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

Konstantinos Thomas[1] **Giorgos Filandrianos**[1] **Maria Lymperaiou**[1]
Chrysoula Zerva[2,3,4] and **Giorgos Stamou**[1]
[1]National Technical University of Athens
[2]Instituto de Telecomunicações
[3]Instituto Superior Técnico, Universidade de Lisboa [4]ELLIS Unit Lisbon
{kthomas, geofila, marialymp}@islab.ntua.gr
chrysoula.zerva@tecnico.ulisboa.pt, gstam@cs.ntua.gr

Abstract

Equivocation and ambiguity in public speech are well-studied discourse phenomena, especially in political science and 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 corresponding *clarity classification dataset* which consists of question-answer (QA) 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 for a given question (high-level) and also provides a fine-grained taxonomy of evasion techniques that relate to unclear, ambiguous responses (lower-level). We combine ChatGPT and human annotators to collect, validate and annotate discrete QA pairs from political interviews, to be used for our newly introduced response clarity task. We provide a detailed analysis and conduct several experiments with different model architectures, sizes and adaptation methods to gain insights and establish new baselines over the proposed dataset and task. [1]

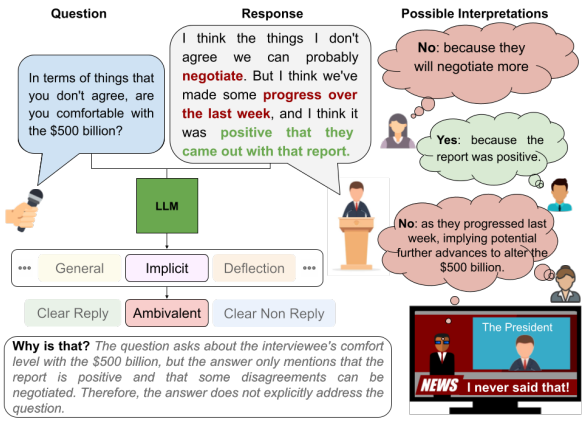


Figure 1: An example from an interview from our dataset with classification along with an analysis from instruction-tuned Llama-70b.

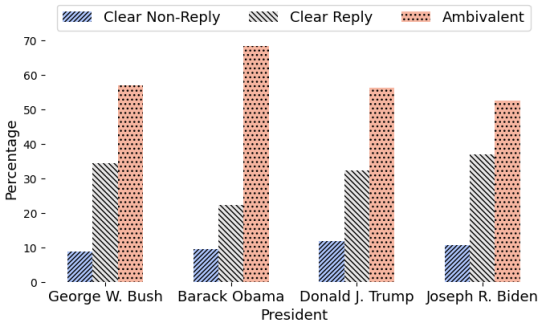


Figure 2: Statistics on answer clarity in political interviews of the latest 4 US presidents.

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 Q&As, concluding that politicians gave clear responses to only 39-46% of questions during televised interviews, while non-politicians had a significantly higher 70-89% reply rate. In Figure 2 we present statistics derived from our human annotations regarding response clarity among US presidents, revealing that politicians often avoid providing clear responses to journalists' questions.

This phenomenon is known 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 contradictions, inconsistencies, subject switches, incomplete sentences, misunderstandings, obscure mannerisms of speech (Watzlawick

[1]Code and Data can be found here: https://github.com/konstantinosftw/Question-Evasion.

Findings of the Association for Computational Linguistics: EACL 2024, pages 5204–5233
November 12-16, 2024

et al., 1964; Bavelas et al., 1988; Rasiah, 2010), rendering political speech susceptible to multiple interpretations from the perspective of the public. Figure 1 presents an example of an interview featuring various interpretations, generated labels, and corresponding explanations using our proposed dataset.

While the topic has been studied extensively in the field of linguistics, politics and communication, with several typologies proposed for classifying question responses (Harris, 1991; Bull and Mayer, 1993; Rasiah, 2010), there has been little attempt to analyse whether such typologies are applicable to larger scale data and consistent with varying human perspectives and biases. In other words, the possibility of automatically classifying response clarity has not been explored in NLP, potentially due to the complexity of the task itself, as well as the underlying need to encode and reason over long context. However, recent advancements in language modelling boosted model performance for long-context inputs (Dai et al., 2019; Wei et al., 2022, 2023), paving the way for framing the task of **automatically measuring response clarity**.

Related to this endeavour, there is related work focusing on the responder's intent interpretation (Ferracane et al., 2021), or the answerability of questions for question-answering (QA) tasks (Min et al., 2020; BingningWang et al., 2020; Rogers et al., 2020; Sun et al., 2022; Wang et al., 2022). However, in both research directions, the focus deviates from directly assessing the clarity of the response, being obfuscated by perceptions of intent or question clarity. We address this by proposing the task of **response clarity evaluation**, focusing exclusively on assessing the effect of the response, building on relevant discourse typologies.

We carry out a detailed analysis of proposed typologies, considering their overlap and consistency, the distribution 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 accounts for a three-way evaluation of response *clarity* in terms of the number of interpretations the intended response holds. The second and more fine-grained level covers eleven common *evasion* phenomena in political literature, which explain in more detail the categorization of responses in the three-scale clarity classes. We use this taxonomy to annotate a dataset of political QA pairs and perform an analysis of the perspective variability among human annotators. We then evaluate different LLMs, exploring various 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 labels) 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 §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 3,445 QA pairs from political interviews.
- We experiment with several language models and methods to gain insights establish performance baselines for the proposed task.

2 Related work

2.1 Equivocation in Social Sciences

Political equivocation, aptly generalised by Dillon (1990) as "the routine strategy for responding to a question without answering it", provides a range of frameworks to analyse evasive responses (Wilson, 1990; Bull, 2009; Bull and Strawson, 2019). Harris (1991) makes a distinction between direct and indirect answers while others focus on how complete the information conveyed by the response is (Bull, 1994, 2003). Wilson (1990); Harris (1991); Bull (2003) 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*, 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 (Table 1). Lastly, ③ *Intermediate replies* are those utterances that fall somewhere between replies and non-replies, i.e. responding completely but to one part of a multi-part question while ignoring the rest;

responding partially to a single-part question; answering a question in a suggestive manner without giving a straightforward answer.

Bull (2003) breaks the 12 evasion techniques of Table 1 further into 28 more fine-grained micro-categories; for example "*Makes political point*" includes the micro-categories "*External attacks on the opposition or other rival groups*", "*Talks up one's own side*", "*Presents policy*". Rasiah (2010) separates the Replies into Direct and Indirect, keeps the Intermediate Replies category as is, while also breaking down the Non-reply category (which he labels "Evasions") into four degrees of evasiveness, whether the evasion was overt or covert and what types of 'agenda shifts' occurred.

1. **Ignores the question.** Makes no attempt to answer the question, or even to acknowledge it 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 or 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: Equiv. typology by Bull and Strawson (2019).

Tailoring these typologies into a response clarity taxonomy suitable for an NLP dataset, it is imperative to modify them considering the following:

- Our focus is slightly different: we target a taxonomy that classifies the clarity of responses (hence an indirect response falls under a different category than a direct one).
- We seek a good per class representation in our dataset to allow computational modelling using LLMs. It is thus necessary to condense classes to avoid overly sparse categorisation while retaining the essential per class characteristics (i.e., we provide meaningful labels).
- Labelling of the responses is conducted by non-expert human annotators so that our annotations also account for the perception and reasoning of the general audience of political interviews rather than a minority of experts. The difficulty of the classification, and thus the resulting error rate, increases as we increase the set of labels they choose from.
- Most interviewers pose multi-barrelled questions. We break those multi-part questions into singular QA pairs and label each one separately, to retain this fine-grained information.

Section 3 discusses the taxonomy we adopted, aiming to optimise for the annotation task.

2.2 Equivocation in NLP

While *equivocation* has not been adequately studied in NLP, there are related areas, such as question answerability, political discourse analysis and deceptive intent detection.

2.2.1 Answerability in question answering

There have been several tasks proposed related to 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 our task shares a connection with question answerability, our focus is on annotating *response clarity* in relation to a given question. This distinc-

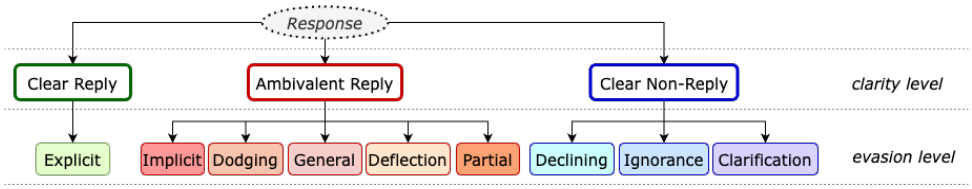


Figure 3: Our proposed taxonomy of response clarity classification.

tion shifts the goal from evaluating *question* clarity leading to a unique task and reasoning process.

2.2.2 Discourse analysis of political speech

Beyond evasion, discourse phenomena in political speech (including interview responses) have been analysed in prior NLP works. Majumder et al. (2020) construct a large-scale dataset of political dialogues 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 closest to our target). Understanding political agendas requires contextualization, depending on which politician expresses a certain claim: Pujari and Goldwasser (2021) propose 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 emphasising certain parts of speech (Trotta and Tonelli, 2021).

Another relevant dimension that has been explored in the context of automated discourse analysis is detecting the intent of the responder. (Girlea, 2017) trained Relational Dynamic Bayesian Networks on psycholinguistic features of non-political dialogues to identify linguistic cues associated with deception. In a work lying closer to ours, (Ferracane et al., 2021) crowdsourced annotators to label political interview answers, firstly as "answer", "shift" or "didn't answer" and ultimately whether that act had honest or deceptive intent. They thus aim to collect diverse, subjective opinions on the (dis)honesty of responders providing a valuable two-way view on the topic that involves both the responder and the audience (annotator). We instead opt for avoiding assumptions on speaker intent, and focusing only on discourse techniques the speaker used, since they are better defined in related literature, and allow us to directly evaluate the clarity of a response. For example, an on-topic response that is slightly open to interpretation would be labelled as "Implicit reply" under the "Ambivalent reply" category by our typology. While for (Ferracane et al., 2021), this would fall under the parent category of "Answer", and either "direct" or "over-answer", depending on whether the annotator felt that the speaker was *purposefully* ambiguous or not. This decision on the annotation focus allows us also to annotate a more extensive dataset ($\approx 3.4K$ pairs) due to its less subjective nature, which considers the level of clarity and completeness of responses.

3 Proposed Taxonomy

The typologies discussed in §2.1 are comprehensive and well-researched, but often exhibit compatibility issues (Bull, 1994; Bull and Strawson, 2019; Rasiah, 2010) as distinctions between categories vary among experts and sub-domains. For instance, a somewhat vague reply may be deemed as evasive by some while indirect yet coherent by others, especially since ambivalent responses are particularly prone to confirmation bias (Nickerson, 1998). To enhance objectivity, we focus on the Clarity/Ambiguity dimension, rather than a Reply/Non-reply distinction. This approach shifts annotators' attention from the bias-prone task of trying to decipher if an answer is "valid" or "invalid", to whether a response can be interpreted unambiguously or accepts a wider range of interpretations.

Extensive typologies such as Bull (2009) include over 30 types of replies, resulting in a sparse dataset with few examples per category that further complicates the annotation task. We thus aimed to consolidate these typologies into fewer essential categories, while maintaining crucial distinctions.

Another necessary adjustment involved breaking down multi-part questions into their constituent questions, which led to the elimination of the category of "intermediate replies". As discussed in §2.1, most interviewers pose multi-barrelled questions and vagueness in a single answer towards a multi-part question results in classifying the entire response as an intermediate reply. To avoid skewing the dataset towards intermediate replies,

we broke multi-barrelled questions into separate questions and asked the annotators to label each sub-question and answer separately.

Taking all of the above into consideration, we arrived at a two-level hierarchical taxonomy. The higher level includes 3 main response categories, namely ① *Clear reply*, containing replies that admit only one interpretation; ② *Clear non-reply*, containing responses where the answerer openly refuses to share information, and ③ *Ambivalent reply*, where a response is given in the form of a valid answer but allows for multiple interpretations. At the second level these 3 categories further split into 9 sub-categories illustrated in Figure 3. As a brief exemplification, "*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 there is a rather clear implication on the culprit. Yet, the answer does not commit to explicitly stating that "the kids ate it" - which would have made for an Explicit reply - but rather prompts for a reasoning step to reach the final assumption. Instead, "*A. I don't know*", for the same question, would be labelled as a *Clear non-reply* and specifically *Claims ignorance*, since the respondent explicitly refuses to provide information; also, "*A. You should not keep your chocolates all around the house*" would be considered a *Deflection*, i.e. an *Ambivalent* answer, as it provides none of the requested information, yet it leverages the subject to pivot on a different point. For further analysis and examples see Table 4 in App. A.2.

4 Dataset creation

As a first step, we collect presidential interviews of US Presidents, provided by the official Whitehouse website [2]. This resulted in 287 unique interviews spanning from 2006 until 2023 which we further analyse in App. A.1. We extracted a total of 3,445 questions and responses from these interviews, as described in the following sections.

We leverage ChatGPT to decompose the original interviews into QA pairs, aiming to separate multi-barrelled questions into separate sub-questions and their respective response sub-parts. We use the automatically generated list of (sub-)questions to generate annotation instances, and then, upon validating the decomposition, annotators label the response to each sub-question separately. Thus, for a given interview question, we may have several QA instances in the final dataset corresponding to distinct sub-questions, and the classification of the respective sub-responses. We henceforth refer to the generated sub-questions and sub-responses as *singular QA pairs*, "*sQAs*" for short.

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 QAs as well as the decomposed sQAs, and asked to label the response for each sub-question separately. We opted for providing the sQAs alongside the full text to reduce the effort of manually extracting distinct sQAs from the original interviews, which would significantly increase the annotation time per sample. We further introduce *counterfactual sQAs* to measure the annotators' potentially exclusive reliance on sQAs, as explained in App. A.3. We were thus able to verify that all annotators followed our instructions and the introduction of sQAs aids instead of hindering the annotation process. The prompt provided to ChatGPT to create the original sQAs and counterfactual sQAs is shown in App. H.

We employ 3 human annotators alongside an expert with a background in political science and political discourse analysis who acts as a validator of the outcome annotations. As a first "training" stage, we provide the annotators with a tutorial that includes annotated examples from each category of the taxonomy to allow them to familiarise themselves with the concepts introduced. Then, the annotators are prompted to perform a series of annotation tasks in the following order: they have to ① evaluate the sQAs produced by ChatGPT as valid or not, and then ② label each of the individual questions and answers, using the proposed taxonomy or indicate an erroneous question in sQAs. Finally, they should ③ add any missing questions, as well as the corresponding label. On average, each annotator evaluated 1150 samples. More information is provided in App. A.3.

Validation set & inter-annotator agreement As the proposed task is challenging and annotator perspectives may influence their decisions, we use a subset of the data (317 common QA pairs) as *validation* for which we collect overlapping annotations from all 3 non-expert annotators. We calculate the inter-annotator agreement between the non-experts, for both the fine-grained 'evasion' taxonomy categories (Figure 3, lower level classes) and the higher-level 'clarity' categories. We thus

[2]Interviews from https://www.whitehouse.gov/.

aim to both confirm the validity of our annotations and explore which labels draw more disagreements, potentially being more dependent on diverging perspectives and biases of annotators or being inherently harder to distinguish. Table 2 shows the annotators' agreement via Fleiss Kappa κ scores (Fleiss et al., 1971) when given samples from two different 'clarity' classes (row, column). Similarly, Figure 4 concerns the 'evasion' level classification.

	Clear R.	Clear Non-R.	Ambiv.
Clear R.	1	0.97	0.65
Clear Non-R.	0.97	1	0.71
Ambiv.	0.65	0.71	1

Table 2: Fleiss κ (higher values are better) between annotators for the 'clarity' classification level.

For the 'clarity' category, the Fleiss Kappa κ indicates moderate to high agreement among non-expert annotators at 0.644, compared to 0.48 for the more challenging 'evasion' classification, signifying moderate agreement. There is near perfect agreement between annotators regarding Clear Reply and Clear Non-Reply (κ=0.97), while, rather intuitively, confusions occur when distinguishing between *Ambivalent* category and any of the rest. Figure 4 sheds more light on the confused labels: it seems that annotators diverge more when discriminating between *General* (*Ambivalent*) vs *Explicit* (*Clear Reply*) (κ=0.58) and *Partial* (*Ambivalent*) vs *Explicit* (*Clear Reply*) (κ=0.68), or 'Declining' (Clear Non-reply) vs'Dodging' (Ambivalent) (κ=0.77). On the contrary, there is a clear distinction between 'Claim ignorance', 'Decline to answer' 'Clarification' categories and 'Explicit' replies ($\kappa \geq$0.92). Moreover, there is also high disagreement within Ambivalent labels, such as 'General' vs 'Implicit', 'General' vs 'Deflection', and 'General' vs 'Dodging' categories.

Handling disagreements As we intend to use the described *validation* set in the evaluation stage (i.e. as our test set), we opt for resolving the disagreements and obtaining a single gold label for all these 317 *validation* samples. 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 resolves the conflict by assigning the final gold label to the respective samples.

Notably, deviating annotations are not necessarily invalid and can represent a variability of perspectives that could be useful to model instead of resolve. Recent work has highlighted the importance of access to multiple perspectives for complex NLP tasks, encouraged by the emergence of datasets that maintain several annotations per instance to motivate training models under uncertainty or annotation variation (Baan et al., 2022, 2023; Plank, 2022; Giulianelli et al., 2023). Hence, and while capturing diverting perspectives is out of scope for this work, we release the full annotations alongside the single-label dataset, to allow for future research into models that can address multi-label scenarios.

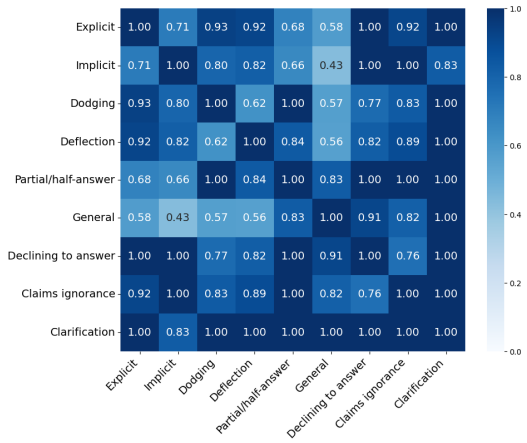


Figure 4: Annotators' agreement using Fleiss κ for labels assigned to the 'evasion' classification level.

Exploratory data analysis revealed shifts in evasion patterns, such as an increased reply rate at the end of the presidential service for some presidents (e.g. D. Trump), while the opposite behaviour is derived for others (e.g. G. Bush). Additionally, evasion correlates with the presence of multi-part questions Interestingly, while in joint interviews, presidents tend to alter their reply strategy compared to when being interviewed on their own. We provide more details in App. A.1.

5 Experiments

5.1 Experimental setup

We test various models on our disagreement-resolved *validation* set to showcase the impact of different modelling choices and establish baselines. Details regarding experiments in App. B.

Modeling variants We compare (i) encoder models: DeBERTa (He et al., 2021), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019); (ii)LLMs: Llama2 (Touvron et al., 2023), Falcon (Almazrouei et al., 2023); and (iii) ChatGPT

(gpt3.5_turbo) [3]. Additionally, we compare varying adaptation strategies, namely inference via zero (ZS) or few-shot (FS) and chain-of-thought (CoT) prompting variants (prompts provided in App. H), as well as instruction-tuning on the target labels using LoRA tuning (more details in App. H.1).

Our CoT approach employs a breakdown of instructions, as well as the "Let's think step by step" phrase (Kojima et al., 2022), asking the model to first reason about QAs and then classify based on the taxonomy. We compare two CoT flavors: ① *standalone CoT* classifies only one sQA at a time, and ② *multiple CoT* attempts to classify all sQAs pertaining to a multi-barrelled question in one go. For the instruction-tuning part, we rely on LoRA fine-tuning (Hu et al., 2021). The details of the experiments are provided in App. B, while the instruction format is outlined in App. H.1.

Classification variants We explore two different classification variants to evaluate responses: ① **Direct clarity classification**: we tune and prompt models to directly predict one of the 3 labels of the clarity level: Clear reply, Ambivalent Reply and Clear non-reply. ② **Evasion-based clarity classification**: we infer the clarity labels in two steps. First, we tune and prompt the models to predict the 9 evasion sub-categories (leaves of the taxonomy tree) and then we infer the 3 labels by traversing the taxonomy hierarchy upwards.

5.2 Results and Discussion

Classification results for different training and inference strategies are provided in Table 3. More detailed analysis can be found in App.F [4].

For the **ZS** setup, we exclusively present results for the larger models due to the very low performance of the smaller ones (Llama 7B/13B and Falcon 7B), which frequently hallucinated and rarely predicted labels within the taxonomy. ChatGPT significantly outperforms the other two models across metrics for both classification variants, and it is positively influenced by the two-step evasion-based strategy. While Falcon also benefits from generating fine-grained labels, Llama exhibits the opposite behaviour, performing worse on the 9-way classification task and thus moving up in the hierarchy leading to increased misclassifications. Instead, Llama has a better representation of the high-level

Classification variant	Model	Acc.	Prec.	Recall	F1
Prompting					
direct clarity	ZS Llama-70b	0.467	0.429	0.235	0.259
	ZS Falcon-40b	0.240	0.252	0.247	0.144
	ZS ChatGPT	0.649	0.476	0.413	0.413
	FS Llama-7b	0.23	0.159	0.474	0.219
	FS Llama-13b	0.211	0.105	0.302	0.156
	FS Llama-70b	0.667	0.333	0.333	0.333
	FS Falcon-7b	0.203	0.107	0.267	0.152
	FS Falcon-40b	0.29	0.13	0.336	0.186
	standalone CoT	0.628	0.414	0.376	0.368
evasion-based clarity	ZS Llama-70b	0.385	0.396	0.308	0.261
	ZS Falcon-40b	0.618	0.365	0.387	0.375
	ZS ChatGPT	0.640	0.507	0.497	0.482
	FS Llama-7b	0.274	0.393	0.335	0.262
	FS Llama-13b	0.291	0.452	0.363	0.259
	FS Llama-70b	0.541	0.565	0.452	0.365
	FS Falcon-7b	0.505	0.299	0.211	0.222
	FS Falcon-40b	0.429	0.167	0.25	0.2
	standalone CoT	0.688	0.611	0.514	0.510
	multi CoT	0.549	0.459	0.500	0.462
Tuned models					
direct clarity	DeBERTa-base	0.58	0.521	0.453	0.441
	RoBERTa-base	0.64	0.579	0.516	0.53
	XLNet-base	0.694	0.52	0.523	0.518
	Llama-7b	0.489	0.452	0.529	0.457
	Llama-13b	0.587	0.579	0.7	0.58
	Llama-70b	**0.759**	**0.67**	0.70	0.68
	Falcon-7b	0.288	0.325	0.333	0.175
	Falcon-40b	0.341	0.512	0.534	0.356
evasion-based clarity	DeBERTa-base	0.555	0.53	0.671	0.537
	RoBERTa-base	0.577	0.501	0.534	0.495
	XLNet-base	0.58	0.523	0.586	0.546
	Llama-7b	0.666	0.618	0.616	0.616
	Llama-13b	0.675	0.617	0.616	0.616
	Llama-70b	0.713	**0.67**	**0.71**	**0.682**
	Falcon-7b	0.533	0.429	0.386	0.397
	Falcon-40b	0.621	0.616	0.532	0.558

Table 3: Results for ZS, FS & CoT prompting inference, as well as for fine/instruction-tuned models. The best results for each prompting/training variant are underlined and best results overall are also in **bold**.

labels, performing better on the direct clarity classification. For **FS**, due to the lengthy sQAs of our dataset's interviews, we employ shorter representative examples (Table 4). FS showcased advanced results compared to ZS, with smaller models experiencing a significant reduction in hallucinations. Further analysis is provided in App. D.1.

CoT experiments exhibit a different behaviour for each classification variant. Specifically, CoT improves the performance for the evasion-based strategy only, hinting that the "step-by-step" reasoning process is more meaningful when address-

[3] Specifically, we used version gpt-3.5-turbo-0613.

[4] Note that results for XxBERTa models are overestimated due to constraint input token size.

ing a task with higher dimensionality/complexity of targeted labels. Interestingly, asking to address all sQAs 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.

In general, LLMs mostly struggled with distinguishing between *Clear* vs *Ambivalent* replies, as well as *Partial* vs *General* ones. This resembles challenges (Figure 4) faced by human annotators but interestingly holds even for ZS and CoT models which were not trained on human annotations, suggesting a generalised difficulty in discerning these classes. Further insights are shown in App. E.

Turning to tuned models, we observe a difference in behaviour: for direct clarity, smaller LLM models seem to struggle and are even outperformed by encoder models such as XLNet or BERT variants, with only the 70b Llama outperforming them. Instead, evasion-driven classification consistently improves the performance of Llama variants. Additionally, Llama models outperform Falcon even with fewer parameters (e.g. the 13B Llama model outperforms the 40B Falcon across metrics). This aligns with other works where LLama-13b surpasses Falcon-40b in reading comprehension (Touvron et al., 2023), while all LLama variants exhibit better prior knowledge (Sun et al., 2023), a crucial factor for our task as discussed below. We expand our experiments to assess the generalisation capabilities of the stronger Llama model (70B) using the dataset of (Ferracane et al., 2021), which is annotated with a different strategy, and provide an analysis as detailed in App. G.

Overall, for both prompting and tuning strategies, the evasion-based clarity classification variant 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. In other words, while the 9-way classification is more challenging (see also App. E), disambiguation between the finer-grained labels helps the models improve their accuracy on the higher-level ones. Further analysis of performance per class is provided in App. C.

Answer grounding We aim to separately assess whether models are influenced by the difficulty of identifying the relevant response snippets in the text, i.e. grounding the answer, a task that can be particularly challenging when a single reply addresses multiple questions. As a proxy to test this, we consider single- vs multi-part question subsets (35% vs 65% of the original test-set), assuming that answer grounding is harder for the latter, and we compare models and annotator performance. While Fleiss κ showed minimal disparity between humans across all models, metrics were notably higher for single-part questions, regardless of the method (ZS/FS, CoT, fine-tuning) or the classification variant (evasion-based or direct clarity). Performance improvements reached 0.16 for F-score, indicating the impact of QA complexity on model performance. More detailed results in App D.2.

Model knowledge We explore whether performance in the proposed task is influenced by models' "prior knowledge" of given entities. For instance, Q: "Did the Federal Reserve make the right move?", A: "I think Bernanke is doing a great job" would be correctly classified as *Dodging* by models unaware that Bernanke is the chairman of the Federal Reserve. To explore the prior knowledge hypothesis, we focus on person names and divide the test-set into two parts: one containing person names in either the question or the answer, and one excluding any named person mentions (60% vs 40% of the original test-set). All models performed better on the latter, "no-person" subset, but smaller models exhibited a much sharper improvement of up to 0.20 in F-score (Llama-7b) compared to larger and presumably more "knowledgeable" ones, thus corroborating the findings of Sun et al. (2023). We provide more details in App D.3.

6 Conclusion

We introduce a novel task on response clarity classification focusing on political interviews. Driven by studies of evasion techniques 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 question-answer pairs are manually annotated with the proposed taxonomy labels. We experiment with a range of different model architectures, sizes and adaptation strategies on our dataset, establishing several baselines. We empirically show that fine-grained labels facilitate classification in response clarity, while encoded model knowledge is strongly associated with classification performance. We aspire for this work to motivate future research in the topic, both from the NLP and political sciences communities.

Limitations

Due to the usage of Large Language Models (ChatGPT) in our pipeline, our annotation process is susceptible to hallucinations, which could affect the quality of the sQA 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 injecting counterfactual sQAs. Additionally, we manually inspected the quality of both the ChatGPT-generated sQAs and the human annotations throughout the annotation campaign to ensure high-quality annotations. 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. Overall, 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. Finally, the inherently missing vocal features present in speech, as well as face movements and hand gestures limit the discourse analysis to purely textual cues, potentially missing some evasion-related characteristics.

Potential risks

Potential risks associated with this work relate to 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, especially since the current state of LLMs under usage tends to hallucinate and produce unfaithful outputs. Hence, further work to ensure the reliability and trustworthiness of the underlying models would be crucial for their deployment.

Acknowledgments

The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the 3rd Call for HFRI PhD Fellowships (Fellowship Number 5537).

This work was supported by the Portuguese Recovery and Resilience Plan through project C645008882000000055 (NextGenAI - Center for Responsible AI), by the EU's Horizon Europe Research and Innovation Actions (UTTER, contract 101070631), and by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020.

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.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

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.

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2021. Did they answer? subjective acts and intents in conversational discourse. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1626–1644, Online. Association for Computational Linguistics.

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

Codruta Liliana Girlea. 2017. Deception detection in dialogues.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

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

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

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. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

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. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

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. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1353–1367, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

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.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023. Head-to-tail: How knowledgeable are large language models (llm). *AKA will llms replace knowledge graphs*.

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. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3025–3035.

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.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

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 Exploratory data analysis

In this section, we describe some interesting patterns present in our proposed dataset.

Label distribution We start our analysis from the core of this work, which is the distribution of the final labels of our dataset, which are presented in Figure 5. Overall, *Explicit Replies* is the most prevalent category, followed by evasion categories with significantly lower frequency each. Specifically, Explicit Replies contribute to 1051 samples in total, followed by Dodging (704 samples), Implicit (488 samples), General (386 samples), Deflection (381 samples), Declining to answer (145 samples), Claims ignorance (119 samples), Clarification (92 samples) and finally Partial/half-answer (79 samples).

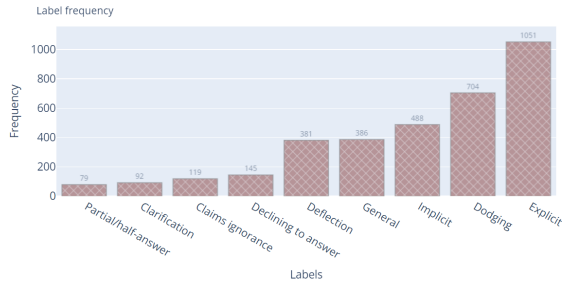


Figure 5: Label distribution in the dataset.

We also analyze the label distribution per president in Figure 6, offering a more detailed insight compared to Figure 2. According to the per president distribution, we conclude that in our collected interviews Donald J. Trump tends to provide more Explicit Replies than the rest of the US presidents, as indicated by the light-colored square of Figure 6.

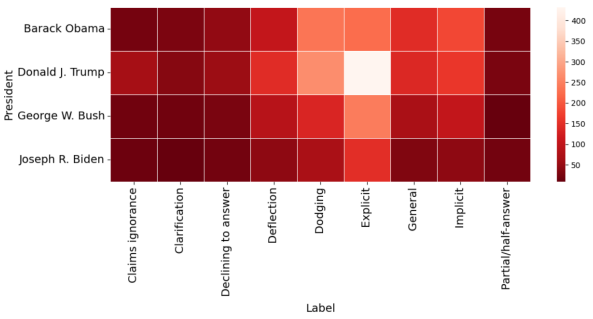


Figure 6: Label distribution per president.

In the following paragraphs we will delve into the insights behind these label distributions.

Temporal insights Moving on to temporal characteristics, in Figure 7 we provide some temporal statistics regarding the interview distribution.

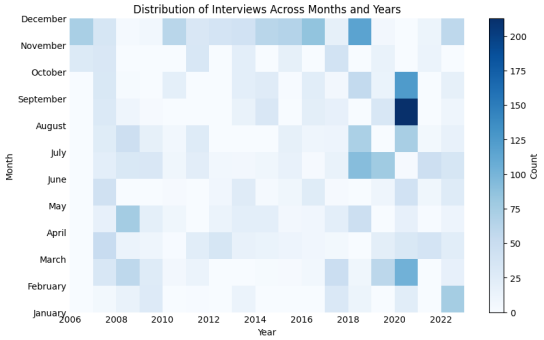


Figure 7: Visualization of interview distribution across months and years in the corpus

In Figure 8 we present the label distribution per year in our dataset. We observe an elevated number of Explicit Replies in 2020, as indicated by the light-colored cell. This observation can be grounded to president-related information, as this can be a strong characteristic in conjunction to label distribution.

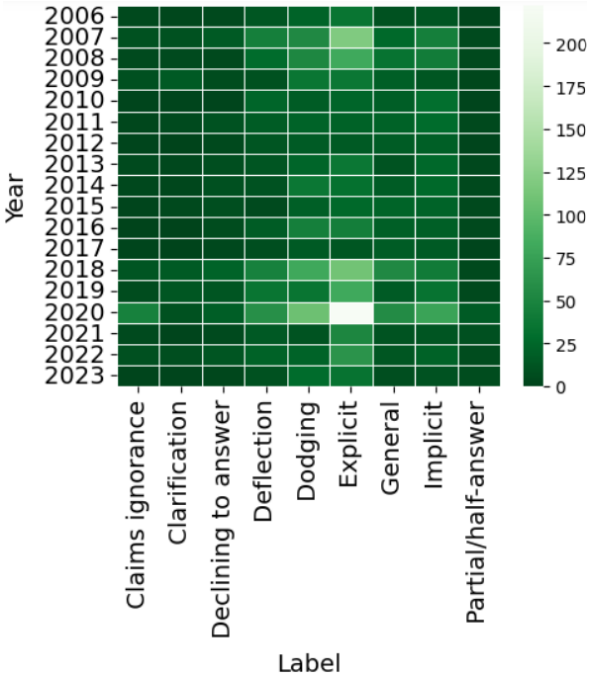


Figure 8: Label distribution across years

So, in association with US presidents, in Figure 9 we demonstrate the timeframe associated with each president's service. We can now conclude that the higher number of Explicit Replies of Figure 8 coincides with Trump's service, which is related to more Explicit Replies, as indicated in Figure 6.

Consequently, temporal evasion characteristics are highlighted in Figure 10.

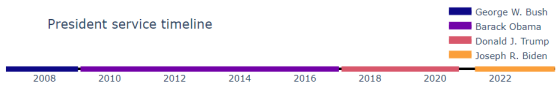


Figure 9: Service timeline for each US president

To this end, some interesting patterns can be derived from Figure 10, especially if we focus on the start and the end of each president's service period. For example, George W. Bush and Joseph R. Biden tend to significantly decrease their ratio of Explicit Replies over implicit replies and evasion strategies, while the opposite pattern occurs for Donald J. Trump. Regarding Barack Obama, his ratio is almost the same at the end of his service in comparison to the beginning, even though fluctuations are observed during his entire service period.

Geographical insights Location-related patterns are examined in Figure 11 in order to derive whether evasion phenomena occur in conjunction to certain locations. Specifically, the horizontal axis represents the location where a presidential speech took place, while the vertical axis corresponds to the percentage of Clear Replies (left), Ambivalent Replies and Clear Non-Replies (right) and the ratio of these two cases (bottom). All percentages are normalized according to the total number of interviews given to each of those locations according to our data. Focusing on the Explicit Reply ratio over all other cases (bottom plot), the resulting long-tailed distribution denotes that in most cases there are few Explicit Replies compared to evasion techniques or Implicit Replies. Overall, we cannot extract a specific pattern location-wise, meaning that the evasion rate is not strongly associated with location.

QA decomposition We also analyze the distribution of sQAs, so that we discover the impact of the number of decomposed QA pairs on other dataset characteristics. This distribution is showcased in Figure 12, where single QA instances dominate the dataset (the highest bar corresponds to 1 sQA, which is equivalent to the initial question and answer, and not decomposed by ChatGPT). As a general tendency, longer QAs -and therefore larger numbers of sQAs- are rare, as proven by the lower bars of Figure 12. This observation eases the annotation process, since longer QA pairs are harder to decompose by ChatGPT, and are consequently evaluated and annotated by humans.

An interesting insight that can be derived from the sQAs count per interview is the corresponding label distribution. This analysis is presented in Figure 13 (we only consider the more frequently occurring sQA numbers as per Figure 13, i.e. instances with 2, 3, 4 sQAs or no sQA as in the case of non-decomposed QA pairs). Interestingly, the top-5 frequent categories are the same for sQAs of counts 2, 3, 4 (Dodging, Implicit, General, Deflection, and Declining to answer categories). Moreover, Explicit Replies are absent from sQAs of count 2, 3, 4, even though they are frequent labels in the dataset (Figure 5). This pattern differs for QA pairs with no decomposition (upper left plot): Explicit Replies are significantly more frequent, followed by other frequently occurring evasion categories (Deflection, General, Dodging). This analysis also suggests an important insight: politicians tend to provide clear replies in answers targeting short, single-barrelled questions while concealing evasion strategies within answers for multi-part questions, where grounding the requested information to the answer given is significantly harder.

Moving forward to a per-president analysis, details regarding the number of questions for all 4 US presidents existing in the interviews under consideration are provided in Figure 14.

We can then proceed by examining the per-president decomposition of questions. The related analysis is presented in Figure 15.

Barack Obama receives more multi-part questions, therefore scoring high in instances where there are 3 or 4 sQAs (bottom plots of Figure 15). This can be possibly related to the elevated number of Ambivalent Replies and low number of Explicit Replies (Figure 2) in association with the connection between evasion frequency and number of sQAs per instance (Figure 13). On the other hand, Donald J. Trump scores higher in instances where single QA pairs occur, or are broken down into 2 parts (2 sQAs), as indicated by the top plots of Figure 15. This could be related to the comparatively lower number of Donald J. Trump Ambivalent replies (Figure 2) and the higher number of Explicit Replies (Figure 6).

To this end, our QA decomposition is deemed as an interesting initial tool towards the possibility of evasions: in cases where many multi-part questions occur, it is possible that evasion strategies may also appear, while the opposite holds in cases with

Figure 10: Percentages of Explicit Replies (left), Implicit/Non-Replies (right) and ratio of Replies over Implicit/Non-Replies (bottom) for each US president during their service.

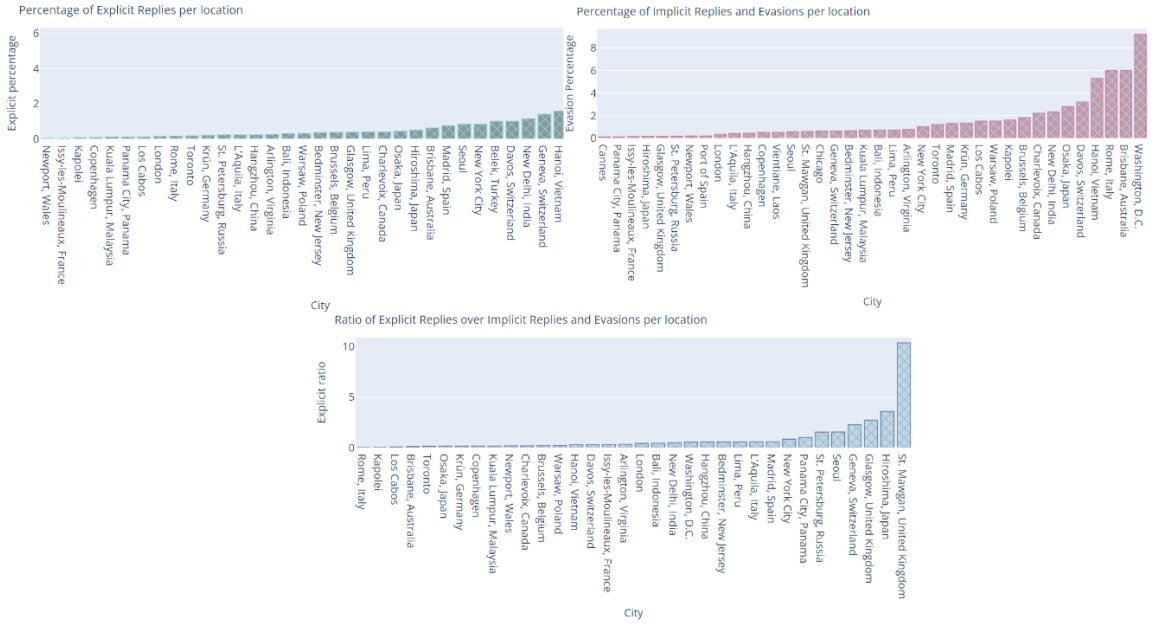


Figure 11: Percentages of Explicit Replies (left), Implicit/Non-Replies (right) and ratio of Replies over Implicit/Non-Replies (bottom) per location.

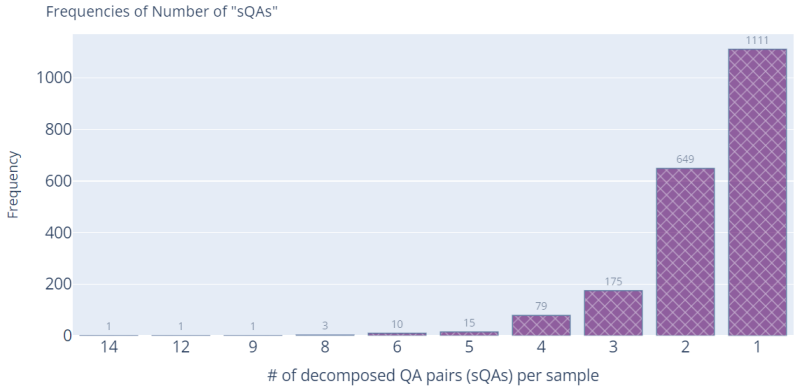


Figure 12: Distribution of sQAs length frequency.

single QA pairs.

Political opponents In Figure 16 we present the distribution of labels when a politician is interviewed on their own versus when they are interviewed with a political opponent. Politicians are more or less consistent towards their Explicit Replies and evasion percentages, as proven by the similar bar height in both cases (with or without an opponent).

Delving deeper into the opponent-related analysis, in Figure 17 we present label percentages with and without political opponent per president. Different patterns arise for each of them: for example, George Bush (Figure 17a) tends to provide more Explicit Replies when being interviewed together with a component than when on his own. On the contrary, Barack Obama (Figure 17b) provides more Explicit Replies when being interviewed on his own. Similarly, Donald J. Trump (Figure 17c) replies explicitly when no opponent is participating in the interview. Smaller differences in Explicit reply percentages under the two interview scenarios are observed for Joseph R. Biden(Figure 17d), even though he tends to provide slightly more Explicit Replies in interviews with a political opponent. Donald J. Trump and Joseph R. Biden tend to employ evasion

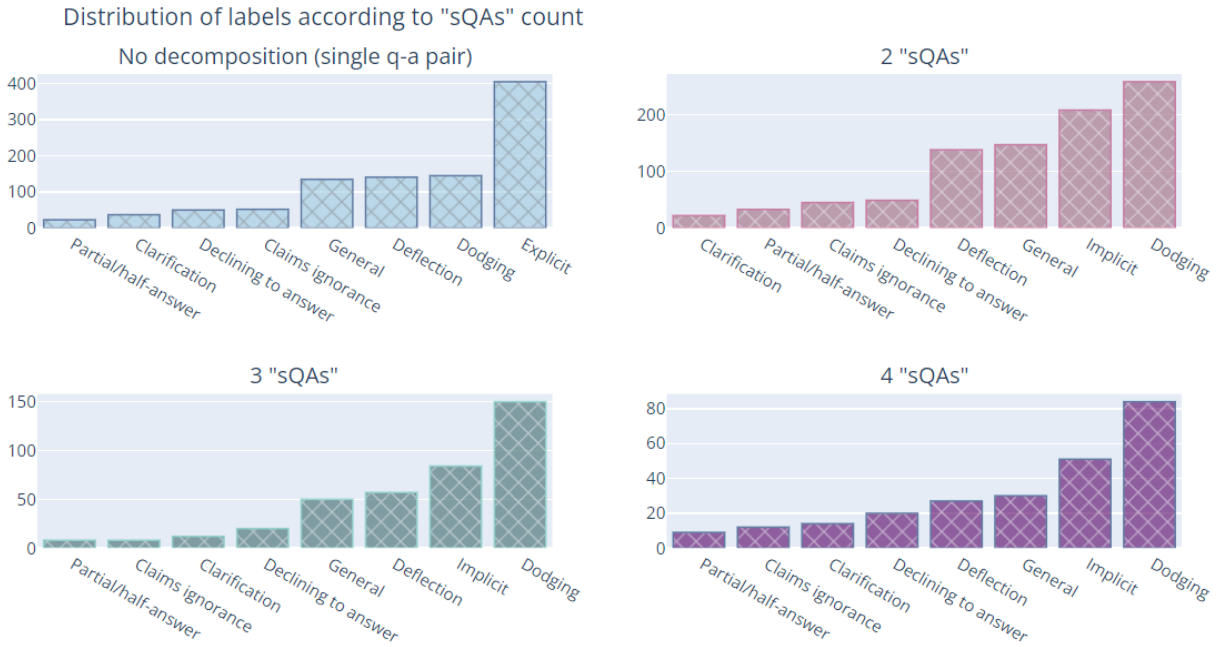


Figure 13: Label frequency per sQAs length.

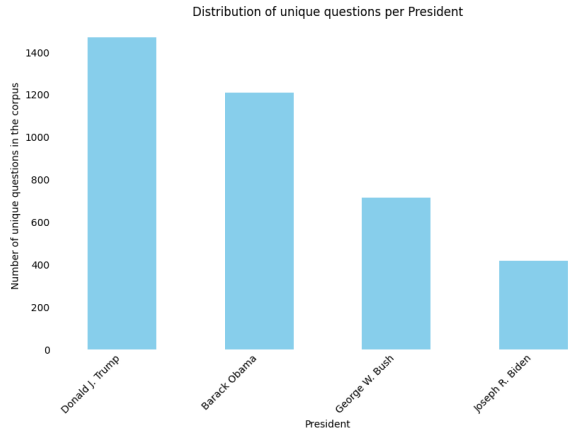


Figure 14: Visualization of distribution of unique questions per President in the corpus

strategies in similar percentages with and without political opponents; some notable exceptions can be observed for Dodging categories, for which the percentages for Biden are higher in presence of a political opponent, while the opposite holds for Trump. In total, the label distributions for Barack Obama and Donald Trump are somewhat similar (note the ranking of labels, as well as the differences between bars with/without opponent), indicating a common behavior in handling interviews with/without political opponents. George Bush holds a diverging distribution, in terms of presenting a larger gap between his top-1 category (Explicit Replies) and the rest; especially when being interviewed on his own, he tends to exploit significantly less evasion techniques in comparison to the rest of the presidents.

Overall, our presented dataset accompanied by this exploratory analysis can be utilized by political scientists, assisting them in extracting interesting insights from political interviews.

A.2 Examples from the proposed taxonomy

In Table 4, 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 classes.

These examples were used in the annotators' "training" phase, during which they were familiarized with the introduced problem, as well as the proposed taxonomy. The same examples were used as demonstrations for few-shot prompting, inserted in the same order as in Table 4.

A.3 Annotation details

Annotators' 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 are females, while the expert annotator

Figure 15: Distribution of per president interviews for different sQA counts.

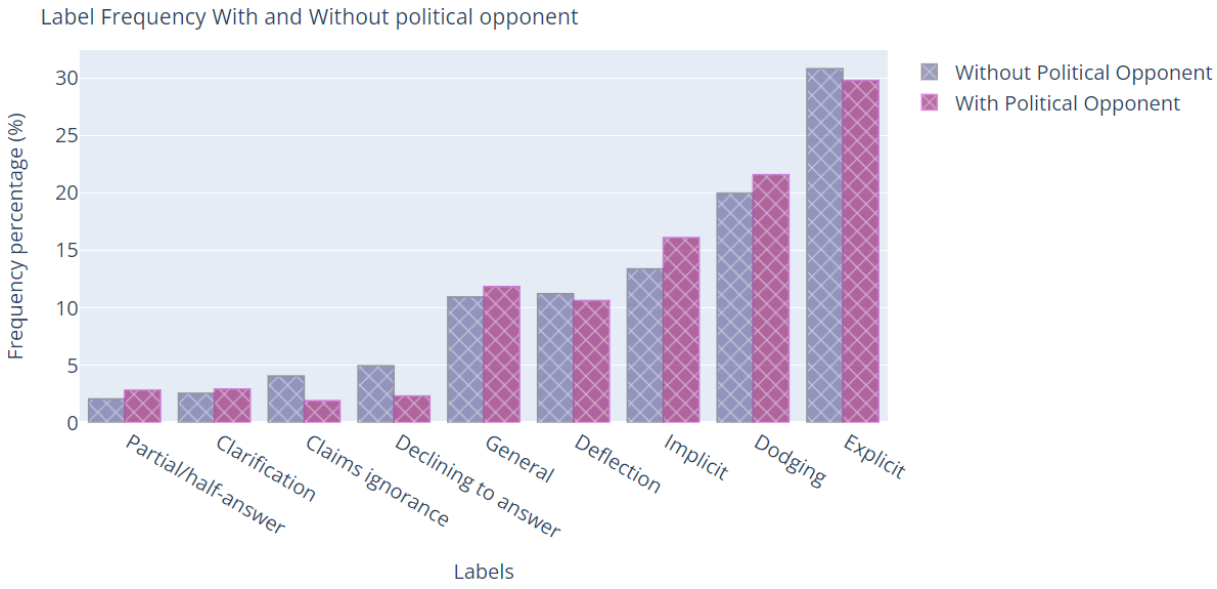
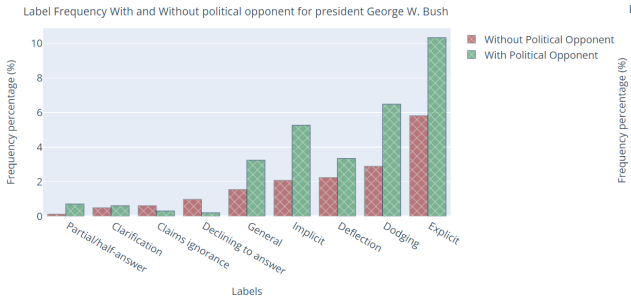


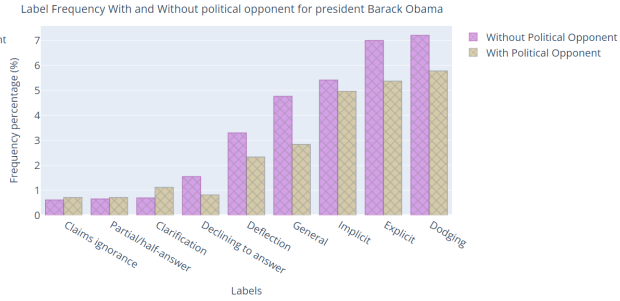
Figure 16: Label percentages for interviews with and without the presence of a political opponent.

is male, and all of them are fluent or native English speakers. We do not disclose geographical characteristics to fully preserve anonymity. Moreover, we did not collect any information regarding age or race/ethnicity.

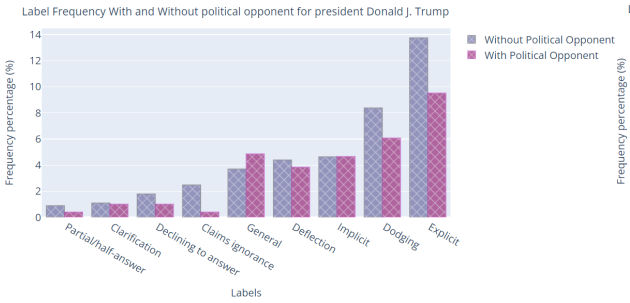
Quality of annotations was ensured via a well-crafted process of designing and monitoring the annotation process. First of all, we collect a descriptive set of instructions: as an introduction, we provided our annotators the examples of Table 4 to familiarize with the nature of the categories. Then, we released a short quiz to validate that they properly learned the fundamentals. After this stage, we proceeded with real examples from our dataset, demonstrating some examples of successful and unsuccessful sQAs in comparison to the initial interviews. Then, we also demonstrated examples with their labels to allow annotators to learn the distinguishing features between each category, especially the usually confused ones (as per Figure 4). Since this step is the most critical for the annotation process, we conducted daily sessions for one week, also distributing short quizzes after each session. The expert monitored and graded the learning process and the quizzes, verifying that the annotators were ready to perform annotations on their own, while also resolving any related questions in



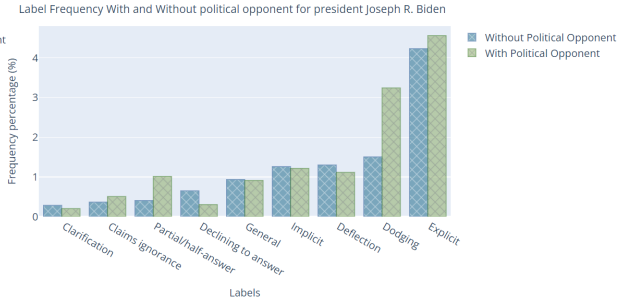
(a) Label distribution for G. Bush.



(b) Label distribution for B. Obama.



(c) Label distribution for D. J. Trump.



(d) Label distribution for J. R. Biden.

Figure 17: Label distribution with and without opponent for each US president of our dataset.

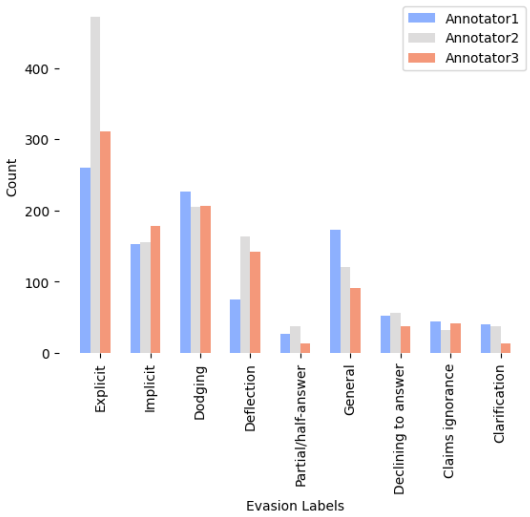


Figure 18: Visualization of distribution of evasion label per annotator in the corpus

Label distribution per annotator Figure 18 depicts the distribution of evasion labels for each non-expert annotator (note that interview samples were randomly distributed to annotators). 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, which is in accordance to the higher number of explicit replies in general (Figure 5).

the meanwhile. Weekly checks on the annotation quality were performed by comparing a subset of the annotations with the annotations provided by the expert. In these intermediate evaluations, no annotator was significantly deviating from the expert. We denote that we consider a non-negligible deviation when the Fleiss score between the expert and any annotator was ≤ 0.7.

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

	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*
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 4: Descriptions and examples of political evasion techniques based on the proposed taxonomy

Labelling platform Our labelling process was conducted in the open source Label Studio[5] platform. We provide some screenshots of the labelling pages in Figures 19, 20 (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.

Annotators have to first evaluate the decomposition quality of sQAs (Figure 19) as provided by ChatGPT. In case of erroneous decomposition, they have to add the corresponding multi-parts missing ("Any Additional Missed Questions?"), among with their taxonomy label. If extraneous multi-parts are generated by ChatGPT, they can be reported (annotators can click the Error button denoting that "Question does not exist in the original text!"), so that this multi-part pair is disabled from the annotation process.

Annotations on presidential speech Extending the findings presented in Figure 2, Table 5 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 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 corre-

[5] https://labelstud.io/

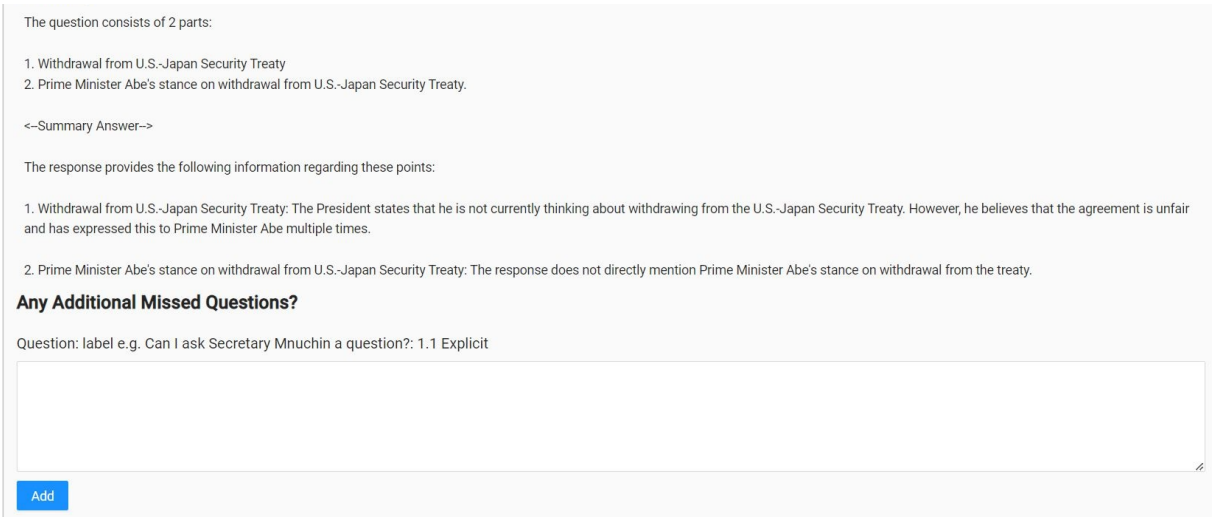


Figure 19: Screenshot from labelling platform: The sQAs for the provided QAs are given to the annotators. They have to highlight each of the enumerated responses and assign one of the labels of the taxonomy (as presented in Figure 20) to each of them.

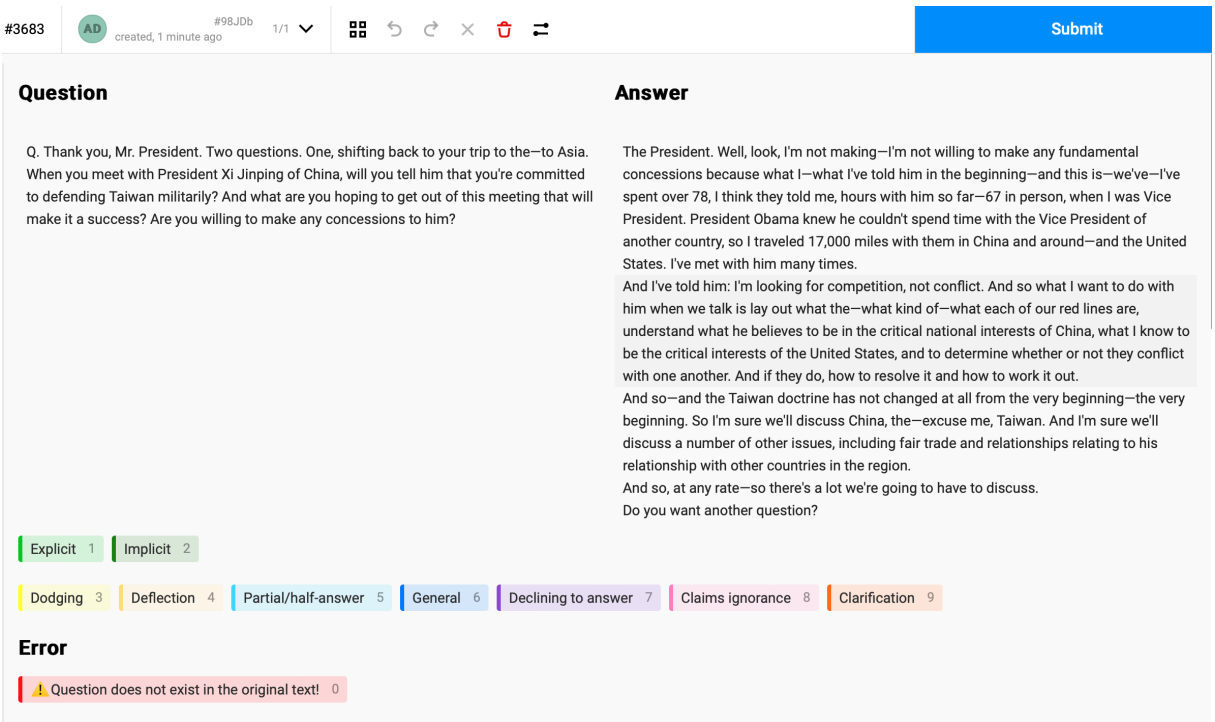


Figure 20: Screenshot from labelling platform: annotators have to read the original Question and Answer as provided. The classes corresponding to our proposed taxonomy are demonstrated as well.

sponds 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.

Dialogue separation The annotators were tasked with identifying potential errors generated by ChatGPT. In Figure 19, they were presented with the option: 'Error, Question does not exist in the original text.' Additionally, if any multi-part pairs were missing, annotators were encouraged to provide them, as shown in Figure 20 with the prompt 'Any additional missing questions?' During the analysis of dialogue separation performed by GPT-3.5-turbo, it was found that 88.6% of the segmented sections were accurately separated, with no errors detected in the sub-questions within the two in-

Response	G. W. Bush	B. Obama	D. J. Trump	J. R. Biden
Clear Reply	34.31	22.38	32.6	37.34
Clear Non-Reply	8.68	9.5	11.77	10.53
Ambivalent	57.0	**68.12**	55.62	52.13
Explicit	34.31	22.38	32.6	**37.34**
Implicit	14.43	18.02	12.08	10.78
Dodging	19.05	23.17	20.08	17.54
Deflection	12.32	10.3	11.02	10.78
Partial/half-answer	1.4	2.28	1.96	5.01
General	9.8	14.36	10.49	8.02
Declining to answer	3.64	4.55	4.08	4.76
Claims ignorance	2.52	2.18	4.91	3.51
Clarification	2.52	2.77	2.79	2.26

Table 5: Statistics of answer clarity and evasion techniques in political interviews per president.

correct segments. Conversely, only 11.4% of the segments contained at least one error in the dialogue separation process. Specifically, 91.41% of the sub-questions were deemed accurate, 7.31% were labelled as 'Error, Question does not exist in the original text,' and 1.27% were initially missing questions that were later provided by the annotators.

Counterfactual Singular QAs (sQAs) Considering that annotators should consult the initial interview text instead of exclusively relying on the more easily readable QA ChatGPT sQAs, we test their cautiousness by inserting 31 additional samples containing counterfactual sQAs in place of the original ones –without them knowing. Those sQAs are purposely unfaithful to the original QAs, guiding an annotator towards believing the responses belong to a different category compared to the actual one. We prompt ChatGPT to select an incorrect (counterfactual) label in order to generate a suitable sQA, which is shown to users instead of the original (the class label is not shown). [6] We manually verify the suitability of each counterfactual sQA. The sQA should be marked as erroneous, and the annotator should write down the decomposed answers occurring, together with their labels.

SQAs insights We computed for each annotator the ratio of selecting the counterfactual label instead of the correct one and found it to be ≤ 0.08. We thus assert that annotators do not solely rely on ChatGPT sQAs and confirm the validity of the process, since they were not significantly influenced by the counterfactual sQAs.

[6] We provide the counterfactual sQA prompt at §H

B Experimental Details

In our experiments, we utilized three distinct datasets: training, development, and validation sets. The original dataset was divided into two parts, allocating 2700 samples to the training set and reserving approximately 750 samples for the development set. For a realistic evaluation, we employed a separate validation dataset comprising 274 samples, which were meticulously annotated by a team of annotators. Any inconsistencies were resolved by a domain expert. This method ensures a robust assessment of the models using ground truth labels validated by an expert. The distribution of each category across these datasets is depicted in Table 6 for clarity labels and Table 7 for evasion labels.

Label	Train	Development	Validation
Clear Reply	796	255	86
Ambivalent Reply	1617	421	207
Clear Non-Reply	284	72	24

Table 6: Distribution of Instances Across Clarity Labels in Training, Development, and Validation Sets.

Label	Train	Validation	Test
Explicit	796	255	90
Implicit	381	107	59
General	313	73	50
Partial/half-answer	69	10	3
Dodging	563	141	61
Deflection	291	90	27
Clarification	69	23	4
Declining to answer	117	28	11
Claims ignorance	98	21	10

Table 7: Distribution of Instances Across Evasion Labels in Training, Validation, and Testing Sets.

B.1 Evaluation

Throughout our paper, we utilize classification metrics for evaluation. Specifically, accuracy, precision, and recall are employed, as well as F1 scores. Regarding F1, we use both the macro and the weighted average strategies. The macro F1 score is calculated as the average of the F1 scores for each class (see Eq. 1), without considering the class distribution, whereas the weighted F1 score accounts for class frequency, giving more weight to larger classes (see Eq. 2).

$$F1_{macro} = \frac{1}{N}\sum_{i=1}^{N} F1_i \quad (1)$$

Classification variant	Model	Acc.	Prec.	Recall	F1
direct clarity	Llama-7b	0.489	0.581	0.489	0.504
	Llama-13b	0.587	0.719	0.587	0.594
	Llama-70b	**0.75**	**0.757**	**0.75**	**0.75**2
	Falcon-7b	0.294	0.537	0.294	0.186
	Falcon-40b	0.341	0.656	0.341	0.244
evasion-based clarity	Llama-7b	0.662	0.669	0.662	0.665
	Llama-13b	0.675	0.68	0.675	0.677
	Llama-70b	0.713	0.743	0.713	0.72
	Falcon-7b	0.533	0.537	0.533	0.533
	Falcon-40b	0.618	0.633	0.618	0.622

Table 8: Classification results using a weighted strategy, which averages F1 scores, weighted by class size. The best results for each strategy are underlined and the best results overall are also in **bold**.

$$F1_{weighted} = \sum_{i=1}^{N} \left(\frac{n_i}{N} \times F1_i \right), \qquad (2)$$

where n_i is the number of instances in each class.

C Performance Analysis for Each Class

In this section, the performance of the instruction-tuned models, which have shown the best performance compared to other strategies, is presented by class. Table 8 illustrates the performance of these models using a weighted strategy.

Using the weighted strategy, the conclusions remain the same, although the numerical results are slightly improved. Further analysis of the model's performance for each class can be found in Table 17, which showcases the classification report of the tuned Llama-2-70b model with evasion-based clarity for each class, which has shown the best results among the other strategies.

	Prec.	Recall	F1	Sup.
Clear Reply	0.54	0.74	0.62	84
Ambivalent	0.84	0.71	0.77	208
Clear Non-Reply	0.63	0.68	0.65	25
Acc.			0.71	317
Macro avg	0.67	0.71	0.68	317
Weighted avg	0.74	0.71	0.72	317

Table 9: Classification report of the tuned Llama-2-70b model, for each class, demonstrating precision, recall, F1 score, and support.

Notably, the model demonstrates its highest precision with the Ambivalent category at suggesting strong accuracy in identifying relevant instances, albeit with a moderate recall. This is followed by a decent performance in the Clear Non-Reply category, with a balanced precision and recall. The category Clear Reply, while having a high recall, indicating effective identification of most relevant cases, shows the lowest precision, which may indicate a higher rate of false positives. This issue particularly arises from confusion between Clear Replies and Ambiguous responses, and between Clear and General responses, as further analyzed in App. E.

Overall, the model achieves a general accuracy of and similarly balanced macro and weighted average scores. These results indicate a reasonably good model performance, particularly in distinguishing the more frequently occurring Ambivalent category.

D Additional Experiments

D.1 Few-Shot prompting

In the few-shot setup, we showcase the model results irrespective of their size. Unlike in the ZS setup, smaller models demonstrated better adherence to the output template and exhibited fewer hallucinations overall. Since the examples in our dataset are quite lengthy, we opt to select one example for each label to present to the model, along with the corresponding explanation provided in Table 10. This methodology mirrors what the human annotators saw before commencing the annotation procedure. We noticed that Falcon struggled more to respond within the given template compared to the zero-shot approach. Nevertheless, examples in the few-shot setup seemed to aid the Llama-70b model in understanding the task, along with the smaller models. In the FS setup, the Llama-7b model exhibited comparable results to a model ten times larger in the ZS setup. In evasion-based clarity models, examples in the middle are often ignored. Instead, responses tend to align with the labels of the first or last examples. This phenomenon is well-documented in literature (Dong et al., 2022). For example in Llamma-70b, 60% of responses matched the labels of the final four examples, compared to less than 10% in the ground truth.

D.2 Answer Grounding

In this section, we outline the distinctions in model performance between single and multi-part questions. Specifically, we divided the test set into two distinct parts: one consisting of segments of the interview containing only single questions (112 out

Classification variant	Model	Acc.	Prec.	Recall	F1
direct clarity	Llama-7b	0.23	0.159	0.474	0.219
	Llama-13b	0.211	0.105	0.302	0.156
	Llama-70b	**0.667**	0.333	0.333	0.333
	Falcon-7b	0.203	0.107	0.267	0.152
	Falcon-40b	0.29	0.13	0.336	0.186
evasion-based clarity	Llama-7b	0.274	0.393	0.335	0.262
	Llama-13b	0.291	0.452	0.363	0.259
	Llama-70b	0.541	**0.565**	**0.452**	**0.365**
	Falcon-7b	0.505	0.299	0.211	0.222
	Falcon-40b	0.429	0.167	0.25	0.2

Table 10: Classification results for few-shot (FS) inference. The best results for each strategy are underlined and best results overall are also in **bold**.

Classification variant	Model	Acc.	Prec.	Recall	F1
direct clarity	Llama-7b	0.47	0.403	0.48	0.402
		0.53	0.537	0.588	**0.538**
	Llama-13b	0.59	0.547	0.711	0.548
		0.59	0.625	0.694	**0.621**
	Llama-70b	0.74	0.594	0.648	0.612
		0.78	0.705	0.742	**0.72**
	Falcon-7b	0.25	0.319	0.337	0.158
		0.37	0.341	0.329	**0.21**
	Falcon-40b	0.29	0.432	0.468	0.284
		0.44	0.67	0.629	**0.459**
evasion-based clarity	Llama-7b	0.67	0.593	0.59	0.591
		0.64	0.602	0.622	**0.608**
	Llama-13b	0.69	0.592	0.581	0.586
		0.64	0.635	0.679	**0.654**
	Llama-70b	0.7	0.601	0.656	0.62
		0.73	0.75	0.785	**0.761**
	Falcon-7b	0.54	0.442	0.372	0.384
		0.52	0.429	0.413	**0.418**
	Falcon-40b	0.64	0.62	0.47	0.493
		0.58	0.578	0.598	**0.586**

Table 11: Classification results for instruction-tuned models. The best results overall are in **bold**. The first line of each model shows the results for the set containing only multi-part questions, while the second line shows the results for single-part questions.

Classification variant	Model	Acc.	Prec.	Recall	F1
direct clarity	zero-shot	0.668	0.418	0.37	0.37
		0.625	**0.559**	**0.483**	**0.461**
	standalone CoT	0.649	0.347	0.34	0.332
		0.607	**0.537**	**0.441**	**0.418**
evasion based clarity	zero-shot	0.639	0.443	0.442	0.436
		0.661	**0.683**	**0.603**	**0.56**
	standalone CoT	0.712	0.568	0.483	0.489
		0.643	**0.657**	**0.558**	**0.536**

Table 12: Classification results for ChatGPT using zero-shot and chain-of-thought inference for the two subsets (single- and multi-part questions). The best results for each subset are in **bold**. The first line of each model shows the results for the set containing only multi-part questions, while the second line shows the results for single-part questions.

of 317 questions), and the other containing only segments with multi-part questions (205 out of 317 questions). We then compared the performance of each method. Using this methodology, we discovered that regardless of the method employed, every model exhibited lower performance on multi-part questions compared to single ones. The results for instruction-tuned models are shown in Table 11, while those for the prompting techniques applied to the model with the best results are presented in Table 12. For each model or method, there are two lines: the first represents performance on the multi-part question set, and the second represents performance on the single question set.

To further investigate whether this difficulty is also encountered by humans, we compared the Fleiss score of the annotators between these two subsets. We found that the difference was only 0.03, indicating that there was no significant difference in the performance of annotators between single and multi-part questions. This suggests that the challenge of grounding answers to multi-part questions is unique to LLMs.

D.3 Connection to encoded knowledge

We further delve into the integral relationship between clarity classification and the knowledge pertaining to a specific named entity. Named entities frequently have properties that are considered common knowledge and that is why they are not explicitly mentioned in a response. As a result, the systems that try to define the clarity of a response would need to be aware of these properties of the name entities. In our dataset the most occurring named entities are persons' names, that why we focused the experimental analysis on these terms. Specifically, we split our dataset into two distinct parts, one containing only parts of the interview that include at least one person's name either in the interview question or the answer and a second one which contains no person names. The first set consists of 189 questions and the second of 128 questions. The differences between the performances for instruction-tuned models are shown

Classification variant	Model	Acc.	Prec.	Recall	F1
direct clarity	Llama-7b	0.434	0.375	0.439	0.369
		0.57	**0.56**	**0.62**1	**0.565**
	Llama-13b	0.55	0.527	0.663	0.52
		0.639	**0.631**	**0.731**	**0.638**
	Llama-70b	0.752	0.65	0.777	0.69
		0.768	**0.7**	**0.686**	**0.692**
	Falcon-7b	0.266	0.255	0.319	0.148
		0.32	**0.348**	**0.355**	**0.213**
	Falcon-40b	0.328	0.489	0.504	0.331
		0.359	**0.533**	**0.55**	**0.374**
evasion based clarity	Llama-7b	0.635	0.57	0.557	0.563
		0.711	**0.67**	**0.678**	**0.673**
	Llama-13b	0.651	0.573	0.611	0.589
		0.711	**0.674**	**0.636**	**0.653**
	Llama-70b	0.709	0.637	0.706	0.661
		0.719	**0.701**	**0.718**	**0.702**
	Falcon-7b	0.497	0.387	0.319	0.332
		0.586	**0.488**	**0.473**	**0.473**
	Falcon-40b	0.598	0.531	0.45	0.468
		0.656	**0.665**	**0.601**	**0.622**

Table 13: Classification results for instruction-tuned models. The best results overall are in **bold**. The first line of each model shows the results for the subset consisting exclusively of instances that contain named entities, while the second line shows the results for the subset without named entities.

Classification variant	Model	Acc.	Prec.	Recall	F1
direct clarity	zero-shot	0.651	0.416	0.371	0.354
		0.641	**0.53**	**0.449**	**0.463**
	standalone CoT	0.614	0.333	0.326	0.311
		0.648	**0.518**	**0.429**	**0.434**
evasion based clarity	zero-shot	0.635	0.457	0.44	0.42
		0.648	**0.559**	**0.532**	**0.536**
	standalone CoT	0.712	0.568	0.483	0.489
		0.677	**0.657**	**0.535**	**0.551**

Table 14: Knowledge-related classification results for ChatGPT using zero-shot and chain-of-thought inference for the two subset. The best results for each subset are in **bold**. The first line of each model shows the results for the subset consisting exclusively of instances that contain named entities, while the second line shows the results for the subset without named entities.

in Table 13, while those for the prompting techniques applied to the model with the best results are presented in Table 14.

The results show that across all models and methods, the performance on the set without named entities is increased compared with the performance on the set with named entities. Notably, there was a steep improvement in the smaller, less knowledgeable models compared to the others, corroborating the findings of (Sun et al., 2023). In this case, if we apply the same comparison for the human-curated annotations, we can see that there was a difference of 0.1 in Fleiss score between the two subsets, implying that it was slightly more difficult for humans also to annotate the set with named entities compared to the other one.

E Evasion classification

In this section, we present the results of the evasion (low-level) classification problem. Table 15 illustrates the performance of the instruction-tuned model on the evasion classification problem, while Table 16 showcases the performance using zero-shot and chain-of-thought prompting on the ChatGPT which is the best-performing model. The performance of the models on the evasion classification task is lower compared to the clarity classification. Among the instruction-tuned models, Llama-70b exhibits the best performance across all metrics, similar to the evasion classification model.

In ChatGPT, a higher level of performance is observed in the zero-shot setup compared to the chain-of-thought (CoT) for evasion classification, contrary to the evasion-based classification method. Further investigation reveals that employing CoT ChatGPT leads to greater confusion between the classes *General* and *Implicit*, as well as *Implicit* and *Partial/half-answer*, compared to the zero-shot setup, where the primary confusion lies between *Partial/half-answer* and *Explicit*. However, the confusion stemming from the zero-shot setup results in different clarity labels, unlike CoT, which elucidates the performance disparity between the two tasks. It is noteworthy that the challenge of discriminating between these classes persists even for humans, as evidenced by the lowest agreement between annotators for these labels, as indicated in Figure 4. This underscores a general difficulty in distinguishing between these two evasion strategies. This analysis is particularly intriguing, especially given the context where the model has not been exposed to the annotated data of the users.

In order to evaluate the performance of the models at the evasion level, Table 17 displays the classification report of the best performing model, Llama-70b.

The results indicate varying performance across different response types in the model's classification capabilities. For example, the "Explicit" category shows strong performance, resulting in a

Model	Acc.	Prec.	Recall	F1
LLama-7b	0.454	0.498	0.458	0.444
LLama-13b	0.464	0.429	0.49	0.423
LLama-70b	**0.571**	**0.571**	**0.558**	**0.545**
Falcon-7b	0.363	0.226	0.216	0.212
Falcon-40b	0.476	0.558	0.475	0.492

Table 15: Classification results for instruction-tuned models for the evasion classification. The best results are in **bold**.

Model	Acc.	Prec.	Recall	F1
zero-shot	**0.315**	0.266	**0.284**	**0.244**
standalone CoT	0.259	**0.293**	0.279	0.229

Table 16: Classification results for evasion classification using zero-shot and chain-of-thought for prompting chatGPT which is best performing model using only prompting techniques. The best results are in **bold**.

relatively high F1-score, which suggests the model is quite effective at identifying and correctly classifying explicit responses. In contrast, the "Implicit" and "Deflection" categories exhibit lower precision and recall, indicating challenges in accurately detecting and classifying these subtler forms of responses, similar to human annotators, as depicted in Table 4. Notably, the "Clarification" category achieved perfect precision but lower recall, highlighting that while the model is highly accurate when it identifies these responses, it consistently fails to detect them.

F Encoder models

In this section, to evaluate the performance of smaller models on the proposed task, we trained three different architectures: DeBERTa (He et al., 2021), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019), and assessed their performance on the same test set. Specifically, we selected two different sizes for each model: base and large, to examine the impact of size variation on model performance. The primary challenge we encountered was truncation, as the maximum input size for DeBERTa and RoBERTa is 512 tokens. To ensure a fair comparison, we also utilized XLNet, which does not have inherent input size limits. We fine-tuned these models using only non-truncated inputs to reduce noise during training. Specifically, out of the total 2700 samples in the training set, only 1713 (63%) had fewer than 512 tokens. We trained the models for five epochs with a constant learning rate of 10^{-5}. Evaluation of the models was

	Prec.	Recall	F1	Sup.
Explicit	0.68	0.84	0.75	94
Implicit	0.50	0.29	0.36	64
Dodging	0.53	0.68	0.59	60
Deflection	0.33	0.45	0.38	20
Partial/half-answer	0.00	0.00	0.00	6
General	0.55	0.37	0.44	49
Declining to answer	0.46	0.60	0.52	10
Claims ignorance	0.67	0.80	0.73	10
Clarification	1.00	0.50	0.67	4
Acc.			0.57	317
Macro avg	0.57	0.50	0.51	317
Weighted avg	0.56	0.57	0.55	317

Table 17: Classification report of the tuned Llama-2-70b model, for each class, demonstrating precision, recall, F1 score, and support.

conducted using the same test set, without removing 173 out of 317 samples with more than 512 tokens. The evaluation results are presented in Table 20, while Table 21 displays the results of the same models on the subset of 173 samples with non-truncated inputs. For comparison, the results of the instruction-tuned LLama models on this subset are also included. As shown in Table 22, the performance of the models on the subset with truncated inputs is close to random chance.

Another noteworthy finding is that the base models consistently outperformed their respective larger counterparts. Specifically, the output of every large model collapsed to a single label. For instance, RoBERTa-large with evasion-based clarity returned the label "Explicit" for every sample. Similar behaviour was observed for every large variant of the three different models.

To further evaluate the behaviour of encoder models and to explain their performance, we again check the differences in performance between the set of entities with named entities and without. The results are shown in Table 18. The first line of each model displays the results for the set containing only interview parts with named entities, while the second line shows the results for the parts without named entities. The 'large' variations of the models were omitted as they returned only a single class regardless of their input. This shows that the performance of encoders in the subset without named entities was improved for every model, regarding the classification strategy. Again, we evaluate the performance of the encoder in the subset and single-part questions, and the results are depicted in Table 19. The results show that the performance of the models in the subset that contains

Classification variant	Model	Acc.	Prec.	Recall	F1
direct clarity	DebERTa-base	**0.562**	**0.521**	**0.467**	**0.465**
		0.593	0.512	0.439	0.416
	RoBERTa-base	**0.625**	**0.614**	**0.593**	**0.592**
		0.651	0.383	0.405	0.392
	XLNet-base	**0.68**	**0.557**	**0.571**	**0.56**
		0.704	0.481	0.468	0.472
evasion based clarity	DebERTa-base	**0.57**	**0.576**	**0.645**	**0.568**
		0.545	0.498	0.715	0.509
	RoBERTa-base	**0.539**	**0.55**	**0.581**	**0.543**
		0.603	0.401	0.439	0.397
	XLNet-base	**0.594**	**0.552**	**0.617**	**0.574**
		0.571	0.49	0.541	0.51

Table 18: Classification results for encoders. The best results overall are in **bold**. The first line of each model shows the results for the set containing only interview parts that contains named entities, while the second line shows the results for the parts withouts named entities.

Classification variant	Model	Acc.	Prec.	Recall	F1
direct clarity	DebERTa-base	**0.615**	**0.508**	**0.469**	**0.44**
		0.518	0.538	0.438	0.429
	RoBERTa-base	0.629	0.482	0.437	0.438
		0.661	**0.649**	**0.595**	**0.612**
	XLNet-base	0.702	0.45	0.453	0.442
		0.679	**0.626**	**0.588**	**0.604**
evasion based clarity	DebERTa-base	0.576	0.492	0.685	0.51
		0.518	**0.624**	**0.64**	**0.563**
	RoBERTa-base	0.561	0.369	0.4	0.361
		0.607	**0.618**	**0.651**	**0.613**
	XLNet-base	0.527	0.413	0.479	0.43
		0.679	**0.707**	**0.706**	**0.706**

Table 19: Classification results for encoders. The best results overall are in **bold**. The first line of each model shows the results for the set containing only multi-part questions, while the second line shows the results for single-part questions.

multipart questions is near to random chance, probably due to increased input size which increases the probability of truncation. This behaviour is consistent even for the XLNet model, where there is no length restriction in their input, so truncation does not occur. However, an interesting observation is that for single-part questions, the models, especially RoBERTa and XLNet, have comparable performance with generative models such as Llama-70b.

G Comparison with Relevant Tasks

In this section, we compare the focus of our work to the closely related work of Ferracane et al. (2021). The relevance of this analysis stems from the general similarity between our analysis and theirs, despite the diverging task objectives: in our work,

Classification variant	Model	Acc.	Prec.	Recall	F1
direct clarity	DeBERTa-base	0.58	0.521	0.453	0.441
	DeBERTa-large	0.691	0.23	0.333	0.272
	RoBERTa-base	0.64	0.579	0.516	0.53
	RoBERTa-large	0.593	0.198	0.333	0.248
	XLNet-base	0.694	0.52	0.523	0.518
	XLNet-large	0.565	0.188	0.333	0.241
evasion based clarity	DeBERTa-base	0.555	0.53	0.671	0.537
	DeBERTa-large	0.249	0.083	0.333	0.133
	RoBERTa-base	0.577	0.501	0.534	0.495
	RoBERTa-large	0.278	0.093	0.333	0.145
	XLNet-base	**0.58**	**0.523**	**0.586**	**0.546**
	XLNet-large	0.385	0.128	0.333	0.185

Table 20: Classification results for fine-tuned encoder models on the test set. The best results for each strategy are underlined and best results overall are also in **bold**.

Classification variant	Model	Acc.	Prec.	Recall	F1
direct clarity	DeBERTa-base	0.572	0.548	0.469	0.469
	DeBERTa-large	0.647	0.216	0.333	0.262
	RoBERTa-base	0.595	0.569	0.524	0.524
	RoBERTa-large	0.566	0.189	0.333	0.241
	Llama-7b	0.506	0.49	0.529	0.495
	Llama-13b	0.673	0.657	0.74	0.67
	Llama-70b	**0.775**	**0.743**	**0.724**	**0.732**
evasion based clarity	DeBERTa-base	0.561	0.568	0.664	0.569
	DeBERTa-large	0.254	0.085	0.333	0.135
	RoBERTa-base	0.555	0.538	0.548	0.512
	RoBERTa-large	0.277	0.092	0.333	0.145
	Llama-7b	0.678	0.651	0.624	0.633
	Llama-13b	0.707	0.692	0.646	0.665
	Llama-70b	0.724	0.695	0.702	0.698

Table 21: Classification results for fine-tuned encoder models on the 173 samples of the test set that the input was not truncated. For comparison reasons the table is also depicted the performance of the instruction tuned LLama for this subset. The best results for each strategy are underlined and best results overall are also in **bold**.

Classification variant	Model	Acc.	Prec.	Recall	F1
direct clarity	DeBERTa-base	0.59	0.381	0.343	0.309
	DeBERTa-large	0.743	0.248	0.333	0.284
	RoBERTa-base	0.694	0.403	0.41	0.406
	RoBERTa-large	0.625	0.208	0.333	0.256
evasion based clarity	DeBERTa-base	**0.549**	**0.44**	**0.734**	**0.424**
	DeBERTa-large	0.243	0.081	0.333	0.13
	RoBERTa-base	0.604	0.392	0.404	0.383
	RoBERTa-large	0.278	0.093	0.333	0.145

Table 22: Classification results for fine-tuned encoder models on the 144 samples of the test set that the input was truncated. The best results for each strategy are underlined and best results overall are also in **bold**.

we detach our analysis from intents or factuality of question, providing a strict formulation of evasion strategies. To this end, unanswered false presuppositions are not necessarily connected to the intent to deceive. We made this selection not only in order to differentiate from Ferracane et al. (2021), but also to restrict the large set of possible interpretations arising under varying intents. For example, a question containing a false premise, such as "Why is the earth flat?" accompanied with a response "The earth is not flat." does not receive the information requested -the reason *why* the earth is flat- but rather utilizes a factual statement -the earth is *scientifically proven* **not** to be flat- to form the response, which can be classified as an Ambivalent Reply. In case the question contained a valid statement (e.g. "Why is the earth round?") a similarly formatted reply ("The earth is not round") would be again classified as Ambivalent Reply in terms of the information provided, even though it reflects reduced factual knowledge or an intent to deceive from the interviewer's side. However, recognizing intents can be subjective and highly variable, while measuring the degree and the type of information provided, as in our work, formulates a more deterministic and strict framework. At the same time, we do not require detailed knowledge of the facts contained in the question, which may be unavailable even to audience with related background; a separate factuality analysis would reveal potential knowledge gaps highlighting possible interpretations of the question at hand. Overall, our annotated responses contain a specific label regardless the intent and the factuality of the question.

We will further analyze the performance of our models using the dataset referenced in (Ferracane et al., 2021). By applying our models to their dataset, we aim to assess their generalizability across varied contexts. It is important to note that while both datasets predominantly cover the political domain and include press conferences of U.S. Presidents, their formulations are markedly distinct. Specifically, the dataset in (Ferracane et al., 2021) is defined by its goal to determine not only if respondents intend to answer questions but also if their responses are truthful. This subjective approach necessitates a multi-label problem framework where instances might receive conflicting labels, such as "Can't answer Sincere" and "Can't answer Lying." This complexity arises when one annotator perceives deception, while another believes in the sincerity of the response. However, more complex situations may arise, such as when one annotator labels an instance as "Answer" and another labels it as "Can't Answer - Lying." This variation indicates that differences in perceived intent and truthfulness can completely alter the label concerning the answerability of the response, contrary to expectations.

Contrastingly, our model's framework does not consider the intent or truthfulness of responses, focusing solely on whether the response addresses the question. Discrepancies in labeling by annotators are resolved by an expert, streamlining the process and ensuring each instance maintains a singular, clear label. This approach aligns with our primary objective: determining the direct answerability of responses, irrespective of underlying intentions or truthfulness.

Further, we seek to evaluate the efficacy of our top-performing model, trained on our dataset, on the dataset proposed in (Ferracane et al., 2021). Initially, we eliminate all duplicate entries, then process the remaining data through the Llama-70b model, which was trained using evasion-based direct clarity strategies. Figures 21 and 22 illustrate the comparison between the ground truth and our predicted labels across the training and development sets. This comparison is crucial, especially considering the development set's relatively small size—it comprises fewer than 200 instances across 27 labels, with some labels lacking adequate representation.

Firstly, it is evident that this dataset is also highly unbalanced, with 'Answer' being the most frequently occurring label, similar to our own dataset. Additionally, there is a clear alignment between the predicted labels using our taxonomy and the ground truth labels. For instance, instances labeled with "shift-dodge & can't answer lying" are predominantly classified under one of the corresponding labels from our taxonomy, such as "Declining to answer," "Claims ignorance," or "Dodging." To provide a quantifiable measure of the model's performance across both tasks, we evaluate the model's effectiveness solely on instances that have a single ground truth label in both sets, as shown in Table 23, employing a weighted average strategy.

The results indicate that our model can generalize effectively, performing well on a dataset annotated with a different strategy. However, it is important to note that the improved outcomes on this dataset, compared to our own, might be attributed to instances having clear and consistent

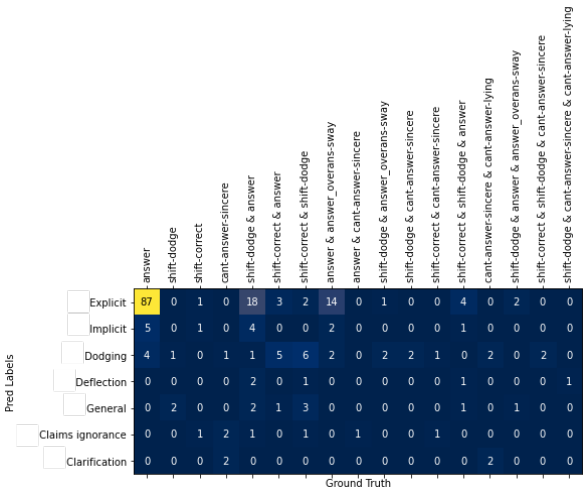


Figure 21: Results of Llama-70b trained using the evasion based clarity for dev set of (Ferracane et al., 2021).

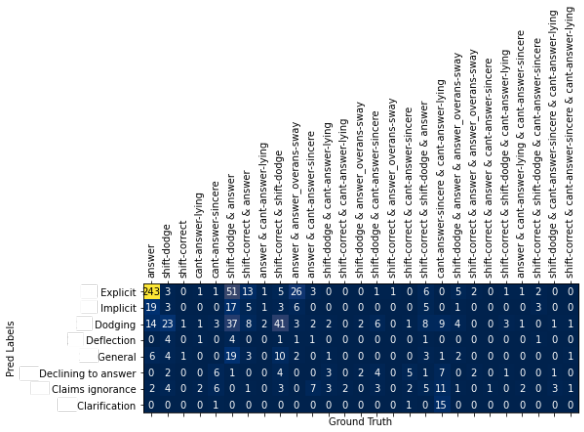


Figure 22: Results of Llama-70b trained using the evasion based clarity for the training set of (Ferracane et al., 2021).

	Acc.	Prec.	Recall	F1
Dev	0.85	0.89	0.85	0.87
Train	0.81	0.85	0.81	0.82

Table 23: The performance of the Llama-70b trained using the evasion based clarity, on development and training sets.

answers across different annotators, suggesting a higher clarity in these instances. Finally, Figures 23 and 24 display the confusion matrices comparing the ground truth with our results for instances with single labels.

H Prompting details

Prompt for generating sQAs The following prompt was provided to ChatGPT to obtain the sQAs of the multi-part 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] ‘‘‘‘‘‘

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-sQAs In addition to this prompt, we create some “counter-sQAs“ to assess the annotators’ reliance on the extracted sQAs rather than the original multi-part pairs as provided in the interviews. The following prompt was appended to the previous one:

message_3 = ‘‘‘‘‘‘

Now, try to create an QAs 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 ’Explicit,” generate an sQA that could make someone believe it is a “General” response or any other label of your choice. The sQA should be at the same length as the original one. Start by selecting the counterlabel and then write the sQs 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:

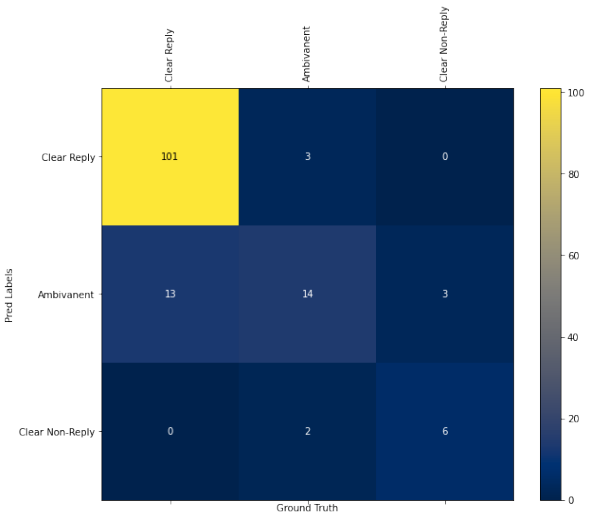


Figure 23: Confusion matrix of Llama-70b trained using the evasion based clarity for dev set of (Ferracane et al., 2021) for the single labelled instances.

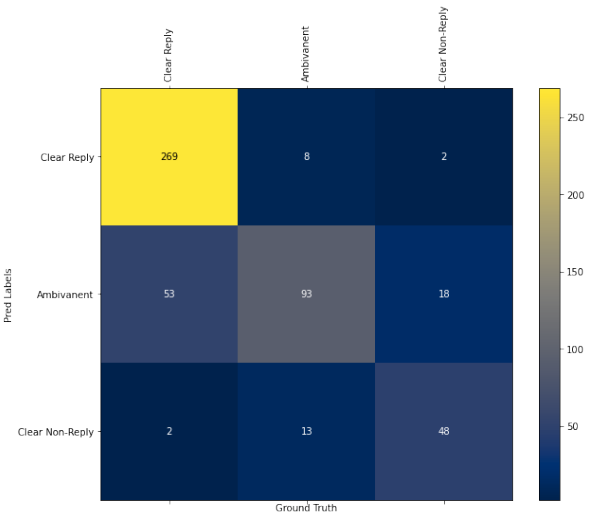


Figure 24: Results of Llama-70b trained using the evasion based clarity for the training set of (Ferracane et al., 2021) for the single labelled instances.

“““

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. Ambiguous - The information requested is given in an incomplete way e.g. the answer is too general, partial, implicit, 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 - The information requested is given in an incomplete way e.g. the answer is too general, partial, implicit, 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: """

Few-Shot (FS) prompt for classification The following prompt was used for addressing the evasion problem in the FS 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:

<Taxonomy>

Here is one small example for each term of the taxonony:
Question:
Do you have your own views about PR at Westminster don't you?
Answer:
I do.
Label: Explicit
Explanation: The answer directly gives the info requested.
Question: Are you going to watch television?
Answer: What else is there to do?
Label: Implicit
Explanation: They suggest planning to watch TV, despite not explicitly stating it.
Question: Do you like my new dress?
Answer: We are late.
Label: Dodging
Explanation: Does not even acknowledge the question and goes straight to another topic.
Question: Did you eat the last piece of pie?
Answer: I have to admit that this was a great recipe, I always like it when there are chocolate chips in the dough.
Label: Deflection
Explanation: Acknowledges the question but goes on a tangent about the chips, without answering.
Question: Did you enjoy the film?
Answer: The directing was great.
Label: Partial/half-answer
Explanation: Directing is only part of what constitutes a film.
Question: What's your favorite film?
Answer: Fight Club, Filth, and Hereditary.
Label: General
Explanation: The reply gives three movies instead of one, which makes the desired information unclear.
Question: The hypothesis I was discussing, wouldn't you regard that as a defeat?
Answer: I am not going to prophesy what will happen.
Label: Declining to answer
Explanation: Directly stating they won't answer.
Question: 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?
Answer: I do not know that date. I will find out and let the House know.
Label: Claims ignorance
Explanation: Claims/admits they don't have the information.
Question: Was it your decision to release the fund?
Answer: You mean the public fund?
Label: Clarification
Explanation: Gives no data, asks for clarification.
Part of the interview
<Part of the interview>
Question
<Question>
Taxonomy code: """

The following prompt was used for addressing the clarity problem in the FS 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:
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 - The information requested is given in an incomplete way e.g. the answer is too general, partial, implicit, dodging or deflection

Here is one small example for each term of the taxonony:
Question:
Do you have your own views about PR at Westminster don't you?
Answer: I do.
Label: Clear Reply
Explanation: The answer directly gives the info requested.
Question: Are you going to watch television?
Answer: What else is there to do?
Label: Ambivalent
Explanation: They suggest planning to watch TV, despite not explicitly stating it.
Question: Do you like my new dress?
Answer: We are late.
Label: Ambivalent
Explanation: Does not even acknowledge the question and goes straight to another topic.
Question: Did you eat the last piece of pie?
Answer: I have to admit that this was a great recipe, I always like it when there are chocolate chips in the dough.
Label: Ambivalent
Explanation: Acknowledges the question but goes on a tangent about the chips, without answering.
Question: Did you enjoy the film?
Answer: The directing was great.
Label: Ambivalent
Explanation: Directing is only part of what constitutes a film.
Question: What's your favorite film?
Answer: Fight Club, Filth, and Hereditary.
Label: Ambivalent
Explanation: The reply gives three movies instead of one, which makes the desired information unclear.

Question: The hypothesis I was discussing, wouldn't you regard that as a defeat?
Answer: I am not going to prophesy what will happen.
Label: Clear Non-Reply
Explanation: Directly stating they won't answer.
Question: 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?
Answer: I do not know that date. I will find out and let the House know.
Label: Clear Non-Reply
Explanation: Claims/admits they don't have the information.
Question: Was it your decision to release the fund?
Answer: You mean the public fund?
Label: Clear Non-Reply
Explanation: Gives no data, asks for clarification.
Part of the interview
<Part of the interview>
Question
<Question>
Taxonomy code: """

H.1 Prompt for LoRA fine-tuning

For the instruction-tuning part, we rely on LoRA fine-tuning (Hu et al., 2021) with $r = 16$, $alpha = 32$ and $dropout = 0.05$ using a subset of 2700 annotated samples as training set and the rest 750 as validation set. The following prompt was used for instruction-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. Inference proceeded without sampling, though we did experiment with sampling, which resulted in slightly lower performance.

---

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>
"""

I 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 230 GPU hours and 440 CPU hours.