"I Never Said That": A Dataset, Taxonomy and Baselines on Response Clarity Classification

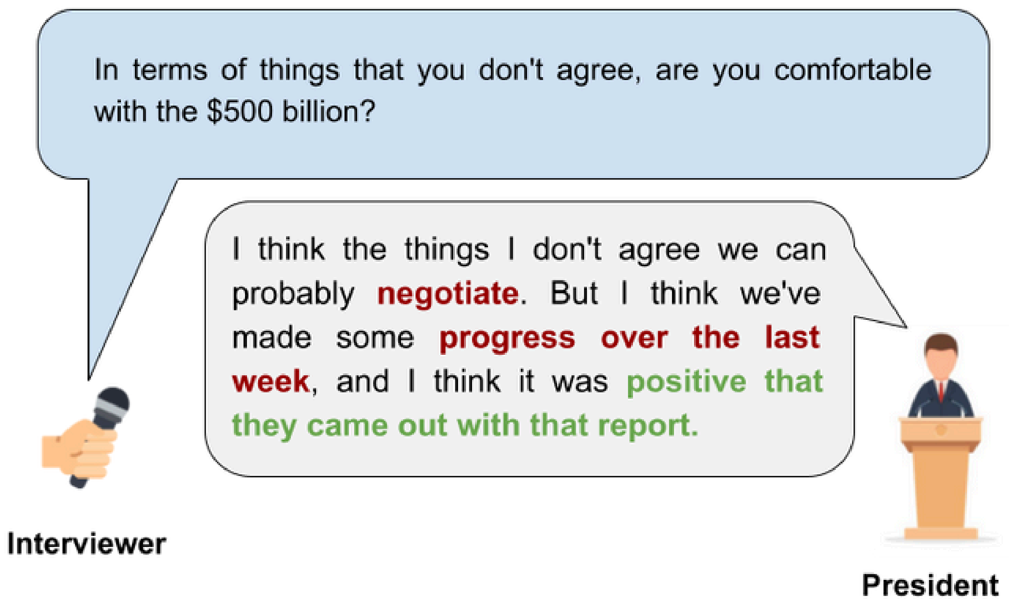
Konstantinos Thomas, Giorgos Filandrianos, Maria Lymperaiou, Chrysoula Zerva, Giorgos Stamou
National Technical University of Athens, Instituto de Telecomunicações,
Instituto Superior Técnico, Universidade de Lisboa, ELLIS Unit Lisbon

We Introduce

- A novel task and dataset on response clarity classification focusing on political interviews.
- A two-level hierarchical taxonomy for clarity classification with different evasion strategies.

Example

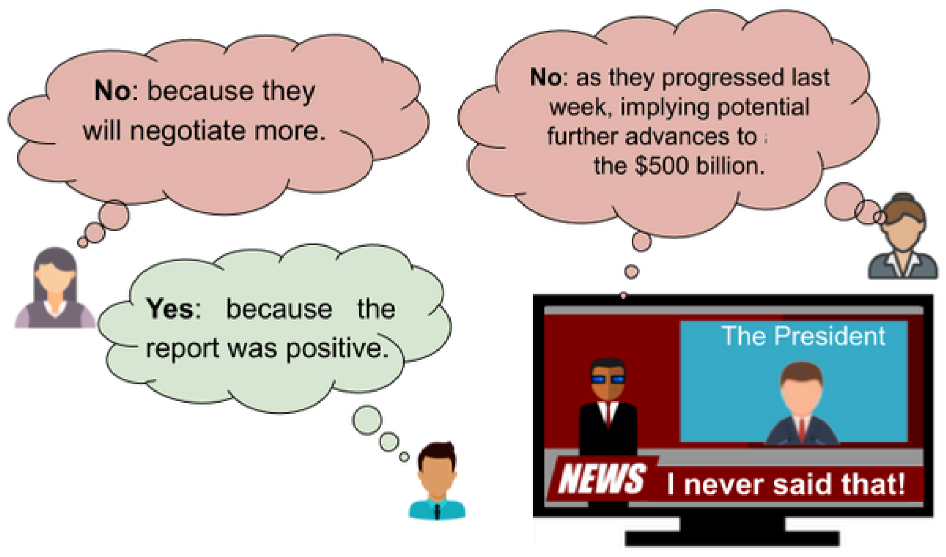
Question & Answer



In terms of things that you don't agree, are you comfortable with the $500 billion?

I think the things I don't agree we can probably **negotiate**. But I think we've made some **progress over the last week**, and I think it was **positive that they came out with that report.**

Possible Interpretations

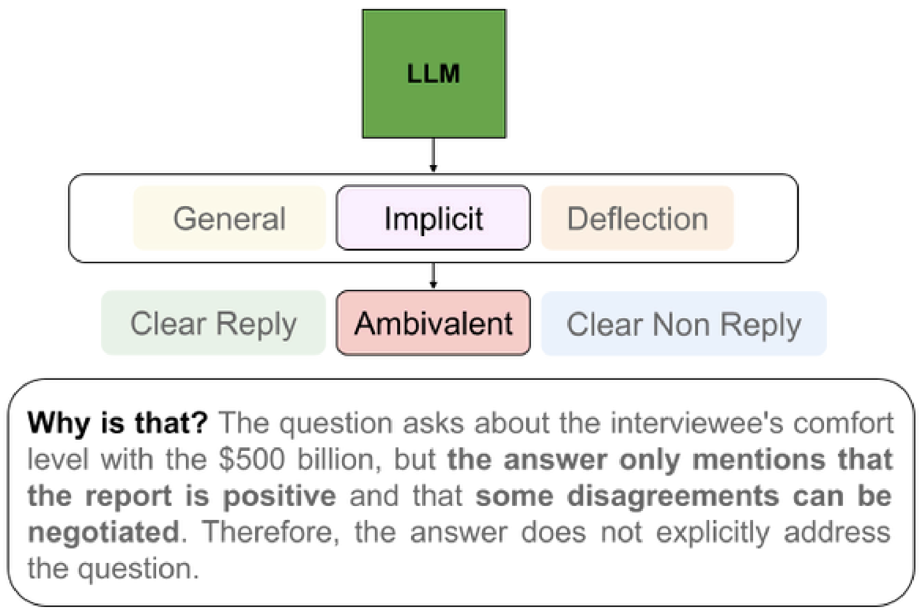


No: because they will negotiate more.

No: as they progressed last week, implying potential further advances to the $500 billion.

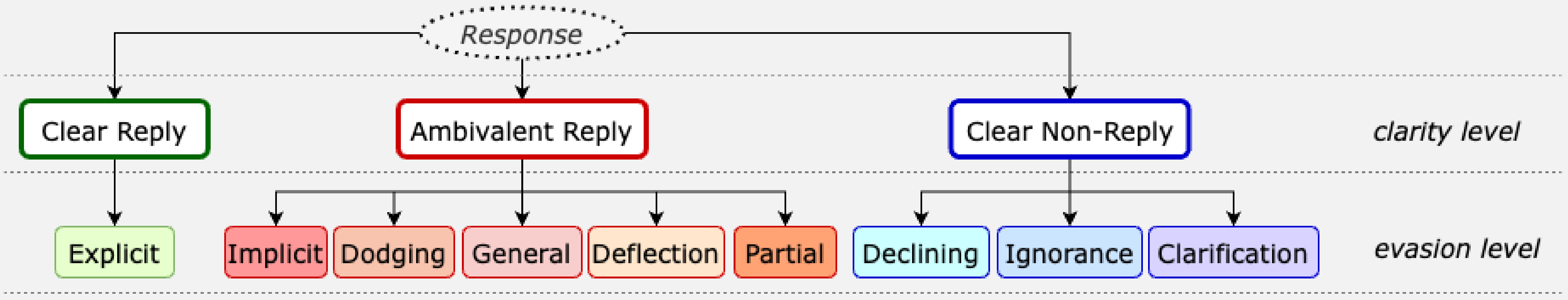
Yes: because the report was positive.

Response Clarity Model



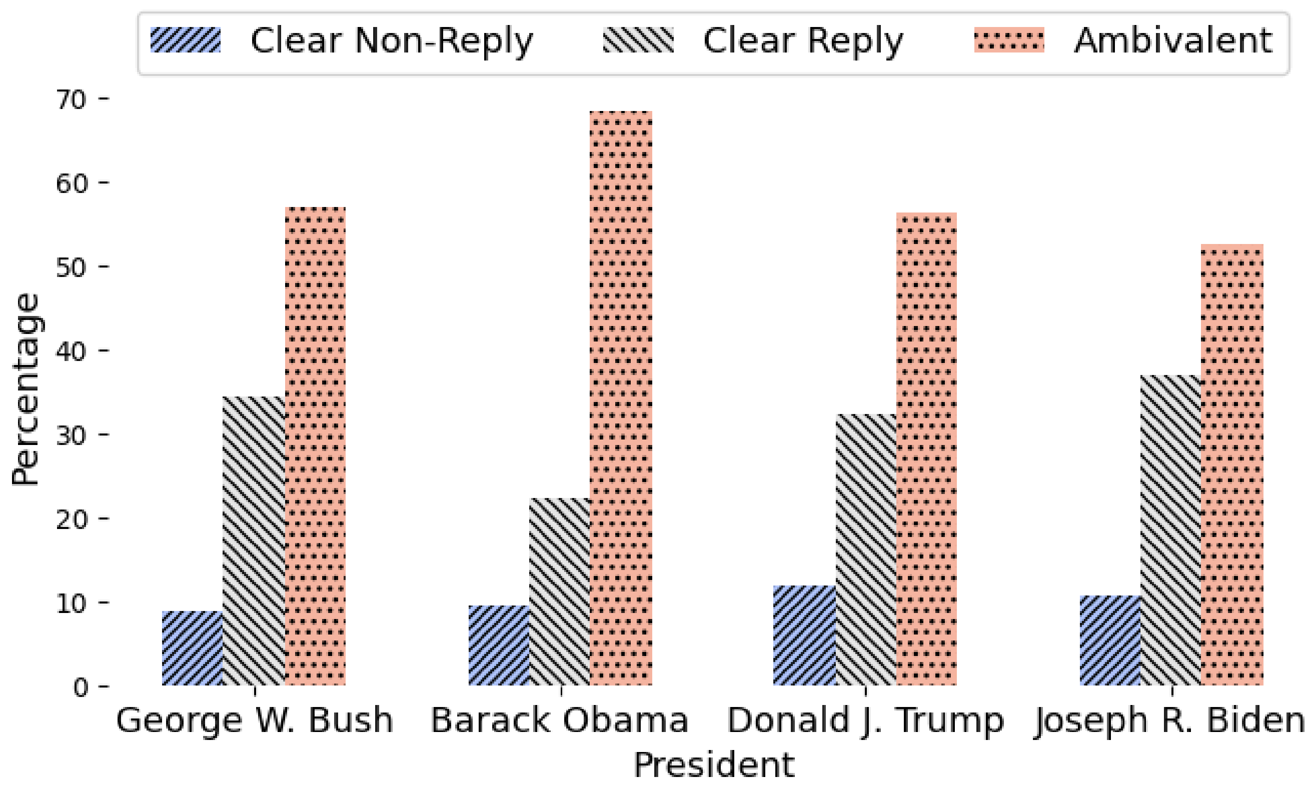
Why is that? The question asks about the interviewee's comfort level with the $500 billion, but **the answer only mentions that the report is positive** and that **some disagreements can be negotiated**. Therefore, the answer does not explicitly address the question.

Taxonomy



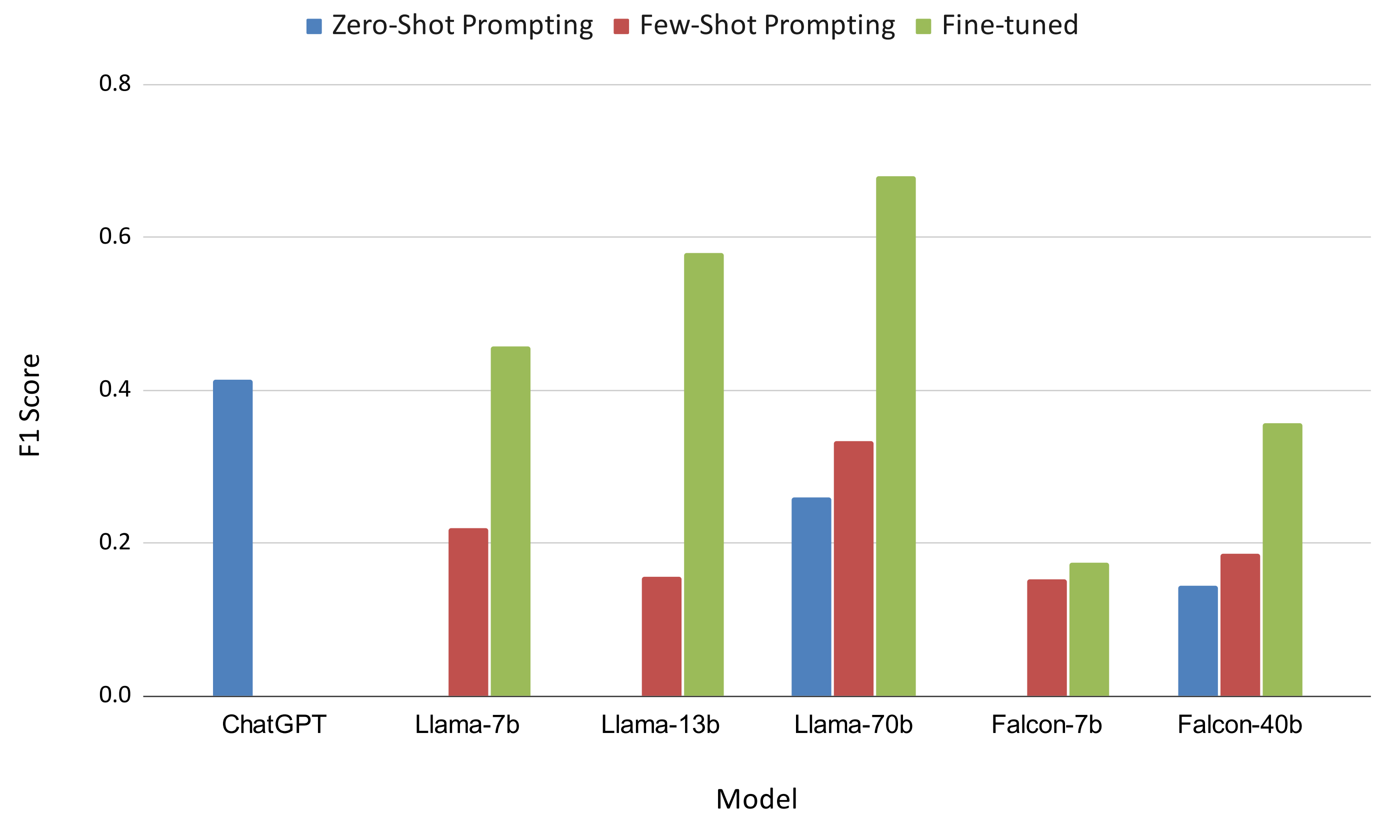
Dataset Analysis

Evasion Rate per President



Dataset Performance Boost

Untuned Prompting vs Fine-Tuned Models



■ Zero-Shot Prompting
■ Few-Shot Prompting
■ Fine-tuned
F1 Score
ChatGPT
Llama-7b
Llama-13b
Llama-70b
Falcon-7b
Falcon-40b
Model