# What Are They Really Answering? QUD Parsing for Political Science

**Nils Grünefeld**
Center for Information and Language Processing, Ludwig Maximilian University of Munich
N.Gruenefeld@campus.lmu.de

## Abstract

Discourse analysis has been a central focus in both theoretical and computational linguistics, yet its potential for political science remains largely untapped. This study explores the application of the Questions Under Discussion (QUD) framework as a method for systematically analyzing political speech, particularly using the example of U.S. presidential debates. QUD parsing provides a structured approach to identifying the implicit questions underlying candidate responses, allowing for a more robust examination of how political figures engage with and redirect moderator inquiries. By applying automated QUD extraction methods, we assess the degree of alignment between stated questions and inferred discourse structures, revealing patterns in rhetorical strategies and response formulation. Furthermore, we integrate text embedding similarity measures to enhance the consistency and scalability of the evaluation process, demonstrating the feasibility of automating QUD-based political discourse analysis. Our findings indicate that the QUD framework effectively captures the underlying logic and coherence of political speech, offering a reproducible method for analyzing debate discourse. These results contribute to the broader study of political communication by providing a linguistically grounded approach to assessing how politicians engage with questions and construct their narratives in high-stakes settings. Our code is available on GitHub.[1]

## 1 Introduction

Discourse-level text analysis has long been a central focus in both classical and computational linguistics, with researchers developing various frameworks to better structure and interpret discourse in a systematic manner. Among these, Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) and Questions Under Discussion (QUD) (Benz and Jasinskaja, 2017) have emerged as particularly influential approaches, offering theoretical foundations that facilitate both the organization of textual content and the extraction of meaningful linguistic relationships. These frameworks have been widely recognized for their utility, not only in theoretical linguistics but also in computational applications, where they have been integrated into a variety of processing pipelines to enhance discourse analysis and improve the interpretability of textual data (Ko et al., 2023).

Despite the demonstrated advantages of discourse analysis frameworks, their application has remained largely confined to the domain of linguistics, with relatively little exploration of their potential utility in other fields. Given their ability to provide structured, reliable, and reproducible methods for analyzing text, such frameworks could offer substantial benefits to disciplines such as political science, where rigorous text analysis is essential for understanding political communication. However, even within the field of natural language processing (NLP), the use of QUD-based methodologies has been notably limited (Wu et al., 2023), and, to the best of our knowledge, no systematic attempt has yet been made to apply QUD parsing in political science.

In this study, we seek to address the broader gap in discourse analysis applications by conducting a pilot investigation into the feasibility of using QUD parsing as a method for analyzing political speech. Our primary hypothesis is that QUD parsing can serve as an effective tool for identifying the implicit questions that underlie candidate responses, thereby providing a structured approach to understanding how political figures engage with and respond to direct inquiries during debates.

To test this hypothesis, we apply our methodology to U.S. presidential debates, which provide a structured and well-documented setting where candidates are explicitly asked questions by moder-

[1] https://github.com/ngruenefeld/qud-for-us-presidential-debates

ators. By employing automatic QUD parsing, we extract the implicit questions that candidates appear to be answering in their responses, allowing us to systematically compare these inferred QUDs with the original moderator questions. This process enables us to assess the degree of alignment between the intended and interpreted discourse structure, ultimately offering a robust, consistent, and reproducible framework for analyzing political speech. Furthermore, in order to enhance the efficiency and reliability of this approach, we explore the possibility of automating the comparison between moderator questions and inferred QUDs by leveraging text embedding similarity measures, which we hypothesize can contribute to a more scalable and consistent methodology for political discourse analysis.

## 2 Background and Related Work

Questions Under Discussion (QUD) parsing is a well-established framework for discourse analysis that has its roots in theoretical linguistics research, with foundational work by Von Stutterheim and Klein (1989); Van Kuppevelt (1995), and more recent developments by Beaver et al. (2017). This framework provides a systematic approach to understanding discourse structure by conceptualizing it as a sequence of implicit questions that are consecutively answered by sentences or statements, allowing for detailed analysis of how information flows through conversation and text.

The field has recently seen significant expansion into computational linguistics applications, with several researchers developing novel approaches to automate and extend the traditional QUD framework. Ko et al. (2023) made substantial progress in developing automated parsers for QUD extraction, while Suvarna et al. (2024) contributed theoretical extensions by incorporating dependency structures between questions, allowing for more nuanced analysis of complex discourse relationships. Work by Wu et al. (2023) has established rigorous automatic evaluation frameworks for QUD parsing, providing essential metrics and methodologies for assessing parser performance, while Wu et al. (2024) have investigated sophisticated methods for evaluating the comparative saliency of different potential QUDs in a given discourse context.

Despite these recent advances in computational applications, research specifically focused on automatic QUD parsing remains relatively limited in scope, with Ko et al. (2023) representing one of the few comprehensive studies that directly addresses the challenges of automated extraction. The majority of recent work in this field, including the evaluation frameworks developed by Wu et al. (2023), has primarily relied on Large Language Models (LLMs) to perform QUD parsing tasks.

## 3 Dataset

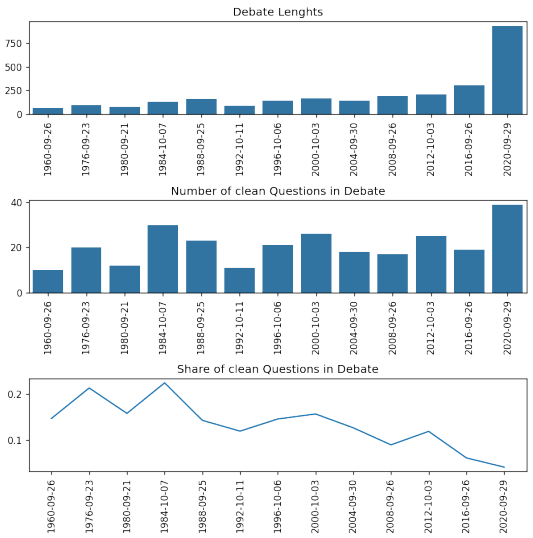



**Figure 1:** Dataset Metadata

Transcripts of all US presidential debates are provided by Martherus (2020) in tabular form, one row containing one utterance by either a moderator or a candidate. To maintain consistency and manage workload, we limit our analysis to the first debate of each election cycle, acknowledging that candidates might adjust their strategies in subsequent debates.

The debates vary in format across election cycles. While all debates comprise a combination of moderator questions, candidate responses to the moderators, and interactions between candidates, they differ significantly in the degree of direct exchanges between candidates. Some debates feature a structured format with minimal cross-talk, whereas others allow for more fluid back-and-forth discussions between participants.

To ensure methodological consistency, our analysis focuses solely on direct moderator questions and the corresponding candidate responses, as these are consistent across all debates. We refer to these as "clean questions". This restriction further enables us to apply the QUD framework more systematically, as it allows us to trace explicit question-

response pairs without the confounding factor of free-form candidate interactions.

A necessary preprocessing step involves manually matching moderator questions with candidate responses. Debate transcripts often include interjections, moderator follow-ups, and overlapping dialogue, requiring careful parsing to extract coherent question-response pairs. This process is conducted manually to ensure accuracy in question-response alignment.

Figure 1 provides an overview of the dataset, including the total duration of each debate, the number of extracted clean questions, and the proportion of clean questions relative to the entire debate. Notably, the share of clean questions decreases consistently over time, reflecting changes in debate structure and candidate interaction. This trend provides valuable insights into how debates have evolved across election cycles.

## 4 Methodology

Our methodology consists of two main components: QUD parsing and evaluating the extracted QUDs.

### 4.1 QUD Parsing

We employ QUD parsing to extract implied questions from candidates' statements during presidential debates. Following Wu et al. (2023), we experimented with several LLMs including Alpaca (Taori et al., 2023), LLaMA (Touvron et al., 2023), and ChatGPT (OpenAI, 2024). For both cost efficiency and accessibility to non-technical researchers, we ultimately selected ChatGPT[2] using few-shot prompting with four examples.

Our approach differs from previous work in a few aspects. Rather than treating QUDs as directed relationships between sentences, we consider them as standalone questions that capture the implicit query being addressed. Furthermore, we expand the scope of QUD parsing beyond single sentences to encompass entire statements, which may span multiple sentences. To facilitate meaningful comparisons between parsed QUDs and moderator questions, we also simplify the latter, as moderator questions tend to be significantly more verbose than parsed QUDs, which could potentially skew automatic evaluation metrics.

The complete set of prompts used for QUD parsing is provided in appendix A.

[2]ChatGPT using version gpt-4o-mini-2024-07-18

### 4.2 Evaluation

Our evaluation framework can be separated into two stages: (1) examining the relationship between QUDs and the answers from which they were parsed, and (2) investigating the alignment between QUDs and the original questions their corresponding answers were meant to address. While previous evaluation frameworks were designed to be domain-agnostic, our approach is specifically tailored to political discourse analysis. Consequently, we developed custom evaluation criteria for each component. The cleaning of questions was validated through manual inspection of a representative subsample.

#### 4.2.1 QUDs vs Answers

We assess the quality of each parsed QUD against its source statement using a three-point scale:

**Poor Alignment (1):** QUD misrepresents or fails to capture the statement's core intent.

**Partial Alignment (2):** QUD captures some key aspects of the statement.

**Strong Alignment (3):** QUD captures the implicit question (nearly) perfectly.

Examples are given in appendix B.

#### 4.2.2 QUDs vs Questions

We evaluate the relationship between each QUD and the original moderator question using a similar three-point scale:

**Poor Alignment (1):** QUD is (nearly) not concerned at all with original question.

**Partial Alignment (2):** QUD lines up partly with the original question.

**Strong Alignment (3):** QUD lines up (nearly) perfectly with the asked question.

Examples are given in appendix B.

To explore the possibility of automating this comparison, we implement two embedding-based approaches using BERT (Devlin et al., 2019) and SentenceBERT (Reimers and Gurevych, 2019). For BERT embeddings, we experiment with three pooling strategies: mean pooling, max pooling, and CLS token embedding. The semantic similarity between QUDs and original questions is then quantified using cosine similarity.

To evaluate the effectiveness of the automatic evaluation, we analyze how well embedding similarities correspond to the manually assigned QUD-Question alignment scores. We examine the relationship between these distributions both qualitatively and statistically. Specifically, we conduct a Student's *t*-test (Student, 1908) to determine whether the embedding similarity distributions significantly differ across alignment score categories, assessing the robustness of automated similarity measures.

Additionally, we investigate the relationship between both embedding similarities and alignment scores and whether or not the candidate ultimately won the election. This analysis allows us to explore potential correlations between debate performance, as measured by alignment and similarity metrics, and electoral success.

# 5 Results

## 5.1 QUD vs Answers

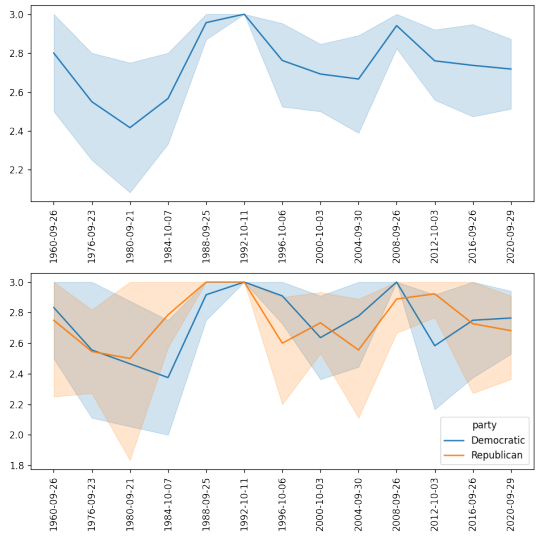



**Figure 2:** QUD-Answer Alignment

All extracted QUDs were coherent and in proper question form. Notably, in some cases, candidate answers contained references to their opponent while remaining on topic - these were still assigned an alignment score of **3**.

The overall effectiveness of our QUD parsing approach is demonstrated by a mean QUD-answer alignment score of **2.7269** (on our 1-3 scale). Breaking this down further, **76.75%** of parsed QUDs achieved strong alignment with their source statements, while **19.19%** showed partial alignment, and **4.06%** exhibited poor alignment.

Figure 2 presents a detailed breakdown of alignment scores across debates and political parties, allowing for comparison of QUD parsing effectiveness across different temporal and partisan contexts.

## 5.2 QUD vs Questions

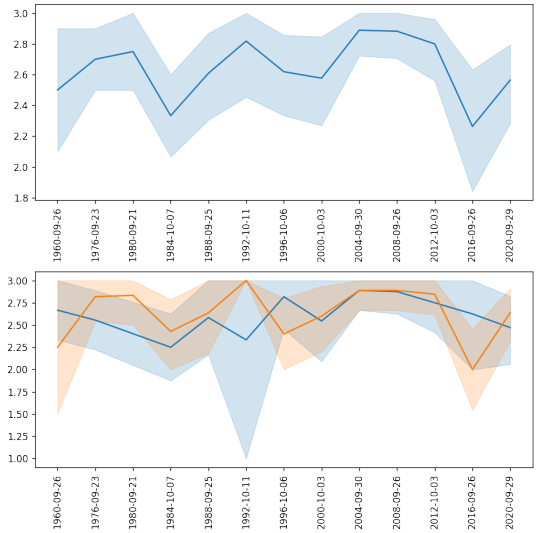


**Figure 3:** QUD-Question Alignment

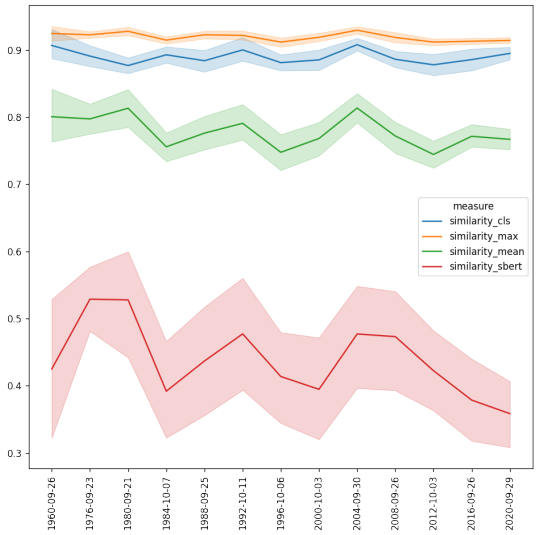



**Figure 4:** Embedding Similarities by Debate

When comparing the parsed QUDs to the original questions their corresponding answers were meant to address, we observed a mean alignment score of 2.6162. The distribution of alignment scores shows that **72.69%** of QUDs achieved strong alignment with the original questions, while **16.24%** showed partial alignment, and **11.07%** exhibited poor alignment. These results are visualized by debate and party in figure 3.

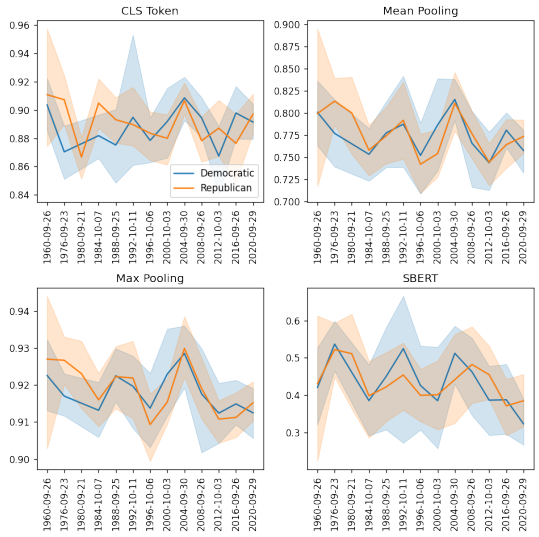



**Figure 5:** Embedding Similarities by Debate and Party

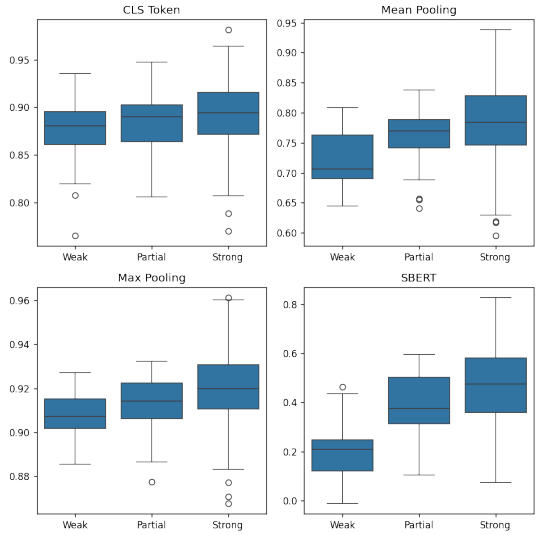



**Figure 6:** Embedding Similarities by QUD-Question Alignment

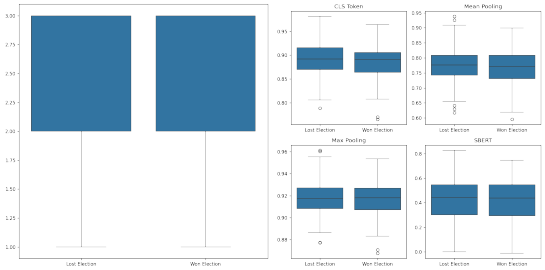
The embedding-based similarity analysis produced varying results across different embedding methods and debates, as shown in figure 4. A breakdown of these similarities by party is presented in figure 5.

The relationship between manual alignment scores and embedding similarities is illustrated in figure 6. As detailed in table 1, *t*-tests comparing the similarity scores across different manual alignment categories revealed significant differences.

Figure 7 presents both alignment scores (figure 7a) and embedding similarities (figure 7b) with respect to electoral outcomes.

| Pooling Method | Alignment | Weak | Partial | Strong |
|---|---|---|---|---|
| CLS Token | Weak | 1.0000 | 0.2127 | 0.0072 |
| | Partial | 0.2127 | 1.0000 | 0.1198 |
| | Strong | 0.0072 | 0.1198 | 1.0000 |
| Mean Pooling | Weak | 1.0000 | 0.0007 | 0.0000 |
| | Partial | 0.0007 | 1.0000 | 0.0161 |
| | Strong | 0.0000 | 0.0161 | 1.0000 |
| Max Pooling | Weak | 1.0000 | 0.0893 | 0.0000 |
| | Partial | 0.0893 | 1.0000 | 0.0012 |
| | Strong | 0.0000 | 0.0012 | 1.0000 |
| SBERT | Weak | 1.0000 | 0.0000 | 0.0000 |
| | Partial | 0.0000 | 1.0000 | 0.0007 |
| | Strong | 0.0000 | 0.0007 | 1.0000 |

**Table 1:** p-values of *t*-tests between QUD-Question Embedding Similarities across Alignments



**(a)** Alignment **(b)** Embedding Similarity

**Figure 7:** Similarity Measures by Election Outcome

## 6 Discussion

### 6.1 QUD vs Answers

The high average alignment between QUDs and their source answers, combined with the absence of significant trends across time periods or political parties, suggests a robust parsing methodology. Particularly noteworthy is the method's effectiveness in handling characteristics of spoken language - fillers such as "uh" and other speech disfluencies do not appear to impact the quality of extracted QUDs. This indicates that our approach successfully bridges the gap between spoken and written discourse analysis.

However, several limitations warrant discussion. The most significant challenge arises when statements address multiple questions simultaneously. In such cases, the parsed QUD typically captures only the primary topic, potentially missing secondary themes or brief references to other questions. For instance, when candidates briefly acknowledge a moderator's question before pivoting to a different topic, the QUD tends to reflect the expanded topic rather than the initial acknowledgment. This is illustrated in the second example in

appendix B, where the alignment ranking becomes ambiguous.

While this might initially appear problematic, it could be argued that responses which only briefly engage with the moderator's question before substantially diverging should indeed be categorized as having partial alignment at best. In this light, the QUDs' tendency to focus on the primary topic of an answer may actually be advantageous for analyzing debate response relevance.

Additional challenges emerge when processing incoherent responses, particularly noticeable in Donald Trump's debate performance. Our method also struggles with instances where candidates explicitly correct or challenge the premises of questions posed to them. Interestingly, we observed that Walter Mondale's responses showed the poorest QUD-Answer alignment scores, though we cannot offer a definitive explanation for this pattern.

Overall, despite these limitations, our QUD parsing approach demonstrates strong performance in capturing the implicit questions being addressed in debate responses. While certain edge cases present challenges, these issues primarily arise in situations where traditional analysis methods would also struggle to interpret the relationship between questions and answers. The method's robustness across different speaking styles and time periods suggests its utility for analyzing political discourse beyond our current application.

### 6.2 QUD vs Questions

Unlike the stable QUD-answer alignment across time, QUD-question alignment shows a slight but noticeable downward trend. This decline is primarily driven by decreasing alignment scores among Republican candidates, while Democratic candidates maintain relatively stable alignment over time. However, this partisan difference is subtle in the manual alignment scores.

This temporal trend becomes more pronounced when examining embedding-based similarity metrics. All pooling methods consistently show this decline, though interestingly, the partisan differences observed in manual alignment scores disappear in the automated analysis. Electoral success appears largely unrelated to both manual alignment scores and embedding similarities, with only BERT CLS token similarities showing marginally lower values for election winners.

The most compelling finding of our analysis, however, is the strong and consistent correlation between manual alignment scores and automated similarity measures. The distributions of embedding similarities differ markedly across alignment categories, with SBERT embeddings showing especially clear differentiation. Statistical analysis through $t$-tests reinforces this observation: all similarity measures except the CLS token embeddings show significant differences across alignment categories, with p-values well below conventional significance thresholds.

These findings have important implications for the automated analysis of political discourse. The strong agreement between manual and automated evaluations enables automatic QUD parsing and evaluation to be effectively employed at scale, opening up possibilities for larger-scale studies of question-response dynamics in political debates while maintaining analytical rigor. Researchers can thus be enabled to efficiently, robustly, and reproducibly examine these dynamics across election cycles, different political contexts, and various types of political discourse beyond presidential debates.

## 7 Conclusion

Our analysis demonstrates that QUD parsing can effectively analyze political discourse in the context of presidential debates. The consistently high QUD-answer alignment scores across different speaking styles and temporal contexts, along with the method's robustness in handling speech disfluencies, suggest that our approach successfully captures the implicit questions being addressed in debate responses. Given this success in handling the varied and often challenging discourse of presidential debates, we believe the method shows strong potential for analyzing political speech more broadly.

However, several limitations and areas for improvement emerged during our investigation. The most significant challenge lies in cases where a single QUD proves insufficient to capture the full complexity of a response. This limitation could potentially be addressed in future work by modifying the framework to allow for multiple QUDs per answer, better reflecting the multi-faceted nature of political discourse. Additionally, while our evaluation criteria provided valuable insights, they may not comprehensively capture all aspects of QUD quality. As noted by Wu et al. (2024), there is inherent difficulty in determining whether a parsed QUD is optimal, as there might always be a more

precise or relevant question to represent the underlying discourse structure.

While our method performed well in analyzing presidential debates, its effectiveness showed some decline as debate formats evolved toward more interactive and less structured exchanges. This observation suggests that the method's utility may vary with the degree of structure in the discourse being analyzed, an important consideration for its application to different forms of political communication.

From an implementation perspective, while our code framework was developed specifically for presidential debates, the underlying approach remains adaptable to different contexts. The main challenge in applying the method to other forms of political discourse would likely stem from the variety of potential data structures and formats in which political speech may be recorded and transcribed. The most significant practical challenge in our study was the manual effort required for matching questions to answers and assigning alignment scores in the dataset preparation phase.

Returning to our initial hypothesis, our results strongly support the feasibility of using QUD parsing as an effective tool for analyzing political speech in general. The high alignment scores between parsed QUDs and their source statements, combined with the strong correlation between manual and automated evaluation metrics, suggest that this approach can indeed provide a structured, reproducible method for analyzing how political figures engage with and respond to questions. While we demonstrated this through the analysis of presidential debates, the method's success in handling various speaking styles and discourse structures indicates its potential utility for analyzing other forms of political communication.

Future research in this direction could focus on developing more sophisticated QUD parsing approaches that can handle multiple questions per response, as well as expanding the evaluation framework to capture additional aspects of QUD quality. Such developments would further strengthen the method's utility for political discourse analysis while maintaining its fundamental advantages of structure and reproducibility.

## 8 Ethical Considerations

In writing this paper, AI (Claude, ChatGPT) was used for drafting text from bullet points, often to overcome writer's block. No part of the final text was adopted from AI without being rewritten or refined. Further, AI assistants (GitHub Copilot, ChatGPT) were used for coding assistance for Python and LaTeX code, respectively.

## References


David I Beaver, Craige Roberts, Mandy Simons, and Judith Tonhauser. 2017. Questions under discussion: Where information structure meets projective content. *Annual Review of Linguistics*, 3(1):265–284.

Anton Benz and Katja Jasinskaja. 2017. Questions under discussion: From sentence to discourse.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Wei-Jen Ko, Yating Wu, Cutter Dalton, Dananjay Srinivas, Greg Durrett, and Junyi Jessy Li. 2023. Discourse analysis via questions and answers: Parsing dependency structures of questions under discussion. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11181–11195, Toronto, Canada. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

James Martherus. 2020. Introducing the transcripts of us presidential debates data set. Available at SSRN: https://ssrn.com/abstract=3611815.

OpenAI. 2024. Gpt-4o system card.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Student. 1908. The probable error of a mean. *Biometrika*, pages 1–25.

Ashima Suvarna, Xiao Liu, Tanmay Parekh, Kai-Wei Chang, and Nanyun Peng. 2024. Qudselect: Selective decoding for questions under discussion parsing.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Jan Van Kuppevelt. 1995. Discourse structure, topicality and questioning. *Journal of linguistics*, 31(1):109–147.

Christiane Von Stutterheim and Wolfgang Klein. 1989. Referential movement in descriptive and narrative discourse. In *North-Holland linguistic series: Linguistic variations*, volume 54, pages 39–76. Elsevier.

Yating Wu, Ritika Mangla, Greg Durrett, and Junyi Jessy Li. 2023. Qudeval: The evaluation of questions under discussion discourse parsing.

Yating Wu, Ritika Rajesh Mangla, Alex Dimakis, Greg Durrett, and Junyi Jessy Li. 2024. Which questions should I answer? salience prediction of inquisitive questions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19969–19987, Miami, Florida, USA. Association for Computational Linguistics.

## A

You are a QUD parsing expert.
Your task is to identify the implicit question that the speaker is answering in the following statement. The question should reflect the main purpose of the statement.

Examples for this question generation are:

**Statement:** Well, Sander, thats a good question, and the answer is; for 40- some years we kept the peace. If you look at the cost of not keeping the peace in Europe, it would be exorbitant. We have reduced the number of troops that are deployed and going to be deployed. I have cut defense spending. And the reason we could do that is because of our fantastic success in winning the Cold War. We never would have got there if we had gone for the nuclear freeze crowd; we never would have got there if we had listened to those that wanted to cut defense spending. I think it is important that the US stay in Europe and continue to guarantee the peace. We simply cannot pull back. Now, when anybody has a spending program they want to spend money on at home, they say, well, lets cut money out of the Defense Dept. I will accept and have accepted the recommendations of 2 proven leaders, General Colin Powell and Secretary Dick Cheney. They feel that the levels were operating at and the reductions that I have proposed are proper. And so I simply do not think we should go back to the isolation days and starting blaming foreigners. We are the sole remaining superpower, and we should be that. And we have a certain disproportionate responsibility. But I would ask the American people to understand that if we make imprudent cuts, if we go too far, we risk the peace. And I dont want to do that. Ive seen what it is like to see a war, to see the burdens of a war, and I dont want to see us make reckless cuts. Because of our programs we have been able to significantly cut defense spending. But lets not cut into the muscle, and lets not cut down our insurance policy, which is participation of American forces in NATO, the greatest peace- keeping organization ever made. Today youve got problems in Europe, still bubbling along even though Europes gone democracys route. But we are there, and I think this insurance policy is necessary. I think it goes with world leadership, and I think the levels weve come up with are just about right.
**Question:** Why is it important for the United States to maintain its current level of defense spending and military presence in Europe?

... [3 more in-context examples]

Please generate one question and one question only without any prefaces.

**Table 2:** Few-shot prompt for QUD generation

You are a QUD parsing expert.
Your task is to clean and extract the central question from a moderator's statement.
The statement may already be clear and concise. In that case, simply repeat it.
The statement may be verbose and contain irrelevant information. In that case, remove the irrelevant information.
Stay as close to the original statement as possible.

Examples for this question generation are:

**Statement:** Mr. Vice President, your campaign stresses the value of your eight year experience, and the question arises as to whether that experience was as an observer or as a participant or as an initiator of policy- making. Would you tell us please specifically what major proposals you have made in the last eight years that have been adopted by the Administration?
**Question:** What major proposals you have made in the last eight years that have been adopted by the Administration?

**Statement:** New question. Are there issues of character that distinguish you from Vice President Gore?
**Question:** Are there issues of character that distinguish you from your opponent?

**Statement:** Mr. Vice President, Im struck by your discussion of women and the sanctity of life. And it leads me to recall your own phrase, that you are haunted by the lives which children in our inner cities live. Certainly the evidence is compelling. Theres an explosion of single parent families. And by any measure, these single parent families, many with unwanted children, are the source of poverty, school drop outs, crime, which many people in the inner city simply feel is out of control. If it haunts you so, why over the eight years of the Reagan- Bush administration have so many programs designed to help the inner cities been eliminated or cut?
**Question:** Why over the eight years of the your administration have so many programs designed to help the inner cities been eliminated or cut?

**Statement:** How do you bring back— specifically bring back jobs, American manufacturers? How do you make them bring the jobs back?
**Question:** How do you bring back American manufacturing jobs?

Please generate one question and one question only without any prefaces.

**Table 3:** Few-shot prompt for question simplification

**B**

| |
|---|
| **Moderator Question:** "Mr. President, I would like to continue for a moment on this uh question of taxes which you have just raised. You have said that you favor more tax cuts for middle- income Americans – even those earning up to $30 thousand a year. That presumably would cost the Treasury quite a bit of money in lost revenue. In view of the very large budget deficits that you have accumulated and that are still in prospect, how is it possible to promise further tax cuts and to reach your goal of balancing the budget?" |
| **Candidate Answer:** "At the time, Mr. Gannon, that I made the recommendation for a $28 billion tax cut – three- quarters of it to go to individual taxpayers and 25 percent to American business. I said at the time that we had to hold the lid an federal spending, that for every dollar of a tax reduction we had to have an equal reduction in federal expenditures – a one- for- one proposition. And I recommended that to the Congress with a budget ceiling of three hundred and ninety-five billion dollars, and that would have permitted us to have a $25 billion tax reduction. In my tax reduction program for middle- income taxpayers, I recommended that the Congress increase personal exemptions from seven hundred and fifty dollars per person to one thousand dollars per person. That would mean, of course, that for a family of four that that family would have a thousand dollars more personal exemption – money that they could spend for their own purposes, money that the government wouldnt have to spend. But if we keep the lid on federal spending, which I think we can – with the help of the Congress, we can justify fully a $28 billion tax reduction. In the budget that I submitted to the Congress in January this year, I re- recommended a 50 percent cutback in the rate of growth of federal spending. For the last ten years the budget of the United States has grown from uh – about 11 percent per year. We cant afford that kind of growth in federal spending. And in the budget that I recommended we cut it in half – a growth rate of 5 to 5 and one- half percent. With that kind of limitation, on federal spending, we can fully justify the tax reductions that I have proposed. And it seems to me with the stimulant of more money in the hands of the taxpayers, and with more money in the hands of business to expand, to modernize, to provide more jobs, our economy stimulated so that well get more revenue and well have a more prosperous economy." |
| **Cleaned Question:** "How is it possible to promise further tax cuts for middle-income Americans and to balance the budget?" |
| **QUD:** "How can a $28 billion tax cut be justified while effectively controlling federal spending?" |
| **QUD-Answer Alignment:** Strong (3) |
| **QUD-Question Alignment:** Strong (3) |

**Table 4:** QUD Parsing Example

**Moderator Question:** "Let me just follow up one quick question. When you hear Vice President Gore question your experience, do you read it the same way, that hes talking about policy differences only?"
**Candidate Answer:** "Yes. I take him for his word. Look, I fully recognize Im not of Washington. Im from Texas. And hes got a lot of experience, but so do I. And Ive been the chief executive officer of the second biggest state in the union. I have a proud record of working with both Republicans and Democrats, which is what our nation needs. Somebody that can come to Washington and say lets forget all the finger pointing and get positive things done on Medicare, prescription drugs, Social Security, and so I take him for his word."
**Cleaned Question:** "Do you consider Vice President Gore's questioning of your experience to be focused only on policy differences?"
**QUD:** "Why should voters trust your ability to work across party lines and effectively address issues like Medicare and Social Security?"

**QUD-Answer Alignment:** Partial (2)
**QUD-Question Alignment:** Weak (1)

**Table 5:** QUD Parsing Example

| |
|---|
| **Moderator Question:** ""<br>**Candidate Answer:** ""<br>**Cleaned Question:** ""<br>**QUD:** ""<br><br>**QUD-Answer Alignment:** Strong (2)<br>**QUD-Question Alignment:** Weak (1) |

**Table 6:** QUD Parsing Example