"I Never Said That": A dataset, taxonomy and baselines on response clarity classification.

Anonymous ACL submission

Abstract

Equivocation and ambiguity in public speech is a well-studied discourse phenomenon, especially in political science for the analysis of political interviews. Inspired by the well-grounded theory on equivocation, we aim to resolve the closely related problem of response clarity in questions extracted from political interviews, leveraging the capabilities of Large Language Models (LLMs) and human expertise. To this end, we introduce a *novel taxonomy* that frames the task of detecting and classifying response clarity and a *clarity classification dataset* which consists of question-answer pairs drawn from political interviews and annotated accordingly. Our proposed two-level taxonomy addresses the clarity of a response in terms of the information provided with respect to a given question (high-level), and also provides a fine-grained taxonomy of evasion techniques that relate to unclear, ambiguous responses (lower-level). Our annotation process leverages ChatGPT towards decomposing political dialogues into discrete question-answer pairs, each of which belongs to a specific response clarity and evasion category. Consequently, human annotators decide upon the correctness of this decomposition, while assigning an evasion label for each question-answer pair. We provide a detailed analysis of the dataset and we conduct several experiments using a range of LLMs to establish new baselines over the proposed dataset. [1]

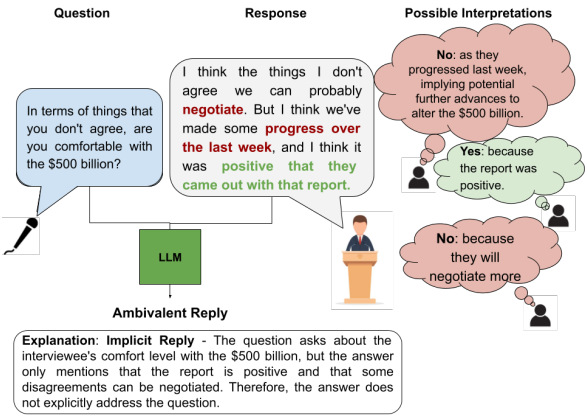

Figure 1: An example from an actual interview in our dataset.

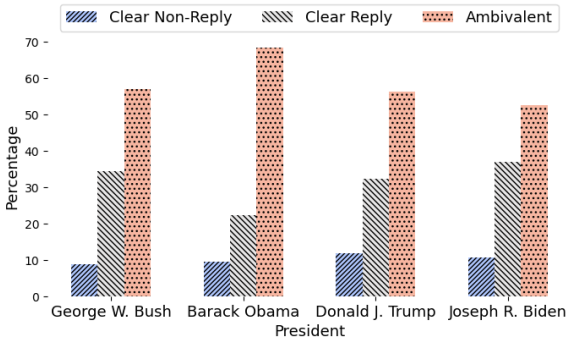

Figure 2: Statistics on answer clarity in political interviews of the latest 4 US presidents

1 Introduction

In the era of mass information dissemination, question evasion and response ambiguity are widespread phenomena in political interviews and debates, rendering their detection an important aspect of political discourse studies. Bull (2003) presents a meta-analysis of five studies on political interview question answering, concluding that politicians gave clear responses to only 39-46% of questions during televised political interviews, while, non-politicians being interviewed on television had a significantly higher, 70-89%, reply rate. In Figure 2 we present some statistics regarding US presidents' response clarity, as occurring from our human annotations, demonstrating that in the majority of cases, politicians exploit meticulously crafted techniques to avoid explicitly responding to journalists' questions. Fig. 1 presents a running example of an interview, featuring various interpre-

[1] All code and data will be made publicly available upon publication.

tations, generated labels, along with corresponding explanations using our proposed dataset.

This phenomenon is referred to as *equivocation* or *evasion* in academic literature and describes a non-straightforward type of communication which is characterised by lack of clarity and includes speech acts such as self-contradictions, inconsistencies, subject switches, incomplete sentences, misunderstandings, obscure style or mannerisms of speech, etc. (Watzlawick et al., 1964; Bavelas et al., 1988; Rasiah, 2010). More formally, it can be considered a type of adversarial attack on a question, where the response adds no information on the queried subject, yet it follows the format of a valid answer.

While the topic has been studied extensively in the field of linguistics, politics and communication, with several typologies proposed to classify responses to a given question (Harris, 1991; Bull and Mayer, 1993; Rasiah, 2010), there has been no attempt to analyse whether such typologies are applicable to a larger scale and consistent with varying human perspectives and biases. In other words, the possibility of automatically classifying the clarity of responses has not been explored in NLP, potentially because of the complexity of the task itself, as well as the underlying need to encode and reason on long context. Nevertheless, recent advancements in language modelling boosted the performance of models for long-context inputs (Dai et al., 2019; Wei et al., 2022, 2023), paving the way for framing the task of **automatically measuring the response clarity to its respective question**.

Related to this endeavour, there have been some recent answerability challenges for question-answering tasks (Min et al., 2020; BingningWang et al., 2020; Rogers et al., 2020; Sun et al., 2022; Wang et al., 2022). However, all these works focus on assessing the clarity of the questions, but not the answers. We fill this gap by proposing the task of **response clarity evaluation**, leveraging the recent advancements of LLMs.

We carry out a detailed analysis of proposed typologies, considering their overlap and consistency, the rate of occurrence of proposed classes in our collected data, and the feasibility of using them in an automated task, resulting in our proposed *two-level response clarity detection taxonomy*. Specifically, the first level of the taxonomy addresses the three-scale evaluation of response *clarity* in terms of the number of interpretations the intended response holds. The second and more fine-grained level refers to common *evasion* phenomena found in political literature, which explain in more detail the categorization of responses in the three-scale clarity categories. We use this taxonomy to annotate a dataset of political questions and answer pairs and carry out an analysis of the variability of perspectives among human annotators. We then evaluate different LLMs, exploring different training and inference frameworks, showing that simple prompting and instruction-tuning techniques using our dataset are highly capable of providing meaningful performance. Moreover, we find that using the labels of the second level (evasion categories) in a two-step classification strategy helps boost performance for clarity classification.

We argue that being able to detect answer ambiguity automatically will facilitate political speech discourse analysis, allowing for comparisons at scale. Additionally, the proposed task can shed light on LLM capabilities of reasoning over long contexts and prove useful for other downstream tasks in NLP such as question answering (see also Section 2.1).

To sum up, our contributions are threefold:

- We propose a new task, response clarity evaluation, which aims to detect the alignment and clarity of a given response with respect to its respective question and provide an empirically and theoretically established taxonomy for it.
- We introduce a human-labelled dataset on the aforementioned task, comprising 3,448 question-answer pairs from political interviews. We make this resource public to facilitate further research on the task.
- We experiment with several LLMs to establish baselines for the proposed task.

2 Related work

2.1 Equivocation in Social Sciences

Political equivocation, generalised by Dillon (1990) as "the routine strategy for responding to a question without answering it", provides a range of proposed frameworks to analyse evasion in responses (Wilson, 1990; Bull, 2009; Bull and Strawson, 2019). Harris (1991) made a distinction between direct and indirect answers, specifying that indirect answers contain the same information as direct answers but this information is provided implicitly and not directly. Beyond direct and indirect answers there are cases where partial, or no useful information is

provided in the response, and several works have attempted to categorise responses along this dimension (Bull, 1994, 2003) For instance, Wilson, Harris and Bull provide criteria for the identification of three main categories (Bull and Mayer, 1993). ① *Replies* correspond to cases where the requested information is given in full. ② *Non-Replies*, which are considered cases where none of the information requested is given in a clear manner (Rasiah, 2010); non-Replies are broken down into twelve further *evasion* sub-categories, presented in Table 1. Finally, ③ *intermediate replies* refer to those utterances which, for a variety of reasons, fall somewhere between replies and non-replies; i.e. responding completely to one part of a multi-part question but ignoring the rest of the query; responding only in part to a single-part question; answering a question through suggestion or implication but not giving a straightforward answer.

1. **Ignores the question.** Makes no attempt to answer the question, or even to acknowledge that a question has been asked.
2. **Acknowledges the question.** Acknowledges that a question has been asked, but equivocates.
3. **Questions the question.** Requests clarification, or reflects the question back to the questioner.
4. **Attacks the question.**
5. **Personalisation.** Makes personal comments, typically in the form of personal attacks.
6. **Declines to answer.**
7. **Makes political points.**
8. **Gives incomplete reply.**
9. **Repeats answer to the previous question.**
10. **States or implies has already answered the question.**
11. **Apologises.**
12. **Literalism.** The literal aspect of a question which was not intended to be taken literally is answered.

Table 1: Equivocation typology proposed by Bull and Strawson (2019)

Bull (2003) breaks the 12 evasion techniques further, into 28 more fine-grained micro-categories. For example *Makes political point* includes the micro-categories of "*External attacks on the opposition or other rival groups*", "*Talks up one's own side*", "*Presents policy*", etc. Rasiah (2010) breaks the replies, which he terms "answers", further in Direct and Indirect answers. He keeps the rest of the Intermediate Responses in one category and also breaks down Non-replies (which he labels "Evasions") into four degrees of evasiveness, as well as whether the evasion was overt or covert and what types of 'agenda shifts' occurred.

Thus, to adapt these typologies to a response clarity taxonomy that can be used for a dataset suitable for NLP, it was necessary to modify it taking into consideration the following factors:

- Our focus is slightly different: on a taxonomy that classifies the clarity of responses (hence an indirect response falls under a different category than a direct one).
- We want to have a good representation of each category in our dataset to allow computational modelling using LLMs. It is thus necessary to condense the categories to avoid overly sparse categorisation while retaining the essential characteristics of each category (i.e., we provide meaningful labels).
- Labelling of the responses is conducted from non-expert human annotators, so that our annotations reflect the views of a larger portion of the population rather than a minority of experts. The difficulty of the classification, and thus the resulting error rate, increases as we increase the label set they choose from.
- Most interviewers pose multi-barrelled questions, leading to a situation where multiple QA pairs are labelled under a single label. We need to break the multi-part questions into singular ones to retain this fine-grained information.

Section 3 discusses the taxonomy we adopted, aiming to optimise for the annotation task, as well as for the selected LLMs.

2.2 Equivocation in NLP

While *equivocation* has not been explicitly studied in NLP, there are still some relevant areas of work, concerning mostly answerability of questions and automated discourse analysis in politics.

2.2.1 Answerability in question answering

There have been several tasks proposed related to question answering (QA) both in open-ended and closed set answer setups. The issue of the *answerability* of a given question an in QA was highlighted in SQuAD 2.0 (Rajpurkar et al., 2018), which introduced adversarially crafted unanswerable questions with respect to a given text span. Lee et al. (2020) expanded the SQuAD 2.0 dataset, also incorporating the rationale for unanswerable

questions. Extending to out-of-domain questions to address practical use cases, Sulem et al. (2021) introduce competitive and non-competitive unanswerable questions. Relevant endeavours question the answerability of information-seeking queries built independently of the passage containing possible answers to those queries (Asai and Choi, 2020). Scalability issues are addressed via synthetic extensions of existing datasets containing both answerable and unanswerable questions (Nikolenko and Kalehbasti, 2020). To the same end, other works develop data augmentation techniques to produce unanswerable queries based on answerable SQuAD 2.0 queries (Zhu et al., 2019; Du et al., 2022). Other datasets targeting answerability issues are ReCO (BingningWang et al., 2020), which provides "yes", "maybe" and "no" labels for questions paired with passages in Chinese, as well as QuAIL (Rogers et al., 2020), which introduces questions of varying certainty according to the accompanying passage.

While the intuition and motivation behind our proposed dataset differs, it also relates to the topic of answerability of a question with respect to a text span: if we consider the provided response as the context, then an incomplete response would in fact result to an unanswerable question. However, our primary target is to annotate responses for clarity with respect to a given question, rather than evaluating question clarity, leading to a different task and reasoning process.

2.2.2 Discourse analysis of political speech

Beyond evasion, discourse phenomena in political speech (including responses in interviews) have been analysed in several NLP works. Majumder et al. (2020) construct a large-scale dataset of political dialogs to study discourse patterns, upon which they train a model that uses external knowledge. Among the analysed discourse patterns, they consider modes of persuasion, entertainment, and information elicitation (the latter being the closer to our target). Understanding political agendas requires contextualization depending on which politician expresses a certain claim, as proposed in Pujari and Goldwasser (2021) proposes the combined use of transformer-based modules to obtain better representations of political agendas based on politician tweets. Finally, non-verbal aspects of political discourse, such as the usage of gestures have been proven to be associated with individuals rather than political parties, while contributing to emphasizing certain parts of speech (Trotta and Tonelli, 2021).

3 Proposed Taxonomy on Response classification

The typologies mentioned in Section 2.1 are thorough and well-studied by experts, however, they are often incompatible with each other. E.g. Bull (1994); Bull and Strawson (2019) consider indirect answers as intermediate replies, while (Rasiah, 2010) considers them as a type of complete reply. Additionally, for some responses, the distinction between categories differs between experts and is highly dependent on the sub-domain and perspective. For example, a somewhat vague answer can be interpreted as evasive by some or an indirect but valid reply by others, depending on their personal views and biases. Such ambivalent responses are especially prone to confirmation bias (Nickerson, 1998). To increase the objectivity of the task, we direct our focus not on the Reply/Non-reply dimension, but rather on the Clarity/Ambiguity dimension. This approach removes the burden from the annotators to subjectively interpret ambiguous answers as valid or invalid, and instead pivots their attention on determining whether a response can be interpreted unambiguously or accepts a wide range of interpretations.

Apart from the aforementioned inconsistencies, the most extensive of the typologies include more than 30 types of replies (Bull, 2009), limiting the number of examples each category would contain and creating sparsity in our dataset. Moreover, the difficulty of the annotation task by non-experts increases with the number of categories. As such, we aimed to consolidate them into fewer essential categories, yet ensuring we maintain key distinctions between labels.

Another crucial adjustment was the question break-down, which also led to the elimination of the category of "intermediate replies", which was skeing the label distributions in the dataset. As mentioned in Section 2.1, most interviewers pose multi-barrelled questions, leading to a situation where vagueness in a single answer on a multipart question results in the full question-answer data point getting classified as an intermediate reply. To mitigate a heavy bias in the dataset towards intermediate replies and the loss of valuable information, we employed an automated process of breaking multi-barrelled questions into separate questions and have the annotators label each sub-question and answer separately.

Taking all of the above into consideration, we

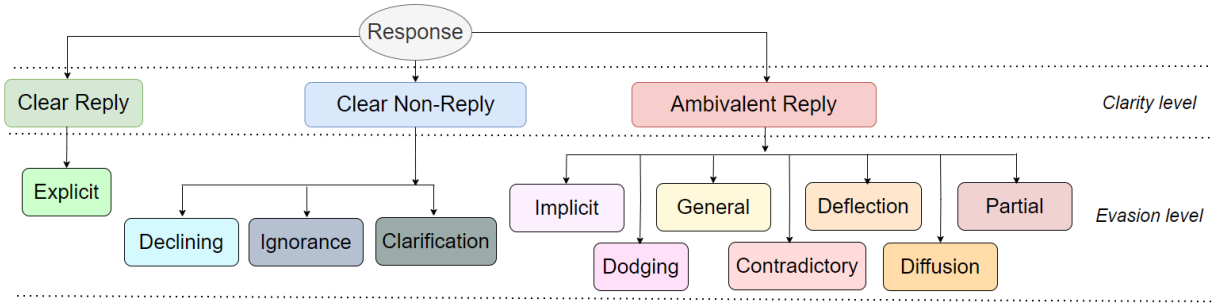

Figure 3: Our proposed taxonomy of response clarity classification.

concluded to a two-level hierarchical taxonomy. The higher level includes 3 main response categories, namely ① *Clear reply*, containing replies that can only be interpreted a single way; ② *Clear non-reply*, containing the responses where the answerer is open about sharing no information, and ③ *Ambivalent reply*, where a response is given in the form of a valid answer but in a can be interpreted in a multitude of ways. At the second level these 3 categories include 11 sub-categories illustrated in Figure 3. As a brief demonstration, "*Q: Have you seen my chocolates? A: The children were in your room this morning.*" would be considered an Implicit reply (under the Ambivalent category) since the suggestion being made is quite clear. Yet the answer does not commit to explicitly stating that the kids probably ate it - which would have made for an Explicit reply - but rather allows the questioner to make the logical step to reach this assumption. While "*A. I don't know*", to the same question, would be labelled as a Clear non-reply and specifically "Claims ignorance", since the answer is straightforward about not being able to provide information. And "*A. You should not keep your chocolates all around the house*" would be considered a "Deflection", again an Ambivalent answer, as the answer gives nothing in terms of the requested information, yet it leverages the subject to pivot on a different point. For further analysis and examples of all sub-categories see Table 6.

4 Dataset creation

As a first step, we collected presidential interviews of US Presidents as provided by the official whitehouse website [2]. This resulted in 287 unique interviews spanning from 2006 until 2023. More statistics regarding the interviews are provided in the Appendix A.1. We then extracted a total of 3,448 questions and responses from these interviews, as described in the following sections.

To prepare the question-answer pairs we leverage ChatGPT to decompose the original interviews. The decomposition focuses on separating the potentially multi-barelled question into separate points (subquestions) and their respective response subparts. We use the automatically generated list of questions to guide the generation of annotation instances in our dataset, where we separately annotate the response to each subquestion. Thus, for a given interview question, we may have several instances in the final dataset each corresponding to a distinct subquestion, and the classification of the respective subresponse. We henceforth refer to the generated subquestions and sub-responses as "summaries".

4.1 Human annotation process

Upon the aforementioned preprocessing of the interview questions, we specify the annotation task where the annotators are provided both with the original question and answer as well as the summary, and asked to label the response for each sub-question separately. We opted for providing the summaries alongside the full text to reduce the effort of manually extracting distinct question-answer pairs from the original interviews, which would significantly increase the annotation time per sample. Nevertheless, explicitly instruct (and monitor) annotators to ensure they carefully consult the original interview, so that we avoid erroneous annotations due to imperfect summaries. We further introduce counterfactual summaries to measure their potentially exclusive reliance on summaries, as explained in Sec. 4.1.1, verifying that they followed our instructions. The prompt provided to ChatGPT to create the original summaries and adversary summaries is demonstrated in Appendix B.

We employed 3 human annotators alongside an

[2]We specifically crawled presidential interviews from https://www.whitehouse.gov/.

expert with background in political science and political discourse analysis who acted as validator of the outcome annotations. Thus, we obtained annotations on 3,448 samples of question-answer summaries derived from the 287 political interviews. As a first "training" stage, we provided the annotators with a tutorial that included annotated examples from each category of the taxonomy to allow them to familiarise themselves with the concepts introduced. Then, the annotators were prompted to perform a series of annotation tasks in the following order: they had to ① evaluate the question-answer summaries produced by ChatGPT as valid or not, and then ② label each of the individual questions and answers, using the proposed taxonomy. Finally, they were asked to ③ add any missing questions, as well as their label. On average, each annotator evaluated 1150 samples (we provide more details in Appendix A.3).

4.1.1 Counterfactual summaries

Considering that annotators should consult the initial QA pairs apart from exclusively relying on the more easily readable QA summaries provided by ChatGPT, we test their cautiousness by inserting 31 additional samples containing counterfactual summaries in place of the original ones—without them knowing. Those summaries are purposely unfaithful to the original QA pairs, guiding an annotator towards believing the responses belong to a different category compared to the actual one. We specifically generate them by prompting ChatGPT to select an incorrect (counterfactual) label and then then generate a suitable summary [3]. We manually verified the misleading aspect for each generated summary. We computed for each annotator the ratio of selecting the counterfactual label instead of the correct one and found that it to be ≤ 0.08. We thus assert that annotators do not solely rely on ChatGPT summaries and confirm the validity of the process, since the annotators were not significantly affected by the counterfactual summaries.

4.2 Validation set & inter-annotator agreement

As the proposed task is rather challenging and annotator perspectives could influence their final decisions we used a subset of the data (317 question-answer pairs) as validation for which we collected annotations from all 3 non-expert annotators. We calculated the inter-annotator agreement between the non-expert annotators, for both the fine-grained 'evasion' categories of our taxonomy (Figure 3, lower level classes) and the higher-level 'clarity' categories. We thus aim to both confirm the validity of our annotations and explore which labels draw most disagreements, potentially being more dependent on different perspectives and biases of annotators. We thus calculate Fleiss Kappa κ (Fleiss et al., 1971) for each label, and show the results in Figure 4 and Table 2 for the low- and high-level categories respectively.

	Clear Rep.	Clear Non-Rep.	Ambivalent
Clear Rep.	1	0.97	0.65
Clear Non-Rep.	0.97	1	0.71
Ambivalent	0.65	0.71	1

Table 2: The Fleiss score between all annotators for the classification between 'clarity' categories ('Clear Reply', 'Clear Non-Reply' and 'Ambivalent')

there is perfect agreement between annotators regarding the Clear Reply and Clear Non-Reply, while, rather intuitively, any confusion occurs mainly between the 'Clear non-reply' and the 'Ambivalent' categories. Looking into the heatmap in Figure 4 for the low-level categorisation can shed more light on the controversial labels. Specifically, we see this confusion stems from the difficulty annotators face when discriminating between 'Implicit' vs 'Explicit', 'General' vs 'Explicit', and 'Decline to answer' vs 'Dodging' categories. Overall, the heatmap shows that there is high confusion between 'General', 'Implicit', 'Dodging' and 'Deflection' categories, while there is a clear distinction of 'Claim ignorance', 'Decline to answer' and 'Clarification' categories with respect to the rest [4].

Handling disagreements As we intend to use the described validation dataset for evaluation (i.e. as our testset), we opted for resolving the disagreements and obtaining a single gold-label for the test set of 317 samples used in our experiments. When a disagreement between non-expert annotators occurs, a majority voting scheme is employed to decide the gold label. If there is no majority label, the expert annotator is tasked to resolve the conflict

[3] The prompt producing counterfactual summaries is demonstrated in Appendix B.

[4] The 'Contradictory' and 'Deflection' categories are not showcased in Figure 4, since there were no corresponding annotations.

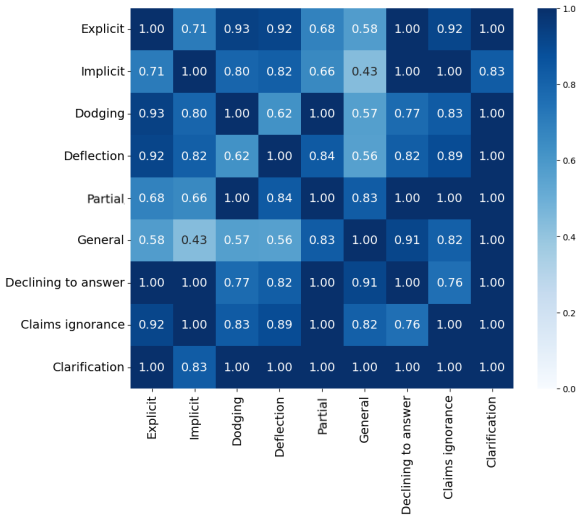

Figure 4: Confusion matrix of assigned taxonomy labels as resulting from our annotators.

by assigning the final gold label to the respective samples.

It is worth noting that deviating annotations are not necessarily invalid and can represent a variability of perspectives that would be useful to model instead of resolving. Recent work has highlighted the importance of access to multiple perspectives for complex NLP tasks and we have seen the emergence of datasets that maintain several annotations per instance to motivate training models under uncertainty or variability of annotations (Baan et al., 2022, 2023; Plank, 2022; Giulianelli et al., 2023). With this in mind, we will release the full annotations alongside the single-label dataset. However, further computational analysis and implementation of baseline models that are trained on multiple annotations per instance were deemed out of the scope of this work but would be an interesting direction for more robust modelling.

5 Experiments

5.1 Experimental setup

We test a variety of models on our introduced validation set (see Section 4.2), aiming to showcase the impact of different architectures and model sizes. Specifically, we leverage the following model architectures: Llama2 (Touvron et al., 2023), Falcon (Almazrouei et al., 2023) and ChatGPT (gpt3.5_turbo). Apart from comparing different model sizes, we aim to compare different training and inference strategies. Namely, we compare zero-shot inference, inference after instruction-tuning on the target labels, and finally, inference via chain-of-thought (CoT) prompting variants (prompts provided in Appendix B).

For the instruction-tuning part, we rely on LoRA finetuning (Hu et al., 2021) with $r = 16$, $alpha = 32$ and $dropout = 0.05$ (Hu et al., 2021) using a subset of 2700 annotated samples as training set and the rest 750 as validation set. Our CoT approach employs a breakdown of instructions, as well as the "Let's think step by step" phrase (Kojima et al., 2023), asking the model to first reason about the question and answer and then to classify with respect to the taxonomy. We compare two CoT flavors: ① *standalone CoT* classifies one , while ② *multiple CoT* is asked to assign a label to all responses that correspond to all sub-questions at one go, instead of handling each subcomponent and summarised question-answer pair independently. The reported results are based on the same test dataset, which consists of 317 samples where disagreements have been resolved.

5.1.1 Classification strategies

We explore two different classification strategies to classify responses with respect to clarity (i.e., the high-level categories):

1. **Direct clarity classification**, where we tune and prompt models to directly predict one of the 3 labels: Clear reply, Ambivalent Reply and Clear non-reply.
2. **Evasion-based clarity classification**, where we infer the clarity labels in two steps: we first tune and prompt the models to predict the 11 sub-categories (positioned on the leaves of the taxonomy tree) and then we infer the 3 labels by traversing the hierarchy of the taxonomy upwards.

5.2 Evaluation

Classification results for the different training and inference strategies are provided in Tables 3, 4, 5.

For the zero-shot setup, we present results exclusively for the larger models, as the performance for the smaller variants (Llama 7B, 13B and Falcon 7B) was very low (the models hallucinated and rarely predicted a label in the provided taxonomy). We can observe that ChatGPT significantly outperforms the other two models across metrics for both classification strategies, and is positively influenced by the two-step evasion-based strategy. While Falcon seems to also greatly benefit from being prompted to generate the fine-grained (and

Classification strategy	Model	Acc.	Prec.	Recall	F1
direct clarity	Llama-70b	0.467	0.429	0.235	0.259
	Falcon-40b	0.240	0.252	0.247	0.144
	ChatGPT	0.649	0.476	0.413	0.413
evasion-based clarity	Llama-70b	0.385	0.396	0.308	0.261
	Falcon-40b	0.618	0.365	0.387	0.375
	ChatGPT	**0.640**	**0.507**	**0.497**	**0.482**

Table 3: Classification results for zero-shot (ZS) inference. The best results for each strategy are underlined and best results overall are also in **bold**.

Classification strategy	Model	Acc.	Prec.	Recall	F1
direct clarity	zero-shot	0.649	0.476	0.413	0.413
	standalone CoT	0.628	0.414	0.376	0.368
evasion-based clarity	zero-shot	0.640	0.507	0.497	0.482
	standalone CoT	**0.688**	**0.611**	**0.514**	**0.510**
	multi CoT	0.549	0.459	0.500	0.462

Table 4: Classification results for chain-of-thought (CoT) inference using ChatGPT. The best results for each strategy are underlined and best results overall are also in **bold**.

Classification strategy	Model	Acc.	Prec.	Recall	F1
direct clarity	Llama-7b	0.500	0.455	0.546	0.466
	Llama-13b	0.621	0.580	0.721	0.602
evasion-based clarity	Llama-7b	0.653	0.586	0.608	0.596
	Llama-13b	0.663	0.613	0.615	0.614
	Llama-70b	**0.716**	**0.683**	**0.694**	**0.684**
	Falcon-7b	0.590	0.570	0.460	0.480
	Falcon-40b	0.615	0.598	0.537	0.557

Table 5: Classification results for instruction-tuned models. The best results for each strategy are underlined and best results overall are also in **bold**.

thus more descriptive) labels, Llama has the opposite behaviour, as it performs worse on the 11-way classification task and thus moving up in the hierarchy leads to increased misclassifications. Instead, it seems that Llama has a better representation for the high-level labels, thus performing better on the direct clarity classification task.

Turning to the CoT experiments we can observe a different behaviour with respect to each classification strategy. Specifically, CoT seems to improve the performance only for the evasion-based strategy, hinting that the "step-bu-step" reasoning process is more meaningful when addressing a more complex task with higher dimensionality/complexity of targeted labels. Interestingly, asking to address all sub-questions and answers in one go (multi-CoT) harms performance instead of improving, potentially because of the impact on the amount of context that needs to be taken into account for generation.

We also perform experiments by instruction tuning variants of Llama and Falcon (the instruction format is provided in Appendix B). Unlike the zero-shot, we observe that evasion-based classification consistently boosts performance for Llama. Additionally, we can observe that the Llama variants outperform Falcon even with fewer parameters (e.g. the 13B Llama model outperforms the 40B Falcon across metrics)

Overall, we observe that evasion-based clarity classification strategy leads to better performance compared to the direct clarity one, indicating that the fine-grained subcategories of the taxonomy assisted in guiding the LLMs towards selecting the correct high-level clarity category more frequently.

6 Conclusion

In this work, we introduce a novel task on response clarity classification in (political) interview scenarios. Driven by popular evasion techniques studied in political sciences, we propose a two-level hierarchical taxonomy for clarity classification that considers different evasion strategies at the lower (leaf) level. We also introduce a new dataset where we annotated question-response pairs with the taxonomy labels. We experiment with a range of different LLM model architectures, sizes and inference strategies on our dataset, providing a wide range of baselines. We show empirically that the two-level taxonomy, with the fine-grained labels for the unclear responses, helps towards the clarity classification, as their use boosts performance across models and inference strategies. We aspire for this work to motivate future research in the topic and we intend to expand our analysis and experiments, potentially exploring different prompting techniques such as in-context learning.

Limitations

Due to the usage of Large Language Models (ChatGPT) in our pipeline, our annotation process is susceptible to hallucinations, possibly affecting the quality of the "summary" extraction and therefore the assignment of correct labels. However, we attempt to mitigate this risk by asserting that our human annotators are attentive and not influenced by injected counterfactual summaries. Additionally, we manually inspected the quality of both the ChatGPT-generated summaries and the human annotations throughout the annotation campaign. Further, despite being crucial for the quality of the derived dataset, the need for human annotators significantly limits the number of samples that can be annotated, especially when considering the complexity of the proposed task. Finally, our dataset and respective analysis are limited to the English language and further work would be needed to generalise the findings to other languages, especially low-resource ones.

Potential risks

Potential risks associated with this work is the possibility of misclassification of a part of political speech due to the usage of neural models (LLMs) as classifiers. This fact may result in erroneously marking politicians' claims as unclear and evasive, if our method is used in real-world scenarios without human monitoring, and especially since the current state of LLMs under usage tend to hallucinate and produce unfaithful outputs.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Akari Asai and Eunsol Choi. 2020. Challenges in information-seeking qa: Unanswerable questions and paragraph retrieval. In *Annual Meeting of the Association for Computational Linguistics*.

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in natural language generation: From theory to applications. *arXiv preprint arXiv:2307.15703*.

Janet Beavin Bavelas, Alex Black, Lisa Bryson, and Jennifer Mullett. 1988. Political equivocation: A situational explanation. *Journal of Language and Social Psychology*, 7:137 – 145.

BingningWang, Ting Yao, Qi Zhang, Jingfang Xu, and Xiaochuan Wang. 2020. Reco: A large scale chinese reading comprehension dataset on opinion.

P. Bull. 2003. *The Microanalysis of Political Communication: Claptrap and Ambiguity*. Routledge.

Peter Bull. 1994. On identifying questions, replies, and non-replies in political interviews. *Journal of Language and Social Psychology*, 13:115 – 131.

Peter Bull. 2009. *Techniques of political interview analysis*, pages 215–228. Cambridge Scholars Publishing.

Peter Bull and Kate Mayer. 1993. How not to answer questions in political interviews. *Political Psychology*, 14:651–666.

Peter Bull and William Strawson. 2019. Can't answer? won't answer? an analysis of equivocal responses by theresa may in prime minister's questions. *Parliamentary Affairs*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Jim T. Dillon. 1990. The practice of questioning.

Hung Du, Srikanth Thudumu, Sankhya Singh, Scott Barnett, Irini Logothetis, Rajesh Vasa, and Kon Mouzakis. 2022. A framework for evaluating mrc approaches with unanswerable questions. *2022 IEEE 18th International Conference on e-Science (e-Science)*, pages 435–436.

J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? evaluating uncertainty in neural text generators against human production variability. *arXiv preprint arXiv:2305.11707*.

Sandra Harris. 1991. Evasive action: how politicians respond to questions in political interviews.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Gyeongbok Lee, Seung-won Hwang, and Hyunsouk Cho. 2020. Squad2-cr: Semi-supervised annotation for cause and rationales for unanswerability in squad 2.0. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5425–5432.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020. Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8129–8141, Online. Association for Computational Linguistics.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions.

Raymond S. Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2:175 – 220.

Liubov Nikolenko and Pouya Rezazadeh Kalehbasti. 2020. When in doubt, ask: Generating answerable and unanswerable questions, unsupervised. *ArXiv*, abs/2010.01611.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682.

Rajkumar Pujari and Dan Goldwasser. 2021. Understanding politics via contextualized discourse processing.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *ArXiv*, abs/1806.03822.

Parameswary Rasiah. 2010. A framework for the systematic analysis of evasion in parliamentary discourse. *Journal of Pragmatics*, 42:664–680.

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to ai complete question answering: A set of prerequisite real tasks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8722–8731.

Elior Sulem, Jamaal Hay, and Dan Roth. 2021. Do we know what we don't know? studying unanswerable questions beyond squad 2.0. In *Conference on Empirical Methods in Natural Language Processing*.

Haitian Sun, William Cohen, and Ruslan Salakhutdinov. 2022. ConditionalQA: A complex reading comprehension dataset with conditional answers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3627–3637, Dublin, Ireland. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Daniela Trotta and Sara Tonelli. 2021. Are gestures worth a thousand words? an analysis of interviews in the political domain. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 11–20, Groningen, Netherlands (Online). Association for Computational Linguistics.

Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa. 2022. Archivalqa: A large-scale benchmark dataset for open domain question answering over historical news collections.

Paul Watzlawick, Janet Beavin Bavelas, and Don D. Jackson. 1964. Pragmatics of human communication: A study of interactional patterns, pathologies and paradoxes.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

John-Charles Wilson. 1990. Politically speaking: The pragmatic analysis of political language.

Haichao Zhu, Li Dong, Furu Wei, Wenhui Wang, Bing Qin, and Ting Liu. 2019. Learning to ask unanswerable questions for machine reading comprehension. In *Annual Meeting of the Association for Computational Linguistics*.

A Dataset details

A.1 Interviews details

In Figure 5 we provide some temporal statistics regarding the interview distribution.

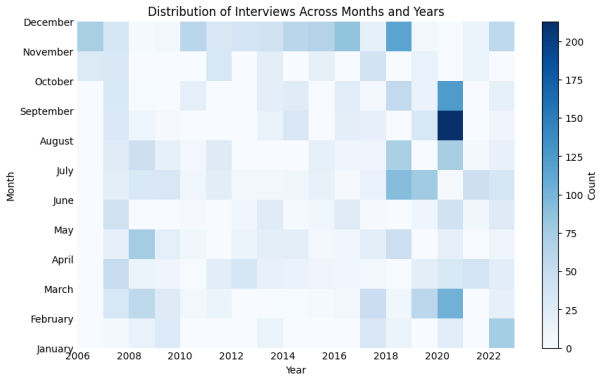

Figure 5: Visualization of interview distribution across months and years in the corpus

Moreover, details regarding the number of questions for all the 4 presidents existing in the interviews under consideration are provided in Figure 6.

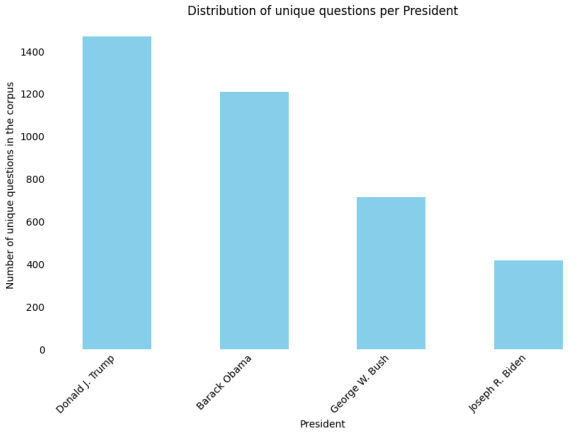

Figure 6: Visualization of distribution of unique questions per President in the corpus

A.2 Examples from the proposed taxonomy

In Table 6, we demonstrate some examples for all the categories mentioned in our proposed taxonomy. We also provide explanations on why these examples were classified in their respective categories.

A.3 Annotation details

Annotator's statistics All three non-expert annotators are of engineering background and participated in this annotation process voluntarily. The reason why we opted for non-expert annotators is because they are more representative of the general public, who are the receivers of political speech and do not have adequate background to immediately capture possible evasions, and therefore cannot fully evaluate the response clarity. The three non-experts were females, while the expert annotator is male. We do not disclose geographical characteristics to fully preserve anonymity.

Label distribution per annotator Figure 7 depicts the distribution of evasion labels for each annotator. The analysis reveals a generally consistent number of labels for each category across annotators. Notably, a slight disparity is observed for the explicit label, with annotator2 exhibiting a significantly different count compared to the other annotators. However, it's important to note that this doesn't necessarily imply a higher likelihood of Annotator2 to annotate instances with this label, as such behavior is not evident in the broader dataset analysis. The observed variation may be attributed to factors such as differing annotation styles or a higher occurrence of explicit responses within Annotator2's set.

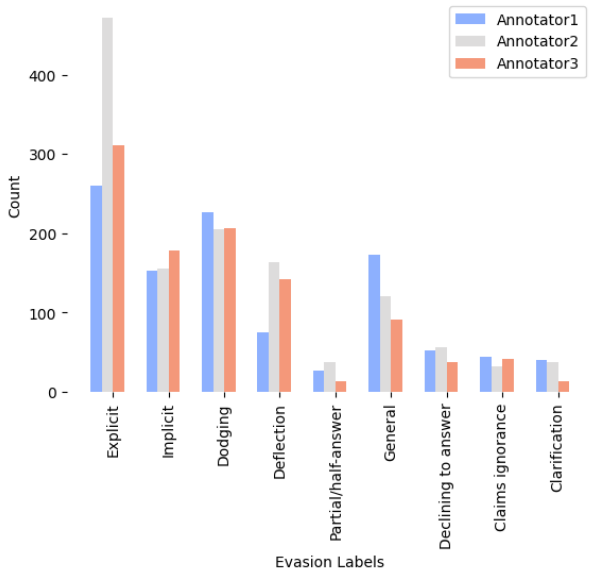

Figure 7: Visualization of distribution of evasion label per annotator in the corpus

Average annotation time per annotator The average time taken by each annotator to complete

	taxonomy	Description	Example
Clear R.	Explicit	The information requested is explicitly stated (in the requested form)	**Q:** er you have your own views about PR at Westminster don't you? **A:** I do. *Why? - directly gives the info requested*
Ambivalent Reply	Implicit	The information requested is given, but without being explicitly stated (not in the expected form)	**Q:** Are you going to watch television? **A:** What else is there to do? *Why? - they suggest planning to watch TV, despite not explicitly stating it*
	General	The information provided is too general/lacks the requested specificity	**Q:** What's your favourite film? **A:** Fight Club, Filth and Hereditary *Why? - the reply gives three movies instead of one, which makes the desired information unclear*
	Partial	Offers only a specific component of the requested information	**Q:** Did you enjoy the film? **A:** The directing was great *Why? - Directing is only part of what constitutes a film*
	Dodging	Ignoring the question altogether	**Q:** Do you like my new dress? **A:** We are late. *Why? - does not even acknowledge the question and goes straight to another topic*
	Deflection	Starts on topic but shifts the focus and makes a different point than what is asked	**Q:** Did you eat the last piece of pie? **A:** I have to admit that this was a great recipe, I always like it when there are chocolate chips in the dough. *Why? - acknowledges the question but goes on a tangent about the chips, without answering*
	Contradictory	The response makes conflicting statements	**Q:** Will you go the the grocery store? **A:** I will go, but I also won't go *Why? - self explanatory*
	Diffusion	Points out that the question is based on false hypotheses and does not provide the requested information	**Q:** Why is the earth flat? **A:** The earth is not flat. *Why? - renders the question invalid by saying that the premise is false*
Clear Non-Reply	Declining to answer	Acknowledge the question but directly or indirectly refusing to answer at the moment	**Q:** The hypothesis I was discussing, wouldn't you regard that as a defeat? **A:** I am not going to prophesy what will happen. *Why? - directly stating they won't answer*
	Claims ignorance	The answerer claims/admits not to know the answer themselves	**Q:** On what precise date did the government order the refit of the HMAS Kanimbla in preparation for its forward deployment to a possible war against Iraq? **A:** I do not know that date. I will find out and let the House know. *Why? - claims/admits they don't have the information*
	Clarification	Does not provide the requested information and asks for clarification	**Q:** Was it your decision to release the fund? **A:** You mean the public fund? *Why? - gives no data, asks for clarification*

Table 6: Descriptions and examples of political evasion techniques based on the proposed taxonomy

the annotation of a segment of an interview was 144.33 seconds (2.4 minutes), excluding instances with exceptionally large durations. This metric directly reflects the inherent complexity of the annotation task. Notably, this average annotation time remained consistent across all annotators.

Labelling platform Our labelling process was conducted in Label Studio platform. We provide some screenshots in Figures 8, 9 (they both belong to the same labelling page). Before the labelling process commenced, we provided detailed guidance to annotators on how to use the platform properly, so that any erroneous annotations because of limited familiarization with the platform are eliminated.

Annotations on presidential speech Extending the findings presented in Figure 2, Table 7 demonstrates more thorough results regarding the clarity of responses, as well as the evasion schemas leveraged by US politicians, as a result of our annotations. All of them tend to provide Ambivalent Replies more often than not, as denoted with red

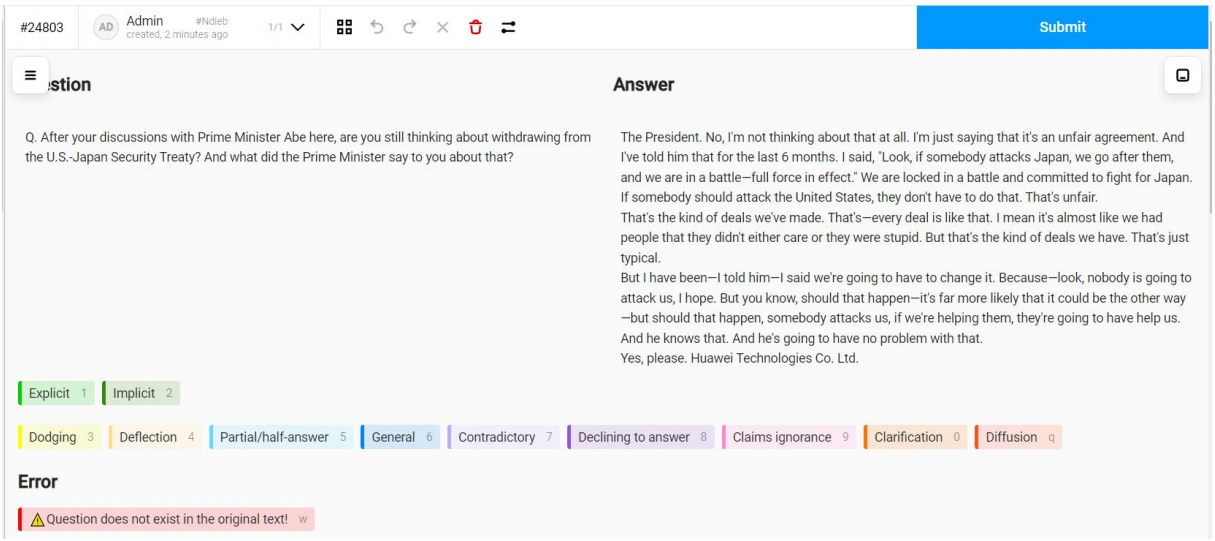

Figure 8: Screenshot from labelling platform: annotators have to read the original Question and Answer as provided. The classes corresponding to our proposed taxonomy are demonstrated as well.

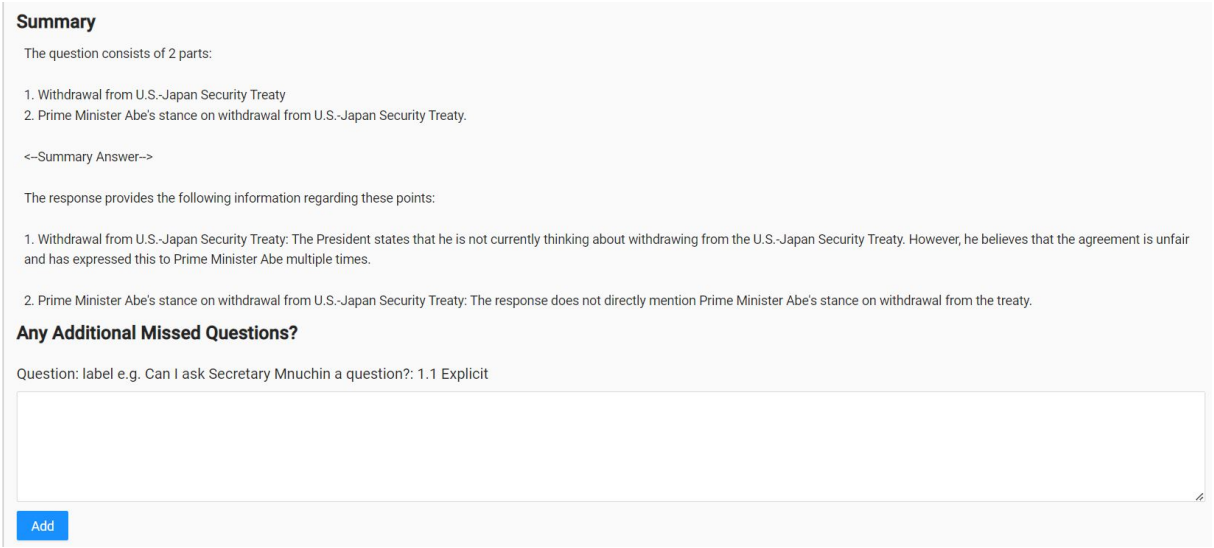

Figure 9: Screenshot from labelling platform: The "summaries" for the provided Questions and Answers are given to the annotators. They have to highlight each of the enumerated responses and assign one of the labels of the taxonomy (as presented in Figure 8) to each of them.

color. Especially Barack Obama utilizes Ambivalent responses more frequently than the rest of the presidents. Blue color denotes the most frequently used evasion technique, which in this case corresponds to 'Explicit replies'; nevertheless, explicit replies only account for about the 1/3rd of the responses for all presidents, leaving much space for evasion schemas to appear. In comparison, Joe Biden tends to provide more explicit replies, as resulting from our annotations.

B Prompting details

Prompt for generating summaries The following prompt was provided to ChatGPT to obtain the "summaries" of the question-answer pairs, as well as to request the appropriate label based on the proposed taxonomy.

message_0 = """

Point out what is this question Q asking. Stating of facts are not considered as questions, but only requests of information do. If it's a multi-part question, break down it the separate components that it asks. Use the following template to show the questions and the questions only.

The question consists of N parts: [add the correct N depending on the question] [Enumerate the question parts and give each part a short title in the beginning of the line] """

Response	G. W. Bush	B. Obama	D. J. Trump	J. R. Biden
Clear Reply	34.36	22.3	32.15	36.93
Clear Non-Reply	8.7	9.51	11.7	10.55
Ambivalent	56.94	**68.19**	56.15	52.51
Explicit	34.36	22.3	32.15	**36.93**
Implicit	14.45	18.04	12.08	10.55
Dodging	18.23	21.61	17.36	14.82
Deflection	12.34	10.31	10.94	10.8
Partial	1.4	2.28	1.89	5.03
General	9.82	14.37	10.42	7.79
Declining to answer	3.65	4.56	4.08	4.77
Claims ignorance	2.52	2.18	4.83	3.52
Clarification	2.52	2.78	2.79	2.26

Table 7: Statistics of answer clarity and evasion techniques in political interviews of the latest 4 US Presidents.

“““

message_1 = “““

Now analyse the information that this answer provides, especially regarding the points being asked, filling the following template.

Template — The response provides the following information regarding these points: [Enumerate the question parts along with their title, followed by the relevant information given per part in the response] — Answer:

“““

message_2 = “““

For each part of the question, and the questions only, use the following taxonomy to describe what type of a reply did the answer provide to it, along with a brief clarification for each choice. Note that if the question does not request elaboration, you should not consider the lack of elaboration in the answer as a lack of information. — Template:

Question part: [number and title]

Verdict: [taxonomy code and title]

Explanation:

—

<taxonomy>

“““

Prompt for generating counter-summaries In addition to this prompt, we create some “counter-summaries“ to assess the annotators’ reliance on the extracted summaries rather than the original question-answer pairs as provided in the interviews. The following prompt was appended to the previous one:

---

message_3 = “““

Now, try to create a summary of the response to intentionally mislead someone into thinking that the answer corresponds to a different category than the one you initially predicted. For instance, if your prediction is ’1.1 Explicit,’ generate a summary that could make someone believe it is a ’2.5 General’ response or any other label of your choice. The summary should be at the same length as the original one. Start by selecting the counterlabel and then write the summary using the following template:

Template

—

The response provides the following information regarding these points:

[Enumerate the question parts along with:

- title
- original label
- counterfactual label
- fake information for each part in the response supporting the counterfactual label.]

—

Answer:

“““

Zero-shot prompt for classification The following prompt was used for addressing the evasion problem in the zero-shot scenario.

---

message_0 = “““ Based on a segment of the interview in which the interviewer poses a series of questions, classify the type of response provided by the interviewee for the following question using the following taxonomy and then provide a chain of thought explanation for your decision:

<Taxonomy>

You are required to respond with a single term corresponding to the Taxonomy code and only.

Part of the interview
<Part of the interview>
Question
<Question>
Taxonomy code: “““

The following prompt was used for addressing the clarity problem in the zero-shot scenario.

---

message_0 = “““ Based on a segment of the interview in which the interviewer poses a series of questions, classify the type of response provided by the interviewee for the following question using the following taxonomy and then provide a chain of thought explanation for your decision:

1. Clear Reply - The information requested is explicitly stated (in the requested form)
2. Clear Non-Reply - The information requested is not given at all due to ignorance, need for clarification or declining to answer
3. Ambivalent Reply - The information requested is given in an incomplete way e.g. the answer is too general, partial, implicit, contradictory, diffused, dodging or deflection

You are required to respond with a single term corresponding to the Taxonomy code and only.

Part of the interview
<Part of the interview>
Question
<Question>
Taxonomy code: “““

Chain-of-Thought (CoT) prompt for classification The following prompt was used for addressing the evasion problem in the CoT scenario.

---

message_0 = “““ Based on a segment of the interview in

which the interviewer poses a series of questions, classify the type of response provided by the interviewee for the following question using the following taxonomy and then provide a chain of thought explanation for your decision:

<Taxonomy>

You are required to respond with a single term corresponding to the Taxonomy code as well as the chain of thought explanation.

Let's think step by step.
Part of the interview
<Part of the interview>
Question
<Question>
Taxonomy code: """

The following prompt was used for addressing the clarity problem in the CoT scenario.

---

message_0 = """ Based on a segment of the interview in which the interviewer poses a series of questions, classify the type of response provided by the interviewee for the following question using the following taxonomy and then provide a chain of thought explanation for your decision:

1. Clear Reply - The information requested is explicitly stated (in the requested form)
2. Clear Non-Reply - The information requested is not given at all due to ignorance, need for clarification or declining to answer
3. Ambivalent Reply - The information requested is given in an incomplete way e.g. the answer is too general, partial, implicit, contradictory, diffused, dodging or deflection

You are required to respond with a single term corresponding to the Taxonomy code as well as the chain of thought explanation.

Let's think step by step.
Part of the interview
<Part of the interview>
Question
<Question>
Taxonomy code: """

Prompt for LoRA finetuning The following prompt was used for LoRa fine-tuning, and it remained consistent across all models and the two methodologies (direct clarity and evasion-based clarity). The only distinction between the two different setups in the prompt was the specific label that the model should generate.

---

message_0 = """Based on a part of the interview where the interviewer asks a set of questions, classify the type of answer the interviewee provided for the following question
Part of the interview
<Interview Part>

Question
<Question>
Label: <Label>
"""

C Computational Resources

All the experiments were conducted on a cluster with 4 NVIDIA A100-SXM4-40GB. The total hours of experimentation for training and inference (both for zero-shot and fine-tuned models) were 210 GPU hours and 427 CPU hours.