an annotator requests an automated clarification from a model l, where l is sampled from a discrete uniform distribution $L \sim \mathcal{U}\{1, \dots, 4\}$. The annotator then makes a binary decision based on an internal quality threshold τ_k, which varies depending on the number of attempts k. If the generated clarification quality e meets or exceeds the threshold ($e \geq \tau_k$), the annotator accepts the output. In this case, we directly observe the quality e as the ROUGE-L score between the model's output and the final human-validated text. Else, if the quality falls below the threshold ($e < \tau_k$), the annotator rejects the output and requests a new generation, leaving the specific value of e unknown. We model the quality e for each model l using a Beta distribution, $e \sim Beta(\alpha_l, \beta_l)$. Observations are $\mathcal{D} = \{(k, l, r, e)_j\}_{j \in data}$ where $r \in \{0, 1\}$ is the acceptance indicator and τ_k is modeled as a Dirac distribution dependent on the attempt index k. The total log-likelihood for our observations $\mathcal{D}$ is defined as follows:

$$\begin{aligned}
\log \mathcal{L} &= \sum \log p(r_j, e_j | k_j, l_j) \\
&= \sum_{j|r=0} \log p(e_j | k_j, l_j) \cdot p(e > \tau_{k_j} | k_j, l_j) \\
&+ \sum_{j|r=1} \log p(e < \tau_{k_j} | k_j, l_j)
\end{aligned} \tag{1}$$

	α	β
`GPT-4.1`	1.75	1.12e-1
`Llama-70B`	1.80	1.22e-1
`Qwen-32B`	1.57	1.19e-1
`Llama-8B`	1.10	1.16e-1

Table 11: α_l and β_l for all four models.

We jointly optimize the parameters α_l and β_l and the acceptance thresholds τ_k via gradient ascent to find the maximum likelihood estimates for each system.

As the Beta distribution's cumulative distribution function (defined as the regularized incomplete beta function $I_x(a,b) = \frac{\mathrm{B}(x;a,b)}{\mathrm{B}(a,b)}$) does not have a closed form and its partial derivatives are not tractable, we approximate its value through numerical integration and compute the gradients by implementing it with torch.

The final α_l and β_l after optimization are reported in Table 11. The associated means computed as $\frac{\alpha_l}{\alpha_l+\beta_l}$ are reported in Section 4.

E Clarification Improvements after finetuning

Figure 14 shows the prompt used for LLM-as-a-Judge for the clarification evaluation. Over the 348 examples of the test set, the preferred clarification was the one of the LLM during the annotation 91 of the time, the one of the finetuned model 231 times, and a draw 26 times.

We use the statistical test χ^2, where H_0 is the fact that all three possible outcomes are equally likely, and the binomial statistical tests by eliminating the draw option and considering H_0 as A and B are equally likely. For both tests, the null hypothesis H_0 can be rejected ($p < 0.001$).

Interestingly, by doing the same evaluation by comparing the finetuned SLMs and annotators' clarifications, the finetuned SLMs are preferred in 194 examples (111 for annotators, 43 draws). this result is also statistically significant ($p < 0.001$). We find that the finetuned SLM tends to rephrase the argumentative unit texts less than the LLMs, and therefore of annotators clarifications (which are usually modifications of LLMs output, as highlighted in appendix B). This makes their clarifications closer to the original meaning of the AUs while correcting their writing mistakes and unnatural phrasings. Table 12 provides examples of SLMs output that avoid the over-analysis of the LLM model.

F Clustering Improvements

Technical details We use the `paraphrase-multilingual-MiniLM-L12-v2` sentence transformer model (Reimers and Gurevych, 2019) as embedding model, UMAP (Sainburg et al., 2021) as a dimension reduction algorithm and HDBSCAN (Campello et al., 2013) as the clustering method.

Experiments and results Table 13 displays the results of the LLM-as-a-Judge. the lines Tax., Eco., Org. and Dem. respectively correspond to the themes *Taxation and Public Spending*, *Ecological Transition*, *Organization of The State* and *Democracy and Citizenship*. The three columns are the three experimental settings:

- **Setting 1**: Initial contributions (A) vs. Argumentative units (B)
- **Setting 2**: Argumentative units (A) vs. Clarifications (B)
- **Setting 3**: Argumentative units (A) vs. AU using clarification-based clusters (B)

in all three settings, D means *Draw*. We evaluate the χ^2 and binomial statistical tests with the same null hypothesis H_0 as done in Appendix E. For both tests, in all settings, and for all themes the null hypothesis H_0 can be rejected ($p < 0.001$).

	Setting 1			**Setting 2**			**Setting 3**		
	A	B	D	A	B	D	A	B	D
Tax	15	81	4	6	92	2	13	87	0
Eco	22	77	1	14	84	2	8	91	1
Org	7	90	3	6	90	4	1	95	4
Dem	10	88	2	1	98	1	4	88	8
Tot.	**54**	**336**	**10**	**27**	**364**	**9**	**26**	**361**	**13**

Table 13: LLM-as-a-Judge preference distribution across four themes and three experimental settings.

Because the text type (segmented AUs) evaluated by the LLM-Judge was identical in Setting 3, the 90% preference confirms that the improvement stems entirely from better cluster organization, not just linguistic clarity.

G GDN-CC-large

Table 14 displays the corpus statistics per theme and in total. This annotation reinforce the analysis made in Section 5, showing that *Democracy and Citizenship* contains proportionally fewer solutions

Initial AU	**LLM output**	**Finetuned SLM output**
Mettre en place des centres de collecte (vêtements matériels puériculture jouets) pour les enfants malades ou placés. Idem pour les animaux. Idem pour le recyclage. *Set up collection centers (clothing, childcare items, toys) for sick or foster children. The same applies to animals. The same applies to recycling.*	Mettre en place des centres de collecte pour vêtements, matériel de puériculture et jouets destinés aux enfants malades ou placés, ainsi que pour les animaux et le recyclage, permettrait d'apporter une aide concrète et ciblée en complément de la fiscalité classique. *Setting up collection centers for clothing, childcare equipment, and toys for sick or foster children, as well as for animals and recycling, would provide concrete and targeted assistance in addition to traditional taxation.*	Il faut mettre en place des centres de collecte pour les enfants malades ou placés, les animaux et le recyclage. *Collection centers must be set up for sick or foster children, animals, and recycling.*
Taxer les propriétaires des grosses cylindrées. *Tax owners of large-engine vehicles.*	Il faut taxer les propriétaires des grosses cylindrées pour encourager des choix de véhicules moins polluants.' *Owners of large-engine vehicles should be taxed to encourage the choice of less polluting vehicles.*	Taxer les propriétaires des grosses cylindrées. *Tax owners of large-engine vehicles.*
Subventions état pour stérilisation chats errants . *Government subsidies for sterilization stray cats .*	L'État devrait accorder des subventions pour la stérilisation des chats errants afin de mieux protéger les animaux. *The government should provide subsidies for the sterilization of stray cats in order to better protect animals.*	Il faut accorder des subventions publiques pour la stérilisation des chats errants. *Public subsidies should be granted for the sterilization of stray cats.*

Table 12: Some example clarifications from the test set given by the LLMs during the annotation and by the finetuned SLM after finetuning.

and more statements. We can also see that *Ecological Transition* and *Taxation and Public Spending* got significantly more contributions (72.4% of all contributions and 73.6% of all argumentative units in the corpus) than the two other themes. These themes are strongly related to the context in which the *Grand Débat National* was organized, following the *Yellow Vest* movement which started due to a raise in taxes.

H Annotation and Finetuning Prompts

We display in this appendix the prompts used during the annotation (Figure 15), for the Argumentative Unit extraction task (Figure 16), the Argumentative Structure detection task (Figure 17) and the Argumentative Unit extraction task (Figure 18). Figure 19 displays the LLM-as-a-Judge prompt used for the evaluation of clustering.

All prompts were given in French to AI models in order to limit code switching, but are here given in English. For `GPT-4.1` prompting as a comparison baseline, we added one example in each prompt before giving the actual contribution to process.

Theme	Contribs.	AUs	statements	solutions	premises
Taxation and Public Spending	95,689	120,902	56,709(28.2%)	118,399(59.0%)	25,666(12.8%)
Ecological Transition	77,944	100,442	50,872(31.3%)	92,587(56.9%)	19,182(11.8%)
Organization of the State	28,584	35,204	19,948(32.6%)	32,826(53.6%)	8,418(13.8%)
Democracy and Citizenship	37,581	44,200	28,217(37.7%)	37,854(50.6%)	8,771(11.7%)
Total	**239,798**	**300,748**	**155,746**(31.2%)	**281,666**(56.4%)	**62,037**(12.4%)

Table 14: Number of contributions, argumentative units, statements, solutions, and premises per theme and in total in the automatically-annotated corpus.

```
### Role
You are an expert in rewriting and clarification. Your task is to judge the
clarity of a given text.

### Strict instructions
The user will give you a text, a segment of that text (which may
potentially be the entire text), and two clarifications, A and B, of that segment.
A clarification must transform the initial text into a clear and self-sufficient
text that can be understood without the context of the initial text.
This clarification may add this context if it is contained in the initial text
and rephrase the segment. However, it must not add
justifications if the text does not mention them. You must judge which of the two
clarifications is the best.
Answer ONLY with “A,” “B,” or “TIE.” Answer “TIE” only if there is
no difference between the two. Prefer “A” or “B.”

### Reference examples:
Example 1 (adding justification):
Text:
I don't understand how anyone can be so out of touch with reality. THE PRESIDENT'S
SALARY MUST BE REDUCED!!
Segment:
THE PRESIDENT'S SALARY MUST BE REDUCED!!
Clarifications:
A- The president's salary must be lowered.
B- The president's salary must be lowered to limit public spending.
Answer: A

Example 2 (one of the two adds important context):
Text:
Inequalities are too great; aid must be increased. The same goes for tax audits.
Segment:
Segment:
The same applies to tax control.
Clarification:
A- Tax control must be strengthened.
B- Tax control must be strengthened because inequality is too high.
Answer: B

Example 3 (equality):
Text:
Lower the allowances of parliamentarians.
Segment:
Lower the allowances of parliamentarians.
Clarification:
- The allowances of parliamentarians must be lowered.
- Lower the allowances of parliamentarians.
Answer: EQUALITY
```

Figure 14: LLM-as-a-judge prompt for clarification evaluation (translated from French to English).

```
You are an argument clarification portal. The user will give you a written opinion
on a given topic, as well as the segmentation of one of the arguments in that
opinion into three types of segments: statement(s), argument(s), and solution(s).
Extract, in one sentence, the clear and self-sufficient argument underlying this
segmentation. Prioritize the solution, and include arguments and statements
only if they seem relevant to you. You can use the context of the entire opinion
to help you, but do not include any information that is not present
in the segments. Respond only with the clear and self-sufficient argument, and
nothing  else. If the argument is already clear and well written, you can refer
directly to that argument.

Given the opinion:
{{contribution}}

on the topic {{theme}}

Extract, in one sentence, the underlying argument consisting of:
- Statements: {{statements}}
- Arguments: {{premises}}
- Solutions: {{solutions}}
```

Figure 15: System Prompt for the clarification task used during the annotation (translated from French to English).

```
I am going to give you an opinion piece in French. Your task is to segment this
text into argumentative units. We define an argumentative unit as one
or more segments of the text that focus on a particular topic. It may consist of
solutions, arguments, or simple statements. An argumentative unit
is not necessarily contiguous: it can join segments that do not follow each other.

This task is EXTRACTIVE. You must COPY and only copy the text of the argumentative
unit exactly as it is written, including capital letters and punctuation.
If the argumentative unit is composed of several non-contiguous segments,
 you can concatenate them by simply separating them with a space. There is at
least one argumentative unit in the text, but no maximum number. Highlight the
argumentative units in the form of a list as shown in the example.
Not all segments of the text are necessarily part of an argumentative unit.

You must give the argumentative units in the form of a list:
- argumentative unit 1
- argumentative unit 2...

Do NOT output ANYTHING OTHER than the list of argumentative units.
Here is the text:
{{contribution}}
```

Figure 16: System Prompt for the Argumentative Units extraction tasks (translated from French to English).

```
I am going to give you a segment of text containing opinions in French. Your task is
to segment this text and assign each segment a type. The possible types are
STATEMENT, PREMISE, and SOLUTION, and ONLY those. Below is the definition of
each type:
- SOLUTION: a proposal for action (whether concrete and feasible or not) to be
taken to solve a problem.
- STATEMENT: the expression of an opinion as an assertion, which does not provide a
solution but rather expresses a feeling.
- PREMISE: a justification, argument, or example that supports an
assertion or solution.

This task is EXTRACTIVE; you must copy the text of each segment exactly
as it is written, including capital letters and punctuation.
The entire text must be segmented. Not all types of
segments are necessarily present, and several segments may be of the same type.
You MUST highlight the segmentation by following the exact format of the example,
including the “-” for each segment.

- [STATEMENT] Statement 1
- [SOLUTION] Solution 1
- [STATEMENT] Statement 2
- [PREMISE] Argument 1...

I will give you the original text and the segment, and you must output the list of
segments and their types in the form “- [TYPE] SEGMENT,” and nothing else.

Original text:
{{contribution}}

Segment:
{{argumentative unit}}
```

Figure 17: System Prompt for the Argumentative Structure detection task (translated from French to English).

```
I will give you a segment of opinion text in French, as well as the original text
from which it is taken. Your job is to rewrite this opinion segment
clearly. In particular, you must first correct spelling,
grammar, and syntax errors. You must also add the context present in the original
text  if it is important for understanding the segment. The final
clarified text must be understandable without access to either the original text or
its sub-segment. If the segment is already clear and well written, you should
simply copy it. You should ONLY highlight the clarification of the segment and
nothing else.

Original text:
{{contribution}}

Segment to be clarified:
{{argumentative unit}}
```

Figure 18: System Prompt for the Argumentative Unit clarification task (translated from French to English.

```
### Role
You are an expert judge of thematic consistency. Your task is to compare two
clusters of texts (A and B).

### Strict instructions
1. A group is “better” if it is more specific, more precise, and if ALL
texts deal with the same subject using the same approach.
2. If a group contains different subjects (even if they are vaguely related to
politics or money), it must be penalized.
3. Respond ONLY with “A,” “B,” or “TIE.” Respond with “TIE” only if there is
no difference between the two. Give preference to “A” or “B.”
No explanations will be tolerated.

### Reference examples:
Example 1 (Divergent topics vs. Identical topics):
A:
- “Reform public inquiries”
- “Limit aid to foreigners”
- “Climate change”
B:
- “Tax kerosene,”
- “Tax polluting companies”
Verdict: B

Example 2 (Similar but different topics vs. Identical topics):
A:
- “Inheritance and estate taxes,”
- “Cost of retirement homes/nursing homes”
B:
- “Eliminate compensation for former presidents,”
- “Reduce benefits for former presidents”
Verdict: B

Example 3 (Total consistency on both sides):
A:
- “Legalize cannabis”
- “Sell cannabis in pharmacies”
B:
- “Increase the minimum wage,”
- “Raise the minimum wage”
Verdict: TIE
```

Figure 19: System Prompt for the LLM-as-a-Judge evaluation of clustering (translated from French to English).