

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

Anonymous ACL submission

## Abstract

001 Equivocation and ambiguity in public speech  
002 are well-studied discourse phenomena, espe-  
003 cially in political science and analysis of politi-  
004 cal interviews. Inspired by the well-grounded  
005 theory on equivocation, we aim to resolve the  
006 closely related problem of response clarity in  
007 questions extracted from political interviews,  
008 leveraging the capabilities of Large Language  
009 Models (LLMs) and human expertise. To this  
010 end, we introduce a *novel taxonomy* that frames  
011 the task of detecting and classifying response  
012 clarity and a corresponding *clarity classifica-*  
013 *tion dataset* which consists of question-answer  
014 (QA) pairs drawn from political interviews and  
015 annotated accordingly. Our proposed two-level  
016 taxonomy addresses the clarity of a response in  
017 terms of the information provided for a given  
018 question (high-level) and also provides a fine-  
019 grained taxonomy of evasion techniques that  
020 relate to unclear, ambiguous responses (lower-  
021 level). We combine ChatGPT and human anno-  
022 tators to collect, validate and annotate discrete  
023 QA pairs from political interviews, to be  
024 used for our newly introduced response clarity  
025 task. We provide a detailed analysis and con-  
026 duct several experiments with different model  
027 architectures, sizes and adaptation methods to  
028 gain insights and establish new baselines over  
029 the proposed dataset and task.<sup>1</sup>

## 1 Introduction

031 In the era of mass information dissemination,  
032 question evasion and response ambiguity are  
033 widespread phenomena in political interviews and  
034 debates, rendering their detection an important as-  
035 pect of political discourse studies. Bull (2003)  
036 presents a meta-analysis of five studies on politi-  
037 cal interview Q&As, concluding that politicians  
038 gave clear responses to only 39-46% of questions  
039 during televised interviews, while non-politicians  
040 had a significantly higher 70-89% reply rate. In

<sup>1</sup>Code and data will be made available upon publication.

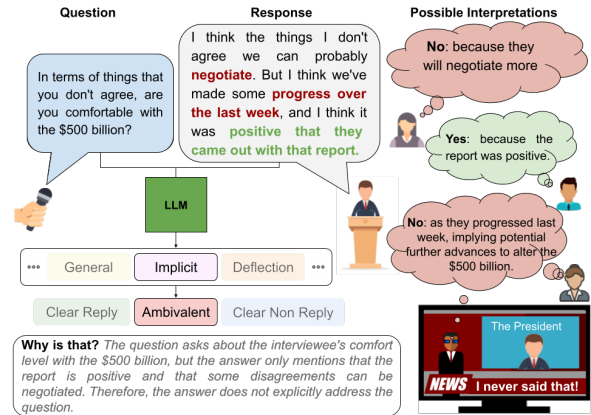


Figure 1: An example from an interview from our dataset with classification along with an analysis from instruction-tuned Llama-70b.

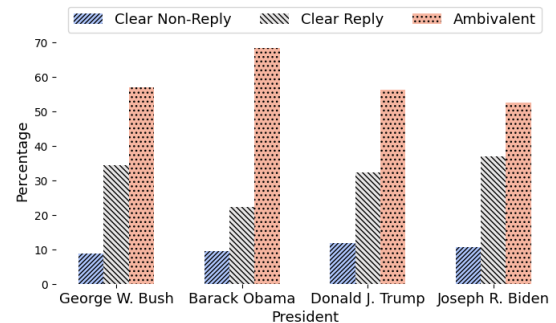


Figure 2: Statistics on answer clarity in political interviews of the latest 4 US presidents.

041 Figure 2 we present statistics derived from our hu-  
042 man annotations regarding response clarity among  
043 US presidents, revealing that politicians often avoid  
044 providing clear responses to journalists' questions.

045 This phenomenon is known as *equivocation* or  
046 *evasion* in academic literature and describes a non-  
047 straightforward type of communication, which is  
048 characterised by lack of clarity and includes speech  
049 acts such as contradictions, inconsistencies, subject  
050 switches, incomplete sentences, misunderstandings,  
051 obscure mannerisms of speech (Watzlawick  
052 et al., 1964; Bavelas et al., 1988; Rasiah, 2010), ren-

053	dering political speech susceptible to multiple inter-	various training and inference frameworks, show-	105
054	pretations from the perspective of the public. Fig-	ing that simple prompting and instruction-tuning	106
055	ure 1 presents an example of an interview featur-	techniques using our dataset are highly capable	107
056	ing various interpretations, generated labels, and cor-	of providing meaningful performance. Moreover,	108
057	responding explanations using our proposed dataset.	we find that using the labels of the second level	109
058	While the topic has been studied extensively in	(evasion labels) in a two-step classification strategy	110
059	the field of linguistics, politics and communication,	helps boost performance for clarity classification.	111
060	with several typologies proposed for classifying	We argue that being able to detect answer ambi-	112
061	question responses (Harris, 1991; Bull and Mayer,	guity automatically will facilitate political speech	113
062	1993; Rasiah, 2010), there has been little attempt	discourse analysis, allowing for comparisons at	114
063	to analyse whether such typologies are applicable	scale. Additionally, the proposed task can shed	115
064	to larger scale data and consistent with varying	light on LLM capabilities of reasoning over long	116
065	human perspectives and biases. In other words, the	contexts and prove useful for other downstream	117
066	possibility of automatically classifying response	tasks in NLP such as question answering (see also	118
067	clarity has not been explored in NLP, potentially	§2.1). To sum up, our contributions are threefold:	119
068	due to the complexity of the task itself, as well		
069	as the underlying need to encode and reason over	• We propose a new task, <i>response clarity evalu-</i>	120
070	long context. However, recent advancements in	<i>ation</i> , which aims to detect the alignment and	121
071	language modelling boosted model performance	clarity of a given response with respect to its	122
072	for long-context inputs (Dai et al., 2019; Wei et al.,	respective question and provide an empirically	123
073	2022, 2023), paving the way for framing the task	and theoretically established taxonomy for it.	124
074	of <b>automatically measuring response clarity</b> .		
075	Related to this endeavour, there is related work	• We introduce a human-labelled dataset on the	125
076	focusing on the responder’s intent interpretation	aforementioned task, comprising 3,445 QA	126
077	(Ferracane et al., 2021), or the answerability of	pairs from political interviews.	127
078	questions for question-answering (QA) tasks (Min		
079	et al., 2020; BingningWang et al., 2020; Rogers	• We experiment with several language models	128
080	et al., 2020; Sun et al., 2022; Wang et al., 2022).	and methods to gain insights establish perfor-	129
081	However, in both research directions, the focus	mance baselines for the proposed task.	130
082	deviates from directly assessing the clarity of the		
083	response, being obfuscated by perceptions of intent	<b>2 Related work</b>	131
084	or question clarity. We address this by proposing		
085	the task of <b>response clarity evaluation</b> , focusing	<b>2.1 Equivocation in Social Sciences</b>	132
086	exclusively on assessing the effect of the response,		
087	building on relevant discourse typologies.	Political equivocation, aptly generalised by Dillon	133
088	We carry out a detailed analysis of proposed ty-	(1990) as “the routine strategy for responding to a	134
089	pologies, considering their overlap and consistency,	question without answering it”, provides a range of	135
090	the distribution of proposed classes in our collected	frameworks to analyse evasive responses (Wilson,	136
091	data, and the feasibility of using them in an au-	1990; Bull, 2009; Bull and Strawson, 2019). Har-	137
092	tomated task, resulting in our proposed <i>two-level</i>	ris (1991) makes a distinction between direct and	138
093	<i>response clarity detection taxonomy</i> . Specifically,	indirect answers while others focus on how com-	139
094	the first level of the taxonomy accounts for a three-	plete the information conveyed by the response is	140
095	way evaluation of response <i>clarity</i> in terms of the	(Bull, 1994, 2003). Wilson (1990); Harris (1991);	141
096	number of interpretations the intended response	Bull (2003) provide criteria for the identification	142
097	holds. The second and more fine-grained level cov-	of three main categories (Bull and Mayer, 1993):	143
098	ers eleven common <i>evasion</i> phenomena in political	① <i>Replies</i> correspond to cases where the requested	144
099	literature, which explain in more detail the cate-	information is given in full. ② <i>Non-Replies</i> , where	145
100	gorization of responses in the three-scale clarity	none of the information requested is given in a clear	146
101	classes. We use this taxonomy to annotate a dataset	manner (Rasiah, 2010); non-Replies are broken	147
102	of political QA pairs and perform an analysis of	down into twelve further <i>evasion</i> sub-categories	148
103	the perspective variability among human annota-	(Table 1). Lastly, ③ <i>Intermediate replies</i> are those	149
104	tors. We then evaluate different LLMs, exploring	utterances that fall somewhere between replies and	150
		non-replies, i.e. responding completely but to one	151
		part of a multi-part question while ignoring the rest;	152

153 responding partially to a single-part question; an- 184  
154 swering a question in a suggestive manner without 185  
155 giving a straightforward answer. 186

156 Bull (2003) breaks the 12 evasion techniques of 187  
157 Table 1 further into 28 more fine-grained micro- 188  
158 categories; for example “*Makes political point*” in- 189  
159 cludes the micro-categories “*External attacks on* 190  
160 *the opposition or other rival groups*”, “*Talks up* 191  
161 *one’s own side*”, “*Presents policy*”. Rasiah (2010) 192  
162 separates the Replies into Direct and Indirect, keeps 193  
163 the Intermediate Replies category as is, while also 194  
164 breaking down the Non-reply category (which he 195  
165 labels “Evasions”) into four degrees of evasiveness, 196  
166 whether the evasion was overt or covert and what 197  
types of ‘agenda shifts’ occurred. 198

- 
1.  **Ignores the question.** Makes no attempt to answer the question, or even to acknowledge it has been asked. 199
  2.  **Acknowledges the question.** Acknowledges that a question has been asked, but equivocates. 200
  3.  **Questions the question.** Requests clarification, or reflects the question back to the questioner. 201
  4.  **Attacks the question.** 202
  5.  **Personalisation.** Makes personal comments or attacks. 203
  6.  **Declines to answer.** 204
  7.  **Makes political points.** 205
  8.  **Gives incomplete reply.** 206
  9.  **Repeats answer to the previous question.** 207
  10.  **States or implies has already answered the question.** 208
  11.  **Apologises.** 209
  12.  **Literalism.** The literal aspect of a question which was not intended to be taken literally is answered. 210
- 

Table 1: Equiv. typology by Bull and Strawson (2019). 211

167 Tailoring these typologies into a response clarity 212  
168 taxonomy suitable for an NLP dataset, it is impera- 213  
169 tive to modify them considering the following: 214  
170

- Our focus is slightly different: we target a tax- 215  
172 onomy that classifies the clarity of responses 216  
173 (hence an indirect response falls under a dif- 217  
174 ferent category than a direct one). 218
- We seek a good per class representation in our 219  
176 dataset to allow computational modelling us- 220  
177 ing LLMs. It is thus necessary to condense 221  
178 classes to avoid overly sparse categorisation 222  
179 while retaining the essential per class charac- 223  
180 teristics (i.e., we provide meaningful labels). 224
- Labelling of the responses is conducted by 225  
181 non-expert human annotators so that our an- 226  
182 notations also account for the perception and 227  
183 228

reasoning of the general audience of political 184  
interviews rather than a minority of experts. 185  
The difficulty of the classification, and thus 186  
the resulting error rate, increases as we in- 187  
crease the set of labels they choose from. 188

- Most interviewers pose multi-barrelled ques- 189  
tions. We break those multi-part questions 190  
into singular QA pairs and label each one sep- 191  
arately, to retain this fine-grained information. 192

Section 3 discusses the taxonomy we adopted, 193  
aiming to optimise for the annotation task. 194

## 2.2 Equivocation in NLP 195

While *equivocation* has not been adequately stud- 196  
ied in NLP, there are related areas, such as question 197  
answerability, political discourse analysis and de- 198  
ceptive intent detection. 199

### 2.2.1 Answerability in question answering 200

There have been several tasks proposed related to 201  
QA both in open-ended and closed set answer se- 202  
tups. The issue of the *answerability* of a given 203  
question in QA was highlighted in SQuAD 2.0 204  
(Rajpurkar et al., 2018), which introduced adver- 205  
sarially crafted unanswerable questions with re- 206  
spect to a given text span. Lee et al. (2020) ex- 207  
panded the SQuAD 2.0 dataset, also incorporating 208  
the rationale for unanswerable questions. Extend- 209  
ing to out-of-domain questions to address practical 210  
use cases, Sulem et al. (2021) introduce competi- 211  
tive and non-competitive unanswerable questions. 212  
Relevant endeavours question the answerability of 213  
information-seeking queries built independently of 214  
the passage containing possible answers to those 215  
queries (Asai and Choi, 2020). Scalability issues 216  
are addressed via synthetic extensions of existing 217  
datasets containing both answerable and unanswer- 218  
able questions (Nikolenko and Kalehbasti, 2020). 219  
To the same end, other works develop data augmen- 220  
tation techniques to produce unanswerable queries 221  
based on answerable SQuAD 2.0 queries (Zhu 222  
et al., 2019; Du et al., 2022). Other datasets target- 223  
ing answerability issues are ReCO (BingningWang 224  
et al., 2020), which provides “yes”, “maybe” and 225  
“no” labels for questions paired with passages in 226  
Chinese, as well as QuAIL (Rogers et al., 2020), 227  
which introduces questions of varying certainty ac- 228  
cording to the accompanying passage. 229

While our task shares a connection with question 230  
answerability, our focus is on annotating *response* 231  
*clarity* in relation to a given question. This distinc- 232

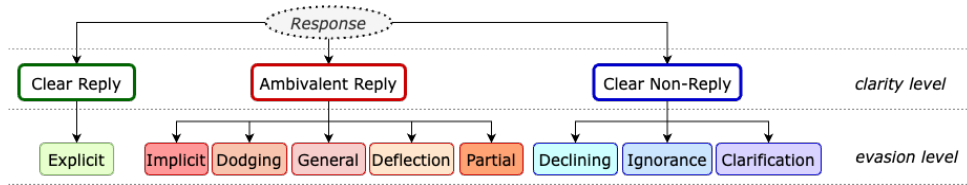


Figure 3: Our proposed taxonomy of response clarity classification.

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<sup>2</sup>. 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

<sup>2</sup>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  $\kappa$  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  $\kappa$  (higher values are better) between annotators for the ‘clarity’ classification level.

For the ‘clarity’ category, the Fleiss Kappa  $\kappa$  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 ( $\kappa=0.97$ ), 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)* ( $\kappa=0.58$ ) and *Partial (Ambivalent)* vs *Explicit (Clear Reply)* ( $\kappa=0.68$ ), or ‘Declining’ (Clear Non-reply) vs ‘Dodging’ (Ambivalent) ( $\kappa=0.77$ ). 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 perspec-

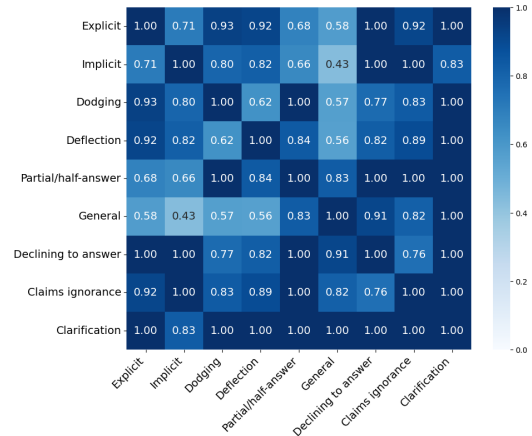


Figure 4: Annotators’ agreement using Fleiss  $\kappa$  for labels assigned to the ‘evasion’ classification level.

tives 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.

**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)<sup>3</sup>. 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<sup>4</sup>.

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

<sup>3</sup>Specifically, we used version gpt-3.5-turbo-0613.

<sup>4</sup>Note that results for XxBERTa models are overestimated due to constraint input token size.

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	<u>0.649</u>	<u>0.476</u>	<u>0.413</u>	<u>0.413</u>	
	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	<u>0.667</u>	<u>0.333</u>	<u>0.333</u>	<u>0.333</u>	
	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		<u>0.640</u>	<u>0.507</u>	<u>0.497</u>	<u>0.482</u>	
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		<u>0.541</u>	<u>0.565</u>	<u>0.452</u>	<u>0.365</u>	
FS Falcon-7b		<u>0.505</u>	<u>0.299</u>	<u>0.211</u>	<u>0.222</u>	
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	<u>0.64</u>	<u>0.579</u>	<u>0.516</u>	<u>0.53</u>	
	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	<b>0.759</b>	<b>0.67</b>	<b>0.70</b>	<b>0.68</b>	
	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		<u>0.58</u>	<u>0.523</u>	<u>0.586</u>	<u>0.546</u>	
Llama-7b		0.666	0.618	0.616	0.616	
Llama-13b		0.675	0.617	0.616	0.616	
Llama-70b		<u>0.713</u>	<b>0.67</b>	<b>0.71</b>	<b>0.682</b>	
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-

554 ing a task with higher dimensionality/complexity  
555 of targeted labels. Interestingly, asking to address  
556 all sQAs in one go (multi-CoT) harms performance  
557 instead of improving, potentially because of the  
558 impact on the amount of context that needs to be  
559 taken into account for generation.

560 In general, LLMs mostly struggled with distin-  
561 guishing between *Clear* vs *Ambivalent* replies, as  
562 well as *Partial* vs *General* ones. This resembles  
563 challenges (Figure 4) faced by human annotators  
564 but interestingly holds even for ZS and CoT models  
565 which were not trained on human annotations, sug-  
566 gesting a generalised difficulty in discerning these  
567 classes. Further insights are shown in App. E.

568 Turning to tuned models, we observe a differ-  
569 ence in behaviour: for direct clarity, smaller LLM  
570 models seem to struggle and are even outperformed  
571 by encoder models such as XLNet or BERT vari-  
572 ants, with only the 70b Llama outperforming them.  
573 Instead, evasion-driven classification consistently  
574 improves the performance of Llama variants. Ad-  
575 ditionally, Llama models outperform Falcon even  
576 with fewer parameters (e.g. the 13B Llama model  
577 outperforms the 40B Falcon across metrics). This  
578 aligns with other works where LLama-13b sur-  
579 passes Falcon-40b in reading comprehension (Tou-  
580 vron et al., 2023), while all LLama variants exhibit  
581 better prior knowledge (Sun et al., 2023), a crucial  
582 factor for our task as discussed below. We expand  
583 our experiments to assess the generalisation capa-  
584 bilities of the stronger Llama model (70B) using  
585 the dataset of (Ferracane et al., 2021), which is  
586 annotated with a different strategy, and provide an  
587 analysis as detailed in App. G.

588 Overall, for both prompting and tuning strate-  
589 gies, the evasion-based clarity classification vari-  
590 ant leads to better performance compared to the  
591 direct clarity one, indicating that the fine-grained  
592 subcategories of the taxonomy assisted in guiding  
593 the LLMs towards selecting the correct high-level  
594 clarity category more frequently. In other words,  
595 while the 9-way classification is more challenging  
596 (see also App. E), disambiguation between the  
597 finer-grained labels helps the models improve their  
598 accuracy on the higher-level ones. Further analysis  
599 of performance per class is provided in App. C.

600 **Answer grounding** We aim to separately assess  
601 whether models are influenced by the difficulty of  
602 identifying the relevant response snippets in the  
603 text, i.e. grounding the answer, a task that can be  
604 particularly challenging when a single reply ad-

605 dresses multiple questions. As a proxy to test this,  
606 we consider single- vs multi-part question subsets  
607 (35% vs 65% of the original test-set), assuming  
608 that answer grounding is harder for the latter, and  
609 we compare models and annotator performance.  
610 While Fleiss  $\kappa$  showed minimal disparity between  
611 humans across all models, metrics were notably  
612 higher for single-part questions, regardless of the  
613 method (ZS/FS, CoT, fine-tuning) or the classifica-  
614 tion variant (evasion-based or direct clarity). Per-  
615 formance improvements reached 0.16 for F-score,  
616 indicating the impact of QA complexity on model  
617 performance. More detailed results in App D.2.

618 **Model knowledge** We explore whether perfor-  
619 mance in the proposed task is influenced by models’  
620 “prior knowledge” of given entities. For instance, Q:  
621 “Did the Federal Reserve make the right move?”, A:  
622 “I think Bernanke is doing a great job” would be  
623 correctly classified as *Dodging* by models unaware  
624 that Bernanke is the chairman of the Federal Re-  
625 serve. To explore the prior knowledge hypothesis,  
626 we focus on person names and divide the test-set  
627 into two parts: one containing person names in ei-  
628 ther the question or the answer, and one excluding  
629 any named person mentions (60% vs 40% of the  
630 original test-set). All models performed better on  
631 the latter, “no-person” subset, but smaller models  
632 exhibited a much sharper improvement of up to  
633 0.20 in F-score (Llama-7b) compared to larger and  
634 presumably more “knowledgeable” ones, thus cor-  
635 roborating the findings of Sun et al. (2023). We  
636 provide more details in App D.3.

## 637 6 Conclusion

638 We introduce a novel task on response clarity classi-  
639 fication focusing on political interviews. Driven by  
640 studies of evasion techniques in political sciences,  
641 we propose a two-level hierarchical taxonomy for  
642 clarity classification that considers different eva-  
643 sion strategies at the lower (leaf) level. We also  
644 introduce a new dataset where question-answer  
645 pairs are manually annotated with the proposed  
646 taxonomy labels. We experiment with a range of  
647 different model architectures, sizes and adaptation  
648 strategies on our dataset, establishing several base-  
649 lines. We empirically show that fine-grained labels  
650 facilitate classification in response clarity, while  
651 encoded model knowledge is strongly associated  
652 with classification performance. We aspire for this  
653 work to motivate future research in the topic, both  
654 from the NLP and political sciences communities.



## 655 Limitations

656 Due to the usage of Large Language Models (Chat-  
657 GPT) in our pipeline, our annotation process is sus-  
658 ceptible to hallucinations, which could affect the  
659 quality of the sQA extraction and therefore the as-  
660 signment of correct labels. However, we attempt to  
661 mitigate this risk by asserting that our human anno-  
662 tators are attentive and not influenced by injecting  
663 counterfactual sQAs. Additionally, we manually in-  
664 spected the quality of both the ChatGPT-generated  
665 sQAs and the human annotations throughout the  
666 annotation campaign to ensure high-quality annota-  
667 tions. Further, despite being crucial for the quality  
668 of the derived dataset, the need for human anno-  
669 tators significantly limits the number of samples  
670 that can be annotated, especially when considering  
671 the complexity of the proposed task. Overall, our  
672 dataset and respective analysis are limited to the En-  
673 glish language and further work would be needed  
674 to generalise the findings to other languages, es-  
675 pecially low-resource ones. Finally, the inherently  
676 missing vocal features present in speech, as well  
677 as face movements and hand gestures limit the dis-  
678 course analysis to purely textual cues, potentially  
679 missing some evasion-related characteristics.

## 680 Potential risks

681 Potential risks associated with this work relate to  
682 the possibility of misclassification of a part of po-  
683 litical speech due to the usage of neural models  
684 (LLMs) as classifiers. This fact may result in erro-  
685 neously marking politicians' claims as unclear and  
686 evasive if our method is used in real-world scenar-  
687 ios without human monitoring, especially since the  
688 current state of LLMs under usage tends to halluci-  
689 nate and produce unfaithful outputs. Hence, further  
690 work to ensure the reliability and trustworthiness  
691 of the underlying models would be crucial for their  
692 deployment.

## 693 References

694 Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Al-  
695 shamsi, Alessandro Cappelli, Ruxandra Cojocaru,  
696 Merouane Debbah, Etienne Goffinet, Daniel Hes-  
697 low, Julien Launay, Quentin Malartic, Badreddine  
698 Nouné, Baptiste Pannier, and Guilherme Penedo.  
699 2023. Falcon-40B: an open large language model  
700 with state-of-the-art performance.

701 Akari Asai and Eunsol Choi. 2020. [Challenges in](#)  
702 [information-seeking qa: Unanswerable questions and](#)

[paragraph retrieval](#). In *Annual Meeting of the Associ-  
ation for Computational Linguistics*. 703  
704

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel  
Fernandez. 2022. [Stop measuring calibration when](#)  
[humans disagree](#). In *Proceedings of the 2022 Con-  
ference on Empirical Methods in Natural Language*  
*Processing*, pages 1892–1915, Abu Dhabi, United  
Arab Emirates. Association for Computational Lin-  
guistics. 705  
706  
707  
708  
709  
710  
711

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ul-  
mer, Haau-Sing Li, Raquel Fernández, Barbara  
Plank, Rico Sennrich, Chrysoula Zerva, and Wilker  
Aziz. 2023. Uncertainty in natural language gener-  
ation: From theory to applications. *arXiv preprint*  
*arXiv:2307.15703*. 712  
713  
714  
715  
716  
717

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. 718  
719  
720  
721

BingningWang, Ting Yao, Qi Zhang, Jingfang Xu, and  
Xiaochuan Wang. 2020. [Reco: A large scale chinese](#)  
[reading comprehension dataset on opinion](#). 722  
723  
724

P. Bull. 2003. *The Microanalysis of Political Communi-  
cation: Claptrap and Ambiguity*. Routledge. 725  
726

Peter Bull. 1994. [On identifying questions, replies,](#)  
[and non-replies in political interviews](#). *Journal of*  
*Language and Social Psychology*, 13:115 – 131. 727  
728  
729

Peter Bull. 2009. *Techniques of political interview anal-  
ysis*, pages 215–228. Cambridge Scholars Publishing. 730  
731

Peter Bull and Kate Mayer. 1993. [How not to answer](#)  
[questions in political interviews](#). *Political Psychol-  
ogy*, 14:651–666. 732  
733  
734

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](#). *Parlia-  
mentary Affairs*. 735  
736  
737  
738

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Car-  
bonell, 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. 739  
740  
741  
742  
743  
744

Jim T. Dillon. 1990. [The practice of questioning](#). 745

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiy-  
ong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and  
Zhifang Sui. 2022. A survey for in-context learning.  
*arXiv preprint arXiv:2301.00234*. 746  
747  
748  
749

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. 750  
751  
752  
753  
754  
755

756	Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrina Erk. 2021. <a href="#">Did they answer? subjective acts and intents in conversational discourse</a> . In <i>Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies</i> , pages 1626–1644, Online. Association for Computational Linguistics.	810
757		811
758		812
759		813
760		814
761		815
762		816
763	J.L. Fleiss et al. 1971. Measuring nominal scale agreement among many raters. <i>Psychological Bulletin</i> , 76(5):378–382.	817
764		818
765		819
766	Codruta Liliana Girlea. 2017. <a href="#">Deception detection in dialogues</a> .	820
767		821
768	Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. <a href="#">What comes next? evaluating uncertainty in neural text generators against human production variability</a> . In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 14349–14371, Singapore. Association for Computational Linguistics.	822
769		823
770		824
771		825
772		826
773		827
774		828
775		
776	Sandra Harris. 1991. Evasive action: how politicians respond to questions in political interviews.	829
777		830
778	Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. <a href="#">Deberta: Decoding-enhanced bert with disentangled attention</a> . In <i>International Conference on Learning Representations</i> .	831
779		832
780		833
781		834
782	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 <i>International Conference on Learning Representations</i> .	835
783		836
784		837
785		838
786		839
787	Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. <i>Advances in neural information processing systems</i> , 35:22199–22213.	840
788		841
789		842
790		843
791		844
792	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 <i>Proceedings of the Twelfth Language Resources and Evaluation Conference</i> , pages 5425–5432.	845
793		846
794		847
795		848
796		849
797		850
798	Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. <a href="#">Roberta: A robustly optimized bert pretraining approach</a> .	851
799		852
800		853
801		854
802		855
803	Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2020. <a href="#">Interview: Large-scale modeling of media dialog with discourse patterns and knowledge grounding</a> . In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 8129–8141, Online. Association for Computational Linguistics.	856
804		857
805		858
806		859
807		860
808		861
809		862
	Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. <a href="#">AmbigQA: Answering ambiguous open-domain questions</a> . In <i>Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)</i> , pages 5783–5797, Online. Association for Computational Linguistics.	863
		864
		865
	Raymond S. Nickerson. 1998. <a href="#">Confirmation bias: A ubiquitous phenomenon in many guises</a> . <i>Review of General Psychology</i> , 2:175 – 220.	866
	Liubov Nikolenko and Pouya Rezazadeh Kalehbasti. 2020. <a href="#">When in doubt, ask: Generating answerable and unanswerable questions, unsupervised</a> . <i>ArXiv</i> , abs/2010.01611.	867
	Barbara Plank. 2022. The “problem” of human label variation: On ground truth in data, modeling and evaluation. In <i>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing</i> , pages 10671–10682.	868
	Rajkumar Pujari and Dan Goldwasser. 2021. <a href="#">Understanding politics via contextualized discourse processing</a> . In <i>Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing</i> , pages 1353–1367, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.	869
	Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. <a href="#">Know what you don’t know: Unanswerable questions for SQuAD</a> . In <i>Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)</i> , pages 784–789, Melbourne, Australia. Association for Computational Linguistics.	870
	Parameswary Rasiyah. 2010. <a href="#">A framework for the systematic analysis of evasion in parliamentary discourse</a> . <i>Journal of Pragmatics</i> , 42:664–680.	871
	Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. <a href="#">Getting closer to ai complete question answering: A set of prerequisite real tasks</a> . <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , 34(05):8722–8731.	872
	Elior Sulem, Jamaal Hay, and Dan Roth. 2021. <a href="#">Do we know what we don’t know? studying unanswerable questions beyond squad 2.0</a> . In <i>Conference on Empirical Methods in Natural Language Processing</i> .	873
	Haitian Sun, William Cohen, and Ruslan Salakhutdinov. 2022. <a href="#">ConditionalQA: A complex reading comprehension dataset with conditional answers</a> . In <i>Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 3627–3637, Dublin, Ireland. Association for Computational Linguistics.	874
	Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023. <a href="#">Head-to-tail: How knowledgeable are large language models (llm). AKA will llms replace knowledge graphs</a> .	875

866	Hugo Touvron, Louis Martin, Kevin Stone, Peter Al-	Xlnet: Generalized autoregressive pretraining for lan-	923
867	bert, Amjad Almahairi, Yasmine Babaei, Nikolay	guage understanding. <i>Advances in neural informa-</i>	924
868	Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti	<i>tion processing systems</i> , 32.	925
869	Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton		
870	Ferrer, Moya Chen, Guillem Cucurull, David Esiobu,	Haichao Zhu, Li Dong, Furu Wei, Wenhui Wang, Bing	926
871	Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller,	Qin, and Ting Liu. 2019. <a href="#">Learning to ask unanswer-</a>	927
872	Cynthia Gao, Vedanuj Goswami, Naman Goyal, An-	<a href="#">able questions for machine reading comprehension</a> .	928
873	thony Hartshorn, Saghar Hosseini, Rui Hou, Hakan	In <i>Annual Meeting of the Association for Computa-</i>	929
874	Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa,	<i>tional Linguistics</i> .	930
875	Isabel Kloumann, Artem Korenev, Punit Singh Koura,		
876	Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Di-		
877	ana Liskovich, Yinghai Lu, Yuning Mao, Xavier Mar-		
878	tinnet, Todor Mihaylov, Pushkar Mishra, Igor Moly-		
879	bog, Yixin Nie, Andrew Poulton, Jeremy Reizen-		
880	stein, Rashi Rungta, Kalyan Saladi, Alan Schelten,		
881	Ruan Silva, Eric Michael Smith, Ranjan Subrama-		
882	nian, Xiaoqing Ellen Tan, Binh Tang, Ross Tay-		
883	lor, Adina Williams, Jian Xiang Kuan, Puxin Xu,		
884	Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan,		
885	Melanie Kambadur, Sharan Narang, Aurelien Ro-		
886	driguez, Robert Stojnic, Sergey Edunov, and Thomas		
887	Scialom. 2023. <a href="#">Llama 2: Open foundation and fine-</a>		
888	<a href="#">tuned chat models</a> .		
889	Daniela Trotta and Sara Tonelli. 2021. <a href="#">Are gestures</a>		
890	<a href="#">worth a thousand words? an analysis of interviews</a>		
891	<a href="#">in the political domain</a> . In <i>Proceedings of the 1st</i>		
892	<i>Workshop on Multimodal Semantic Representations</i>		
893	<i>(MMSR)</i> , pages 11–20, Groningen, Netherlands (On-		
894	line). Association for Computational Linguistics.		
895	Jiexin Wang, Adam Jatowt, and Masatoshi Yoshikawa.		
896	2022. Archivalqa: a large-scale benchmark dataset		
897	for open-domain question answering over historical		
898	news collections. In <i>Proceedings of the 45th Inter-</i>		
899	<i>national ACM SIGIR Conference on Research and</i>		
900	<i>Development in Information Retrieval</i> , pages 3025–		
901	3035.		
902	Paul Watzlawick, Janet Beavin Bavelas, and Don D.		
903	Jackson. 1964. <a href="#">Pragmatics of human communication:</a>		
904	<a href="#">A study of interactional patterns, pathologies and</a>		
905	<a href="#">paradoxes</a> .		
906	Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel,		
907	Barret Zoph, Sebastian Borgeaud, Dani Yogatama,		
908	Maarten Bosma, Denny Zhou, Donald Metzler, Ed H.		
909	Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy		
910	Liang, Jeff Dean, and William Fedus. 2022. <a href="#">Emer-</a>		
911	<a href="#">gent abilities of large language models</a> . <i>Transactions</i>		
912	<i>on Machine Learning Research</i> . Survey Certifica-		
913	tion.		
914	Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert		
915	Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu,		
916	Da Huang, Denny Zhou, et al. 2023. Larger language		
917	models do in-context learning differently. <i>arXiv</i>		
918	<i>preprint arXiv:2303.03846</i> .		
919	John-Charles Wilson. 1990. <a href="#">Politically speaking: The</a>		
920	<a href="#">pragmatic analysis of political language</a> .		
921	Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Car-		
922	bonell, Russ R Salakhutdinov, and Quoc V Le. 2019.		

## 931 A Dataset details

### 932 A.1 Exploratory data analysis

933 In this section, we describe some interesting pat-  
 934 terns present in our proposed dataset.

935 **Label distribution** We start our analysis from  
 936 the core of this work, which is the distribution of  
 937 the final labels of our dataset, which are presented  
 938 in Figure 5. Overall, *Explicit Replies* is the most  
 939 prevalent category, followed by evasion categories  
 940 with significantly lower frequency each. Specifi-  
 941 cally, Explicit Replies contribute to 1051 samples  
 942 in total, followed by Dodging (704 samples), Im-  
 943 plicit (488 samples), General (386 samples), De-  
 944 flection (381 samples), Declining to answer (145  
 945 samples), Claims ignorance (119 samples), Clarifi-  
 946 cation (92 samples) and finally Partial/half-answer  
 947 (79 samples).

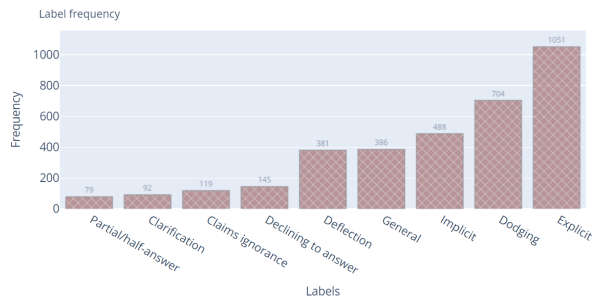


Figure 5: Label distribution in the dataset.

948 We also analyze the label distribution per pres-  
 949 ident in Figure 6, offering a more detailed insight  
 950 compared to Figure 2. According to the per pres-  
 951 ident distribution, we conclude that in our collected  
 952 interviews Donald J. Trump tends to provide more  
 953 Explicit Replies than the rest of the US presidents,  
 954 as indicated by the light-colored square of Figure  
 955 6.

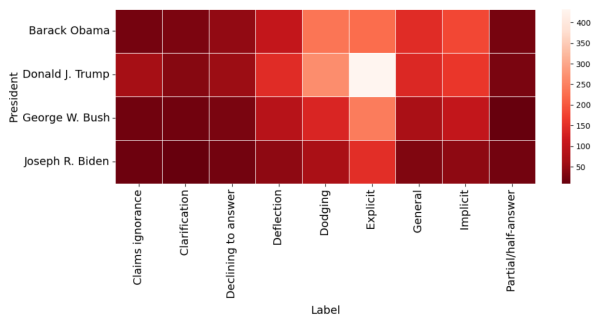


Figure 6: Label distribution per president.

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

958 **Temporal insights** Moving on to temporal char-  
 959 acteristics, in Figure 7 we provide some temporal  
 statistics regarding the interview distribution.

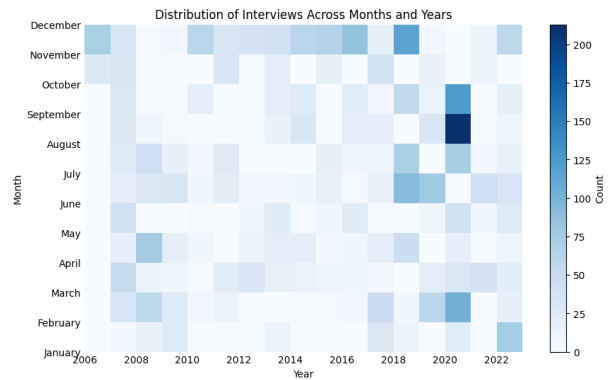


Figure 7: Visualization of interview distribution across months and years in the corpus

960 In Figure 8 we present the label distribution per  
 961 year in our dataset. We observe an elevated num-  
 962 ber of Explicit Replies in 2020, as indicated by  
 963 the light-colored cell. This observation can be  
 964 grounded to president-related information, as this  
 965 can be a strong characteristic in conjunction to la-  
 966 bel distribution.

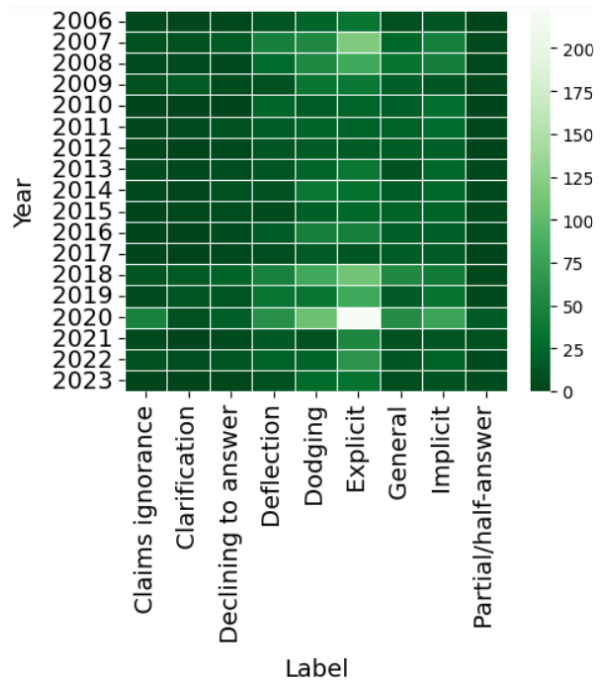


Figure 8: Label distribution across years

967 So, in association with US presidents, in Figure  
 968 9 we demonstrate the timeframe associated with  
 969 each president’s service. We can now conclude that  
 970 the higher number of Explicit Replies of Figure 8  
 971 coincides with Trump’s service, which is related  
 972

to more Explicit Replies, as indicated in Figure 6. Consequently, temporal evasion characteristics are highlighted in Figure 10.

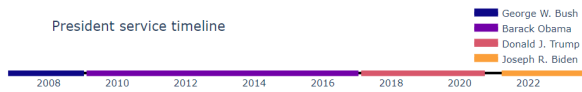


Figure 9: Service timeline for each US president

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 an-

notation 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

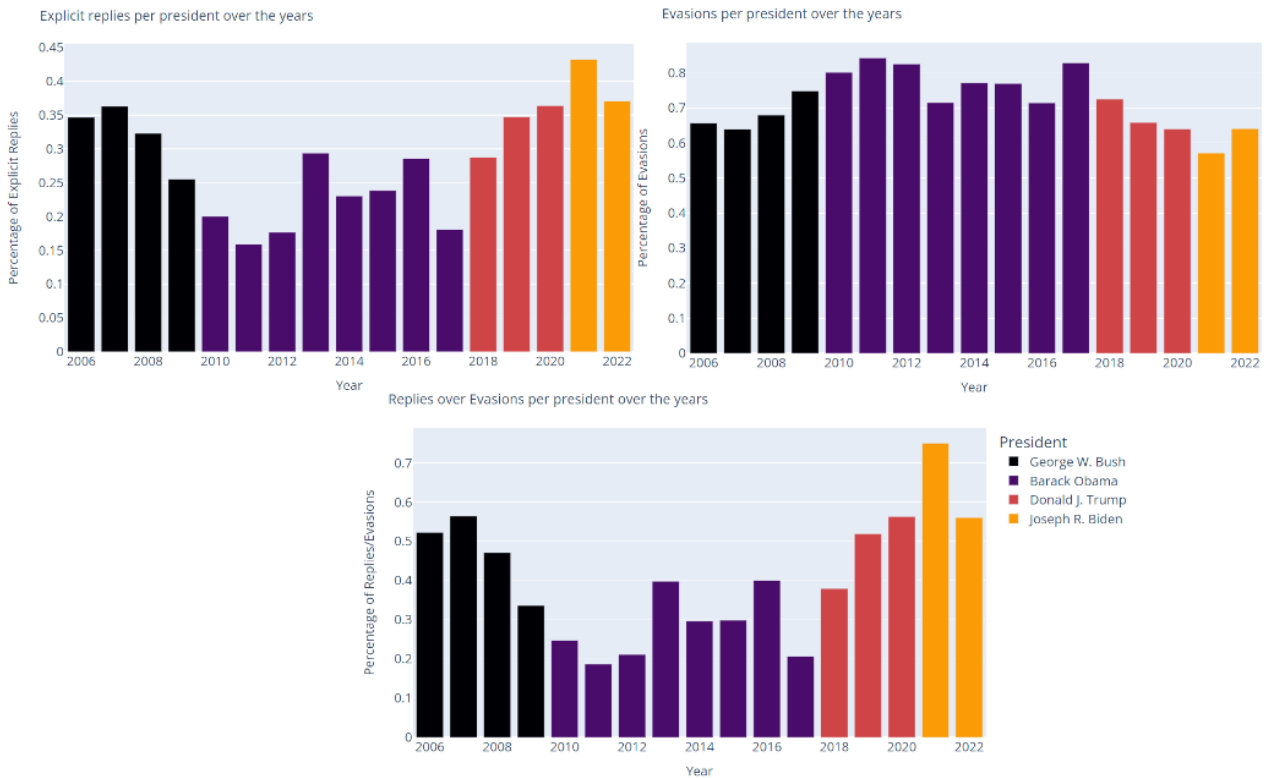


Figure 10: Percentages of Explicit Replies (left), Implicit/Non-Replies (right) and ratio of Replies over Implicit/Non-Replies (bottom) for each US president during their service.

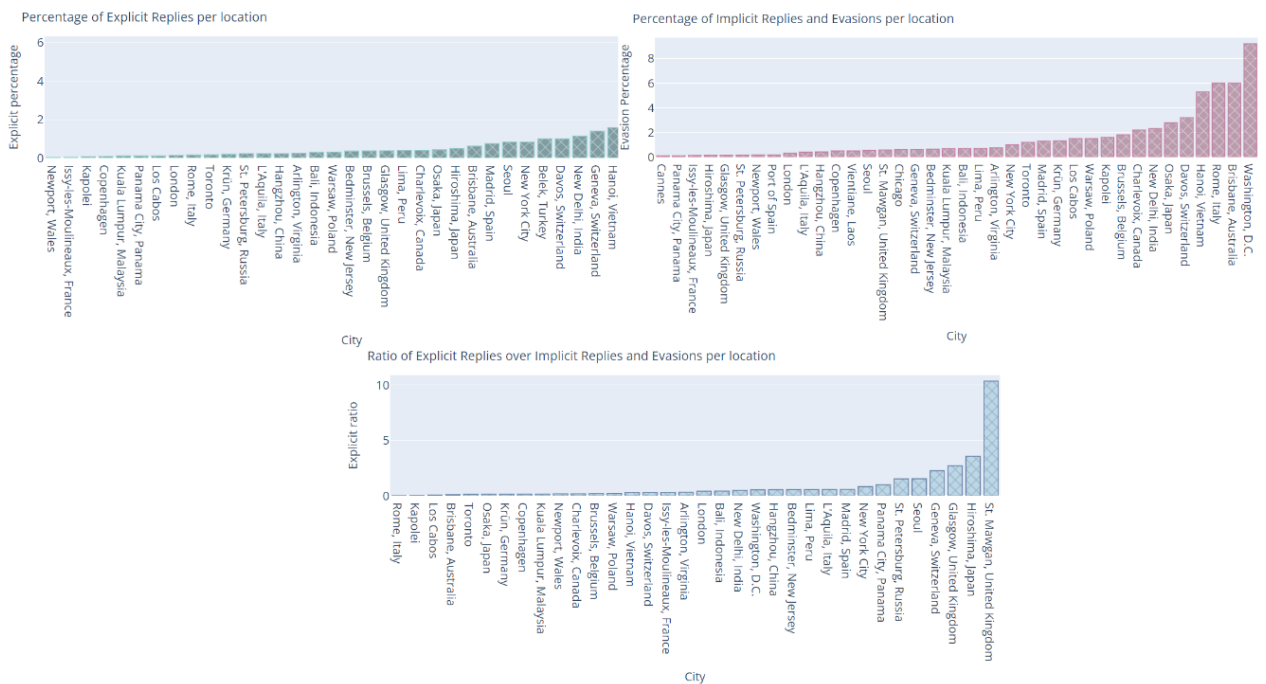


Figure 11: Percentages of Explicit Replies (left), Implicit/Non-Replies (right) and ratio of Replies over Implicit/Non-Replies (bottom) per location.

1070 also appear, while the opposite holds in cases with  
1071 single QA pairs.

**Political opponents** In Figure 16 we present  
the distribution of labels when a politician is in-  
terviewed on their own versus when they are in-  
terviewed with a political opponent. Politicians

1072  
1073  
1074  
1075

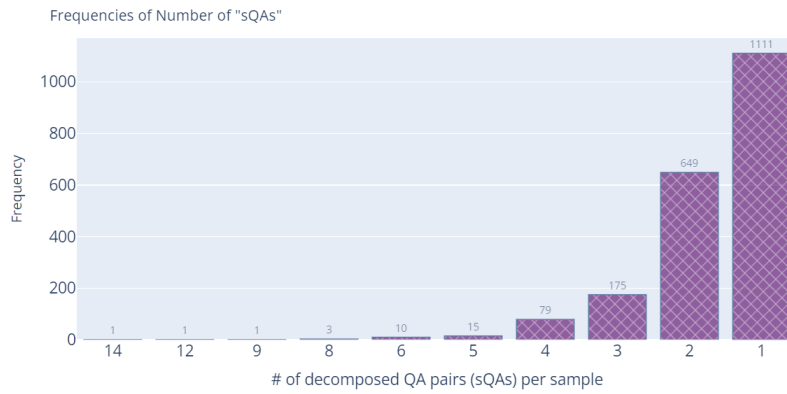


Figure 12: Distribution of sQAs length frequency.

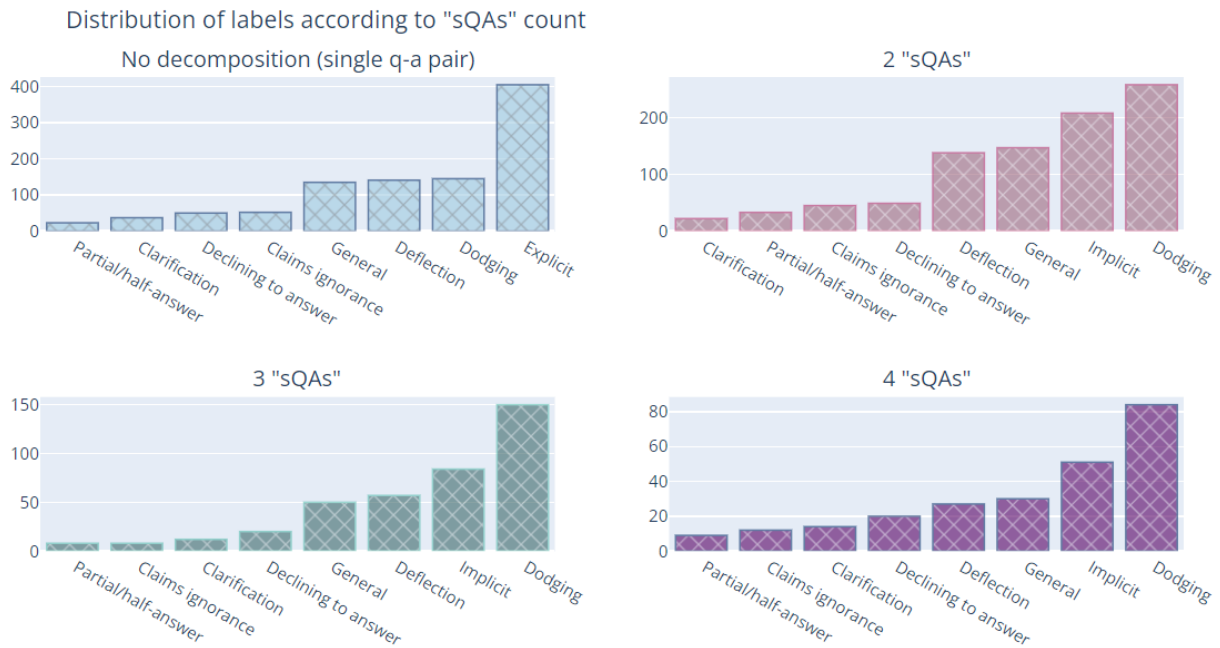


Figure 13: Label frequency per sQAs length.

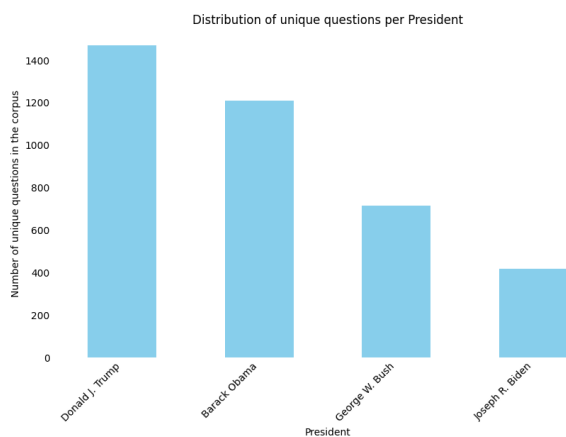


Figure 14: Visualization of distribution of unique questions per President in the corpus

Replies and evasion percentages, as proven by the similar bar height in both cases (with or without an opponent).

1077  
1078  
1079

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

1080  
1081  
1082  
1083  
1084  
1085  
1086  
1087  
1088  
1089  
1090  
1091  
1092  
1093

are more or less consistent towards their Explicit



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

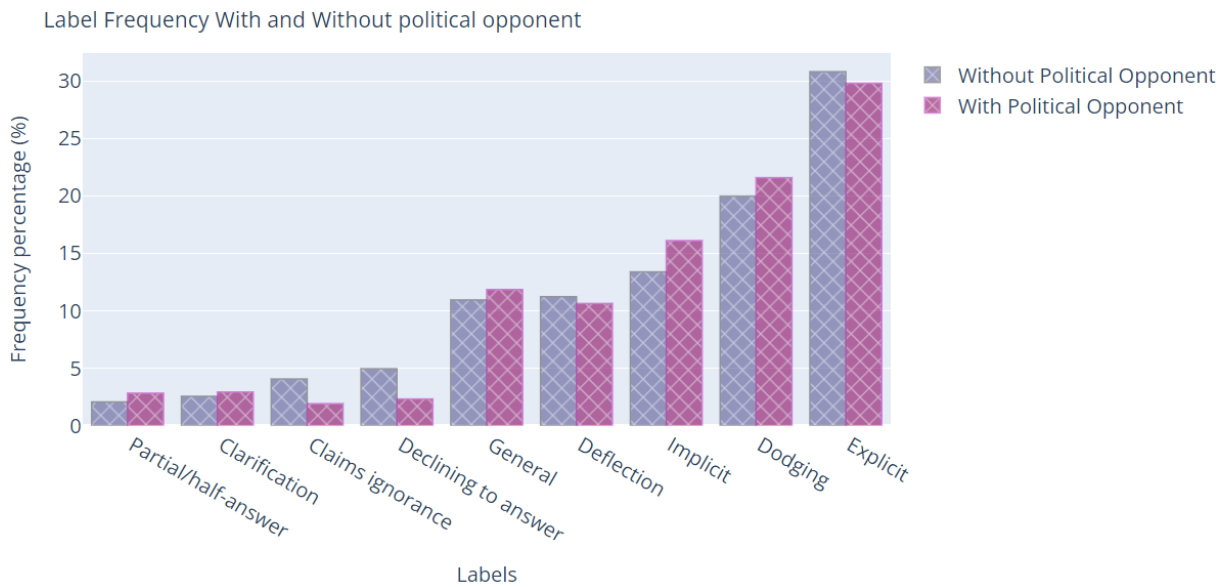


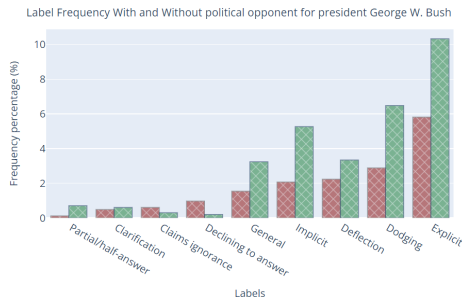
Figure 16: Label percentages for interviews with and without the presence of a political opponent.

1094 for Joseph R. Biden(Figure 17d), even though he  
 1095 tends to provide slightly more Explicit Replies in  
 1096 interviews with a political opponent. Donald J.  
 1097 Trump and Joseph R. Biden tend to employ evasion  
 1098 strategies in similar percentages with and without  
 1099 political opponents; some notable exceptions can  
 1100 be observed for Dodging categories, for which  
 1101 the percentages for Biden are higher in presence  
 1102 of a political opponent, while the opposite holds  
 1103 for Trump. In total, the label distributions for  
 1104 Barack Obama and Donald Trump are somewhat  
 1105 similar (note the ranking of labels, as well as the  
 1106 differences between bars with/without opponent),  
 1107 indicating a common behavior in handling inter-

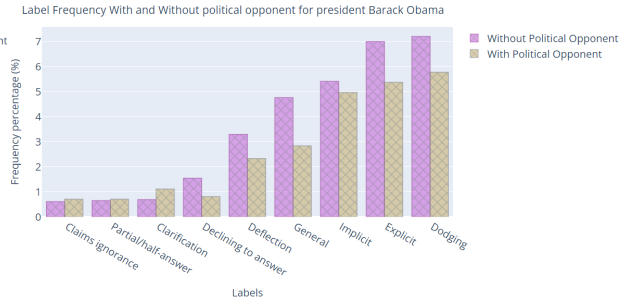
views with/without political opponents. George  
 1108 Bush holds a diverging distribution, in terms of  
 1109 presenting a larger gap between his top-1 category  
 1110 (Explicit Replies) and the rest; especially when  
 1111 being interviewed on his own, he tends to exploit  
 1112 significantly less evasion techniques in comparison  
 1113 to the rest of the presidents.  
 1114

Overall, our presented dataset accompanied  
 1116 by this exploratory analysis can be utilized by  
 1117 political scientists, assisting them in extracting  
 1118 interesting insights from political interviews.  
 1119

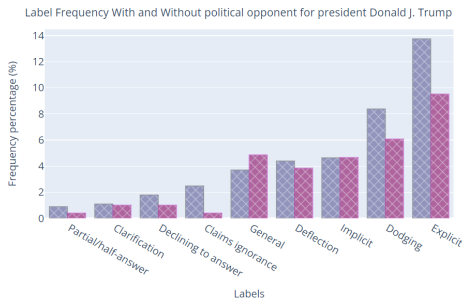




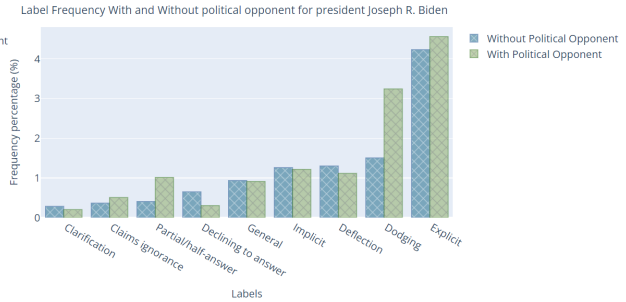
(a) Label distribution for G. Bush.



(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.

## 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 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.

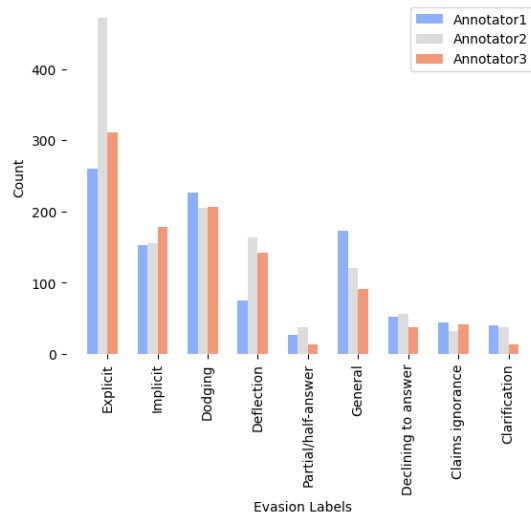


Figure 18: Visualization of distribution of evasion label per annotator in the corpus

**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,

	taxonomy	Description	Example
Clear R.	Explicit	The information requested is explicitly stated (in the requested form)	<b>Q:</b> er you have your own views about PR at Westminster don't you? <b>A:</b> I do. <i>Why?</i> - directly gives the info requested
	Implicit	The information requested is given, but without being explicitly stated (not in the expected form)	<b>Q:</b> Are you going to watch television? <b>A:</b> What else is there to do? <i>Why?</i> - they suggest planning to watch TV, despite not explicitly stating it
Ambivalent Reply	General	The information provided is too general/lacks the requested specificity	<b>Q:</b> What's your favourite film? <b>A:</b> Fight Club, Filth and Hereditary <i>Why?</i> - the reply gives three movies instead of one, which makes the desired information unclear
	Partial	Offers only a specific component of the requested information	<b>Q:</b> Did you enjoy the film? <b>A:</b> The directing was great <i>Why?</i> - Directing is only part of what constitutes a film
	Dodging	Ignoring the question altogether	<b>Q:</b> Do you like my new dress? <b>A:</b> We are late. <i>Why?</i> - 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	<b>Q:</b> Did you eat the last piece of pie? <b>A:</b> I have to admit that this was a great recipe, I always like it when there are chocolate chips in the dough. <i>Why?</i> - acknowledges the question but goes on a tangent about the chips, without answering
	Declining to answer	Acknowledge the question but directly or indirectly refusing to answer at the moment	<b>Q:</b> The hypothesis I was discussing, wouldn't you regard that as a defeat? <b>A:</b> I am not going to prophesy what will happen. <i>Why?</i> - directly stating they won't answer
Clear Non-Reply	Claims ignorance	The answerer claims/admits not to know the answer themselves	<b>Q:</b> 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? <b>A:</b> I do not know that date. I will find out and let the House know. <i>Why?</i> - claims/admits they don't have the information
	Clarification	Does not provide the requested information and asks for clarification	<b>Q:</b> Was it your decision to release the fund? <b>A:</b> You mean the public fund? <i>Why?</i> - gives no data, asks for clarification

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

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 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  $\leq 0.7$ .

**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 ev-

ident 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).

**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.

**Labelling platform** Our labelling process was conducted in the open source Label Studio<sup>5</sup> 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 corresponds to ‘Explicit Replies’; nevertheless, Explicit Replies only account for about the 1/3rd of the responses for all presidents, leaving much space for evasion schemas to appear. In comparison, Joe

<sup>5</sup><https://labelstud.io/>

Biden tends to provide more Explicit Replies, as resulting from our annotations.

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	<b>57.0</b>	<b>68.12</b>	<b>55.62</b>	<b>52.13</b>
Explicit	<b>34.31</b>	<b>22.38</b>	<b>32.6</b>	<b>37.34</b>
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.

**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 incorrect 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

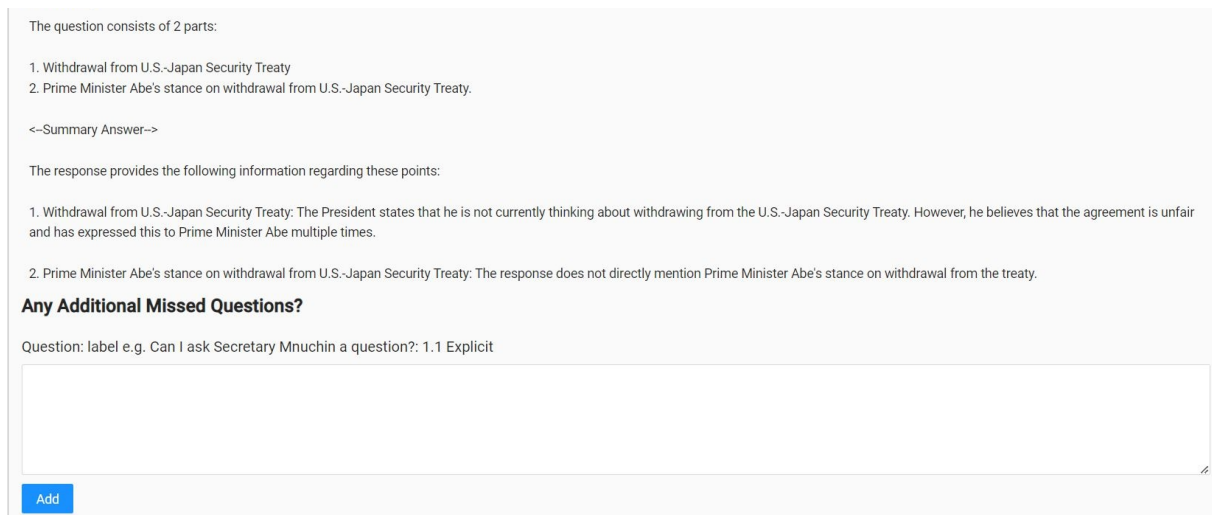


Figure 19: Screenshot from labelling platform: The sQAs for the provided QAs are given to the annotators. They have to highlight each of the enumerated responses and assign one of the labels of the taxonomy (as presented in Figure 20) to each of them.

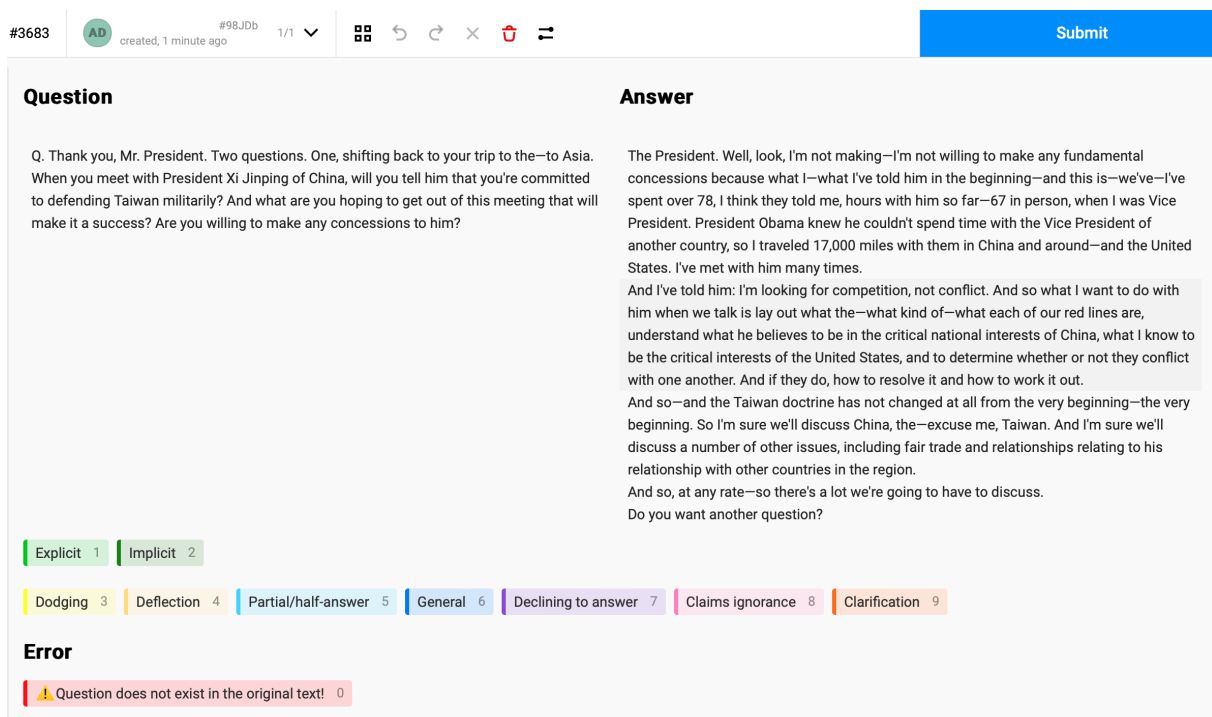


Figure 20: Screenshot from labelling platform: annotators have to read the original Question and Answer as provided. The classes corresponding to our proposed taxonomy are demonstrated as well.

original (the class label is not shown).<sup>6</sup> 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  $\leq 0.08$ .

<sup>6</sup>We provide the counterfactual sQA prompt at §H

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.

## B Experimental Details

In our experiments, we utilized three distinct datasets: training, development, and validation sets. The original dataset was divided into two parts,

1281  
1282  
1283  
1284  
1285  
1286  
1287  
1288

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).

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

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

where  $n_i$  is the number of instances in each class.

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	<b>0.75</b>	<b>0.757</b>	<b>0.75</b>	<b>0.752</b>
	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	<u>0.713</u>	<u>0.743</u>	<u>0.713</u>	<u>0.72</u>
	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**.

## 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 Llama-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 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

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	<b>0.667</b>	<u>0.333</u>	<u>0.333</u>	<u>0.333</u>
	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	<u>0.541</u>	<b>0.565</b>	<b>0.452</b>	<b>0.365</b>
	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**.

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

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

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
	standalone CoT	<b>0.625</b>	<b>0.559</b>	<b>0.483</b>	<b>0.461</b>
evasion based clarity	zero-shot	0.639	0.443	0.442	0.436
	standalone CoT	<b>0.661</b>	<b>0.683</b>	<b>0.603</b>	<b>0.56</b>
	standalone	0.712	0.568	0.483	0.489
	CoT	<b>0.643</b>	<b>0.657</b>	<b>0.558</b>	<b>0.536</b>

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.

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 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

Classification variant	Model	Acc.	Prec.	Recall	F1
direct clarity	Llama-7b	0.434	0.375	0.439	0.369
		<b>0.57</b>	<b>0.56</b>	<b>0.621</b>	<b>0.565</b>
	Llama-13b	0.55	0.527	0.663	0.52
		<b>0.639</b>	<b>0.631</b>	<b>0.731</b>	<b>0.638</b>
	Llama-70b	0.752	0.65	0.777	0.69
		<b>0.768</b>	<b>0.7</b>	<b>0.686</b>	<b>0.692</b>
Falcon-7b	0.266	0.255	0.319	0.148	
	<b>0.32</b>	<b>0.348</b>	<b>0.355</b>	<b>0.213</b>	
Falcon-40b	0.328	0.489	0.504	0.331	
	<b>0.359</b>	<b>0.533</b>	<b>0.55</b>	<b>0.374</b>	
evasion based clarity	Llama-7b	0.635	0.57	0.557	0.563
		<b>0.711</b>	<b>0.67</b>	<b>0.678</b>	<b>0.673</b>
	Llama-13b	0.651	0.573	0.611	0.589
		<b>0.711</b>	<b>0.674</b>	<b>0.636</b>	<b>0.653</b>
	Llama-70b	0.709	0.637	0.706	0.661
		<b>0.719</b>	<b>0.701</b>	<b>0.718</b>	<b>0.702</b>
Falcon-7b	0.497	0.387	0.319	0.332	
	<b>0.586</b>	<b>0.488</b>	<b>0.473</b>	<b>0.473</b>	
Falcon-40b	0.598	0.531	0.45	0.468	
	<b>0.656</b>	<b>0.665</b>	<b>0.601</b>	<b>0.622</b>	

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.

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

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

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.

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	<b>0.571</b>	<b>0.571</b>	<b>0.558</b>	<b>0.545</b>
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**.

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.

Model	Acc.	Prec.	Recall	F1
zero-shot	<b>0.315</b>	0.266	<b>0.284</b>	<b>0.244</b>
standalone CoT	0.259	<b>0.293</b>	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**.

	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.

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 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 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 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 observa-

Classification variant	Model	Acc.	Prec.	Recall	F1
direct clarity	DeBERTa-base	<b>0.562</b>	<b>0.521</b>	<b>0.467</b>	<b>0.465</b>
		0.593	0.512	0.439	0.416
	RoBERTa-base	<b>0.625</b>	<b>0.614</b>	<b>0.593</b>	<b>0.592</b>
		0.651	0.383	0.405	0.392
	XLNet-base	<b>0.68</b>	<b>0.557</b>	<b>0.571</b>	<b>0.56</b>
		0.704	0.481	0.468	0.472
evasion based clarity	DeBERTa-base	<b>0.57</b>	<b>0.576</b>	<b>0.645</b>	<b>0.568</b>
		0.545	0.498	0.715	0.509
	RoBERTa-base	<b>0.539</b>	<b>0.55</b>	<b>0.581</b>	<b>0.543</b>
		0.603	0.401	0.439	0.397
	XLNet-base	<b>0.594</b>	<b>0.552</b>	<b>0.617</b>	<b>0.574</b>
		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 without named entities.

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

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.

tion 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, 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

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	<u>0.64</u>	<u>0.579</u>	<u>0.516</u>	<u>0.53</u>
	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	<b>0.58</b>	<b>0.523</b>	<b>0.586</b>	<b>0.546</b>
	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	<b>0.775</b>	<b>0.743</b>	<b>0.724</b>	<b>0.732</b>
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	<u>0.724</u>	<u>0.695</u>	<u>0.702</u>	<u>0.698</u>

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	<u>0.694</u>	<u>0.403</u>	<u>0.41</u>	<u>0.406</u>
	RoBERTa-large	0.625	0.208	0.333	0.256
	DeBERTa-base	<b>0.549</b>	<b>0.44</b>	<b>0.734</b>	<b>0.424</b>
evasion based clarity	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**.

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, con-

1573  
1574  
1575  
1576  
1577  
1578  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619  
1620  
1621  
1622  
1623  
1624

	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.

trary 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 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

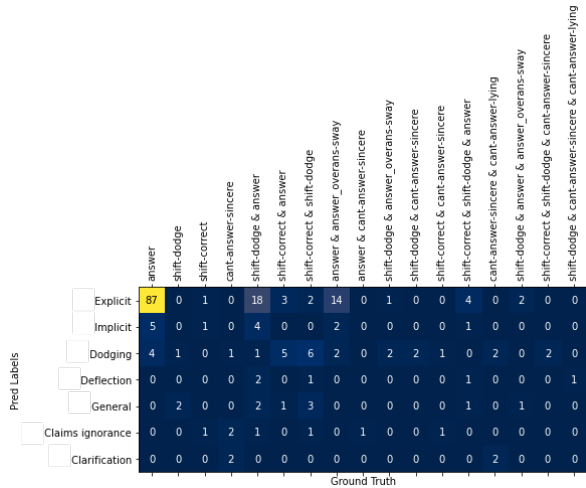


Figure 21: Results of Llama-70b trained using the evasion based clarity for dev set of (Ferracane et al., 2021).

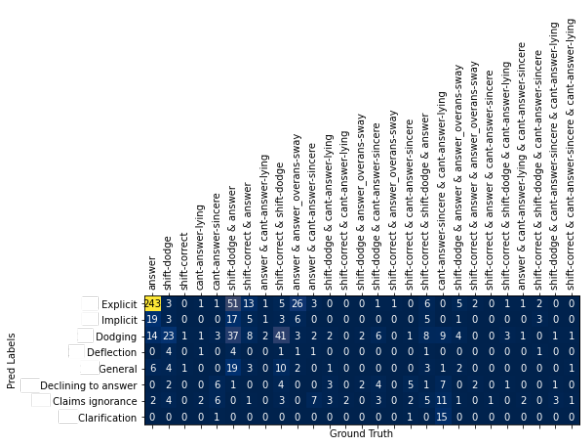


Figure 22: Results of Llama-70b trained using the evasion based clarity for the training set of (Ferracane et al., 2021).

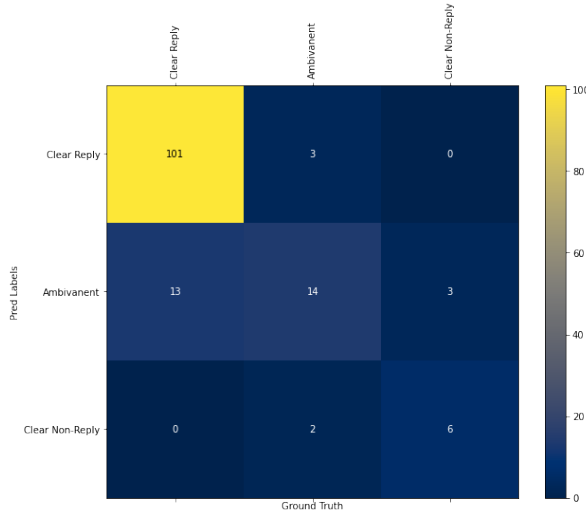


Figure 23: Confusion matrix of Llama-70b trained using the evasion based clarity for dev set of (Ferracane et al., 2021) for the single labelled instances.

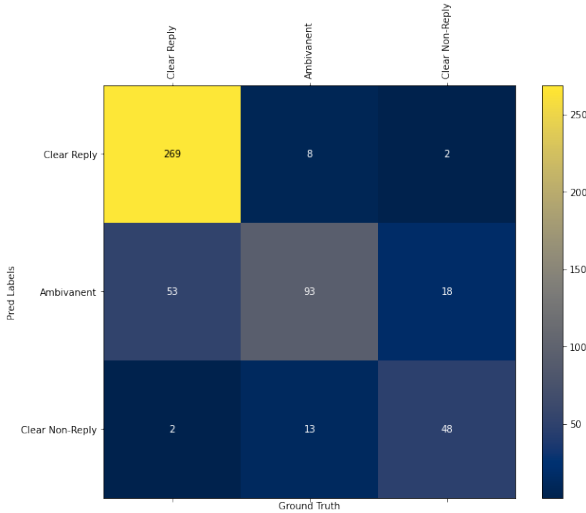


Figure 24: Results of Llama-70b trained using the evasion based clarity for the training set of (Ferracane et al., 2021) for the single labelled instances.

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:

““““

**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

1768	the clarity problem in the zero-shot scenario.			
1769				
1770	message_0 = “” Based on a segment of the interview in			
1771	which the interviewer poses a series of questions, classify the			
1772	type of response provided by the interviewee for the following			
1773	question using the following taxonomy and then provide a			
1774	chain of thought explanation for your decision:			
1775				
1776	1. Clear Reply - The information requested is explic-			
1777	itly stated (in the requested form)			
1778	2. Clear Non-Reply - The information requested is not given			
1779	at all due to ignorance, need for clarification or declining to			
1780	answer			
1781	3. Ambiguous - The information requested is given in			
1782	an incomplete way e.g. the answer is too general, partial,			
1783	implicit, dodging or deflection.			
1784				
1785	You are required to respond with a single term corre-			
1786	sponding to the Taxonomy code and only.			
1787				
1788	### Part of the interview ###			
1789	<Part of the interview>			
1790	### Question ###			
1791	<Question>			
1792	Taxonomy code: “”			
1793				
1794	<b>Chain-of-Thought (CoT) prompt for classifica-</b>			
1795	<b>tion</b> The following prompt was used for address-			
1796	ing the evasion problem in the CoT scenario.			
1797				
1798	message_0 = “” Based on a segment of the interview in			
1799	which the interviewer poses a series of questions, classify the			
1800	type of response provided by the interviewee for the following			
1801	question using the following taxonomy and then provide a			
1802	chain of thought explanation for your decision:			
1803				
1804	<Taxonomy>			
1805				
1806	You are required to respond with a single term corre-			
1807	sponding to the Taxonomy code as well as the chain of			
1808	thought explanation.			
1809				
1810	Let’s think step by step.			
1811	### Part of the interview ###			
1812	<Part of the interview>			
1813	### Question ###			
1814	<Question>			
1815	Taxonomy code: “”			
1816				
1817	The following prompt was used for addressing			
1818	the clarity problem in the CoT scenario.			
1819				
1820	message_0 = “” Based on a segment of the interview in			
1821	which the interviewer poses a series of questions, classify the			
1822	type of response provided by the interviewee for the following			
1823	question using the following taxonomy and then provide a			
1824	chain of thought explanation for your decision:			
1825				
1826	1. Clear Reply - The information requested is explic-			
1827	itly stated (in the requested form)			
1828	2. Clear Non-Reply - The information requested is not given			
1829	at all due to ignorance, need for clarification or declining to			
1830	answer			
1831	3. Ambivalent - The information requested is given in			
1832	an incomplete way e.g. the answer is too general, partial,			
1833	implicit, dodging or deflection			
1834				
		You are required to respond with a single term correspond-		1835
		ing to the Taxonomy code as well as the chain of thought		1836
		explanation.		1837
				1838
		Let’s think step by step.		1839
		### Part of the interview ###		1840
		<Part of the interview>		1841
		### Question ###		1842
		<Question>		1843
		Taxonomy code: “”		1844
				1845
		<b>Few-Shot (FS) prompt for classification</b> The		1846
		following prompt was used for addressing the eva-		1847
		sion problem in the FS scenario.		1848
				1849
		message_0 = “” Based on a segment of the interview in		1850
		which the interviewer poses a series of questions, classify the		1851
		type of response provided by the interviewee for the following		1852
		question using the following taxonomy:		1853
				1854
		<Taxonomy>		1855
				1856
		Here is one small example for each term of the tax-		1857
		onomy:		1858
		Question:		1859
		Do you have your own views about PR at Westminster don’t		1860
		you?		1861
		Answer:		1862
		I do.		1863
		Label: Explicit		1864
		Explanation: The answer directly gives the info requested.		1865
		Question: Are you going to watch television?		1866
		Answer: What else is there to do?		1867
		Label: Implicit		1868
		Explanation: They suggest planning to watch TV, despite not		1869
		explicitly stating it.		1870
		Question: Do you like my new dress?		1871
		Answer: We are late.		1872
		Label: Dodging		1873
		Explanation: Does not even acknowledge the question and		1874
		goes straight to another topic.		1875
		Question: Did you eat the last piece of pie?		1876
		Answer: I have to admit that this was a great recipe, I always		1877
		like it when there are chocolate chips in the dough.		1878
		Label: Deflection		1879
		Explanation: Acknowledges the question but goes on a tangent		1880
		about the chips, without answering.		1881
		Question: Did you enjoy the film?		1882
		Answer: The directing was great.		1883
		Label: Partial/half-answer		1884
		Explanation: Directing is only part of what constitutes a film.		1885
		Question: What’s your favorite film?		1886
		Answer: Fight Club, Filth, and Hereditary.		1887
		Label: General		1888
		Explanation: The reply gives three movies instead of one,		1889
		which makes the desired information unclear.		1890
		Question: The hypothesis I was discussing, wouldn’t you		1891
		regard that as a defeat?		1892
		Answer: I am not going to prophesy what will happen.		1893
		Label: Declining to answer		1894
		Explanation: Directly stating they won’t answer.		1895
		Question: On what precise date did the government order		1896
		the refit of the HMAS Kanimbla in preparation for its forward		1897
		deployment to a possible war against Iraq?		1898
		Answer: I do not know that date. I will find out and let the		1899
		House know.		1900
		Label: Claims ignorance		1901
		Explanation: Claims/admits they don’t have the information.		1902

1903	Question: Was it your decision to release the fund?	Label: Clear Non-Reply	1972
1904	Answer: You mean the public fund?	Explanation: Claims/admits they don't have the information.	1973
1905	Label: Clarification	Question: Was it your decision to release the fund?	1974
1906	Explanation: Gives no data, asks for clarification.	Answer: You mean the public fund?	1975
1907	### Part of the interview ###	Label: Clear Non-Reply	1976
1908	<Part of the interview>	Explanation: Gives no data, asks for clarification.	1977
1909	### Question ###	### Part of the interview ###	1978
1910	<Question>	<Part of the interview>	1979
1911	Taxonomy code: ""	### Question ###	1980
1912		<Question>	1981
1913	The following prompt was used for addressing	Taxonomy code: ""	1982
1914	the clarity problem in the FS scenario.		1983
1915			
1916	message_0 = ""	<b>H.1 Prompt for LoRA fine-tuning</b>	1984
1917	Based on a segment of the interview in which the inter-	For the instruction-tuning part, we rely on LoRA	1985
1918	viewer poses a series of questions, classify the type of response	fine-tuning (Hu et al., 2021) with $r = 16$ , $alpha =$	1986
1919	provided by the interviewee for the following question using	$32$ and $dropout = 0.05$ using a subset of 2700	1987
1920	the following taxonomy:	annotated samples as training set and the rest 750	1988
1921	1. Clear Reply - The information requested is explicitly	as validation set. The following prompt was used	1989
1922	stated (in the requested form)	for instruction-tuning, and it remained consistent	1990
1923	2. Clear Non-Reply - The information requested is not given	across all models and the two methodologies (di-	1991
1924	at all due to ignorance, need for clarification or declining to	rect clarity and evasion-based clarity). The only	1992
1925	answer	distinction between the two different setups in the	1993
1926	3. Ambivalent - The information requested is given in	prompt was the specific label that the model should	1994
1927	an incomplete way e.g. the answer is too general, partial,	generate. Inference proceeded without sampling,	1995
1928	implicit, dodging or deflection	though we did experiment with sampling, which	1996
1929		resulted in slightly lower performance.	1997
1930	Here is one small example for each term of the taxonomy:		1998
1931	Question:	message_0 = ""Based on a part of the interview where the	1999
1932	Do you have your own views about PR at Westminster don't	interviewer asks a set of questions, classify the type of answer	2000
1933	you?	the interviewee provided for the following question	2001
1934	Answer: I do.	### Part of the interview ###	2002
1935	Label: Clear Reply	<Interview Part>	2003
1936	Explanation: The answer directly gives the info requested.		2004
1937	Question: Are you going to watch television?		2005
1938	Answer: What else is there to do?	### Question ###	2006
1939	Label: Ambivalent	<Question>	2007
1940	Explanation: They suggest planning to watch TV, despite not	Label: <Label>	2008
1941	explicitly stating it.	.....	2009
1942	Question: Do you like my new dress?		2010
1943	Answer: We are late.		
1944	Label: Ambivalent	<b>I Computational Resources</b>	2011
1945	Explanation: Does not even acknowledge the question and	All the experiments were conducted on a clus-	2012
1946	goes straight to another topic.	ter with 4 NVIDIA A100-SXM4-40GB. The total	2013
1947	Question: Did you eat the last piece of pie?	hours of experimentation for training and inference	2014
1948	Answer: I have to admit that this was a great recipe, I always	(both for zero-shot and fine-tuned models) were	2015
1949	like it when there are chocolate chips in the dough.	230 GPU hours and 440 CPU hours.	2016
1950	Label: Ambivalent		
1951	Explanation: Acknowledges the question but goes on a tangent		
1952	about the chips, without answering.		
1953	Question: Did you enjoy the film?		
1954	Answer: The directing was great.		
1955	Label: Ambivalent		
1956	Explanation: Directing is only part of what constitutes a film.		
1957	Question: What's your favorite film?		
1958	Answer: Fight Club, Filth, and Hereditary.		
1959	Label: Ambivalent		
1960	Explanation: The reply gives three movies instead of one,		
1961	which makes the desired information unclear.		
1962	Question: The hypothesis I was discussing, wouldn't you		
1963	regard that as a defeat?		
1964	Answer: I am not going to prophesy what will happen.		
1965	Label: Clear Non-Reply		
1966	Explanation: Directly stating they won't answer.		
1967	Question: On what precise date did the government order		
1968	the refit of the HMAS Kanimbla in preparation for its forward		
1969	deployment to a possible war against Iraq?		
1970	Answer: I do not know that date. I will find out and let the		
1971	House know.		