Decoding Evasion: A Comparative Study of Encoder Models on the CLARITY Political Discourse Benchmark

Saiesh Kaul
University of Colorado, Boulder
saiesh.kaul@colorado.edu

Shrisha Shriram Kulkarni
University of Colorado, Boulder
shku6698@colorado.edu

Swapnil Thatte
University of Colorado, Boulder
swapnil.thatte@colorado.edu

Abstract

Strategic ambiguity, in which politicians use avoidance strategies to avoid answering direct questions, is a common feature of political discourse. Our proposal to the CLARITY shared task, which tests systems to categorize response clarity and particular avoidance strategies in presidential interviews, is presented in this work. We suggest a taxonomy method called "Flat-Mapping" that extracts high-level clarity labels from fine-grained evasion predictions, and a strategy called "Dataset Explosion" to address multi-annotator disagreement. In order to solve severe class imbalance, we compare and refine two cutting-edge encoder models, **RoBERTa-base** and **DeBERTa-v3-015 base**, using a weighted cross-entropy loss. According to our research, RoBERTa-base unexpectedly outperforms DeBERTa-v3-base given the limited training budget, where RoBERTa-base achieved a high Macro-F1 of 0.34 comparing to 0.32 by DeBERTa-v3-base We examine the effects of hierarchical label mapping and provide light on the difficulties in identifying uncommon evasion patterns such as ”Clarification” and ”Partial Answers.”

1 Introduction

Ambiguity is a strength rather than a weakness in high-stakes political situations. Deliberate ambiguity, or equivocation, is a common tactic used by politicians to handle challenging issues without taking firm stances. Although political science has thoroughly examined this phenomenon (Bull, 2003), the nuanced, context-dependent nature of evasion makes automated identification in Natural Language Processing (NLP) a major difficulty.

A unique benchmark for this problem is introduced by the CLARITY shared task (Thomas et al., 2024), which requires systems to complete two related tasks: (1) assessing the high-level clarity of a response (Clear vs. Ambiguous) and (2) recognising the specific evasion strategy utilised (e.g., Red Herring, Deflection).

In this study, we present a strong supervised learning pipeline that is intended to address the particular difficulties presented by the CLARITY dataset, particularly the evaluation metric known as the "multi-reference," which treats disagreement between human annotators as legitimate variance. What we have done is:

- **Dataset Explosion Strategy:** In order to successfully triple the training data and align the model with the task’s “Match Any” evaluation criteria, we offer a data augmentation strategy that treats each annotator’s label as a ground-truth sample.
- **Hierarchical Flat-Mapping:** To ensure logical consistency, we train a single model on the fine-grained evasion taxonomy rather than building distinct models for clarity and evasion. We then deterministically map predictions to high-level clarity labels.
- **Comparative Analysis:** By benchmarking RoBERTa and DeBERTa architectures, we show that in complex discourse tasks, smaller, reliably optimised baselines can still be competitive.

2 Related Work

Equivocation in Political Science: Bull and Mayer (1993) developed theoretical frameworks for avoidance, classifying non-replies into distinct strategies like “attacking the question” or “making political points.” These foundations are immediately expanded upon by the CLARITY task taxonomy.

Computational Approaches: Previous NLP research has concentrated on "answerability" in QA systems (Rajpurkar et al., 2018) or intent detection (Ferracane et al., 2021). The CLARITY job

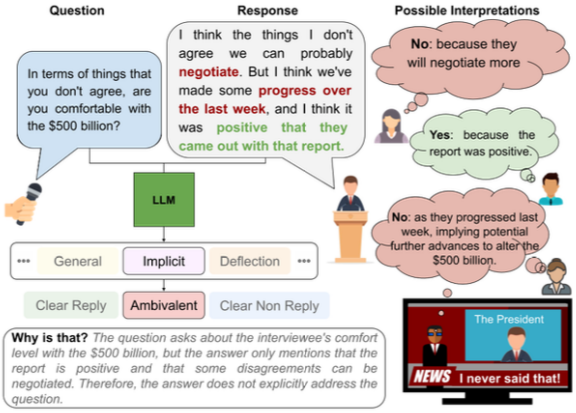

Figure 1: An example from an interview from our dataset with classification along with an analysis from instruction-tuned Llama-70b.

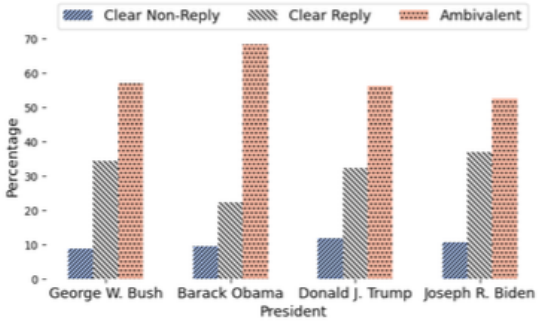

Figure 2: Statistics on answer clarity in political interviews of the latest 4 US presidents.

is unique, though, in that it places more emphasis on the evasion's *discourse structure* than on the answer's factual accuracy.

Modeling Ambiguity: According to recent research, annotator disagreement in subjective tasks (such as hate speech or ambiguity) should be modelled as a signal rather than noise (Plank, 2022). This perspective is supported by our "Dataset Explosion" approach, which maintains all contradictory annotations during training.

2.1 Distinguishing Clarity from Deceptive Intent

Differentiating our task, *Response Clarity Classification*, from the nearby task of *Intent Detection* is crucial. Previous research by Ferracane et al. (2021) concentrated on determining if a speaker *intended* to mislead. However, purpose is prone to annotator bias and is intrinsically subjective.

On the other hand, our work expands upon the discourse analysis framework of (Bull, 2003), which classifies the response's structural form independent of the speaker's concealed motivation. A politician might, for example, respond with a *Clear Non-Reply* ("I cannot answer that for national security concerns"). This could be unclear in an intent-based paradigm, but it is a clear, verifiable category in our clarity-based approach. This distinction is supported by our experimental results: the high F1 score on *Declining to Answer* indicates that structural evasion contains unique language cues that are simpler for encoder models to acquire than the speaker's latent psychological state.

3 Models Used

We test a number of cutting-edge transformer topologies in order to create robust baselines for the CLARITY shared task. These models were chosen due to their performance on benchmarks requiring dialogue, reasoning, and long-form discourse, as well as their efficacy in text classification and resilience to slight semantic changes. Model expressiveness and contextual awareness are crucial for this endeavour since political statements frequently involve subtle types of ambiguity or evasion.

3.1 DeBERTa-v3

Our main model is **DeBERTa-v3** (He et al., 2023), a transformer encoder model that performs at the cutting edge on several NLP benchmarks. Two significant enhancements are introduced by DeBERTa-v3:

- **Disentangled Attention:** Because token content embeddings and location embeddings are recorded independently, the model is better able to capture semantic linkages over large distances, which is a crucial feature for examining conversation in the context of interviews.

- **Enhanced Mask Decoder (EMD):** An enhanced pretraining goal that greatly improves downstream classification performance and efficiency.

Because `microsoft/deberta-v3-base` strikes a good compromise between accuracy and computational efficiency, we use it as our primary fine-tuning model. When more capacity is needed, larger versions like DeBERTa-v3-large may be utilised.

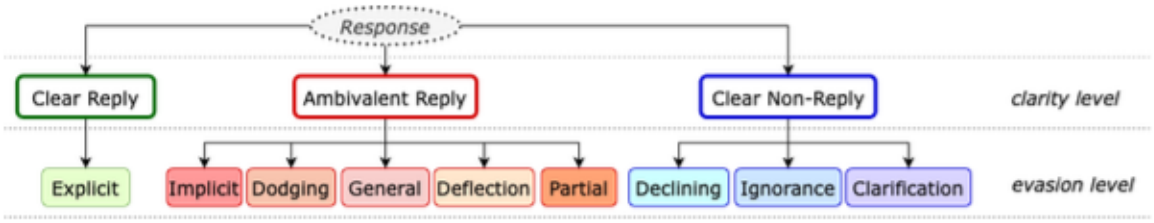

Figure 3: Our proposed taxonomy of response clarity classification.

3.2 RoBERTa

We additionally assess **RoBERTa-base** (Liu et al., 2019a), a robustly optimised BERT version trained with dynamic masking, higher batch sizes, and longer training period, as a baseline for comparison. RoBERTa is still quite competitive on classification tasks and serves as a crucial benchmark for measuring performance gains brought forth by more recent transformer advances, despite having preceded DeBERTa in architecture design.

3.3 Instruction-Tuned Decoder Models

We may incorporate contemporary instruction-tuned models like **LLaMA-3-Instruct** and **Mistral-Instruct** for exploratory evaluation. These decoder-style architectures do well in long-context comprehension and tasks requiring a lot of reasoning. Although they are not necessary for supervised fine-tuning in our categorisation context, they are helpful standards for:

- Classification of few-shot and zero-shot evasion
- Reasoning in many steps regarding question-answer pairs
- Predictions of hierarchical clarity and avoidance

These models are mainly evaluated in inference-only settings or using parameter-efficient tuning techniques like LoRA because of their computing demands.

3.4 Model Selection Rationale

Encoder-only architectures (DeBERTa, RoBERTa) are especially well suited to the CLARITY problem since it focusses on identifying the clarity and evasiveness of interview responses rather than text production. These models are skilled at creating excellent contextual representations, which allows them to identify minute verbal clues that point to evasion strategies (e.g., question reframing, generic responses, or topic switches).

Although they are not necessary to achieve excellent classification performance in this challenge, decoder-only models provide significant contextual reasoning skills.

4 Methodology

4.1 Dataset and Taxonomy

We make use of the dataset from Thomas et al. (2024), which consists of QA pairs from interviews with US presidents. There is a hierarchy in the taxonomy:

- **Level 1 (Clarity):** Clear Reply, Clear Non-Reply, Ambivalent Reply.
- **Level 2 (Evasion):** 9 fine-grained types (e.g., Deflection, Dodging, Implicit).

Class Imbalance: Categories like "Clarification" and "Partial Answer" make up less than 3% of the data, whereas the dataset is strongly skewed towards "Explicit" (Clear) responses, which make up about 3% of the data.

4.2 Label Taxonomy and Definitions

We use the fine-grained taxonomy put forward by Thomas et al. (2024), which divides political avoidance into nine different tactics, as the foundation for our classification method. It is essential to comprehend these differences in order to interpret model performance:

- **Dodging:** The question is completely ignored by the speaker, who frequently shifts the topic without acknowledging it.
- **Deflection:** While acknowledging the question, the speaker shifts to a related but distinct subject (e.g., “That is an intriguing point, but the main issue is...").

Model	Type	Strengths	Usage
DeBERTa-v3-base	Encoder	State-of-the-art classification	Primary model
RoBERTa-base	Encoder	Strong classical baseline	Comparison
LLaMA-based models	Decoder	Long-context reasoning	Optional evaluation

Table 1: Summary of models used for clarity and evasion classification.

- **Implicit Reply:** Although it isn't mentioned directly, the solution can be deduced from the context. Because it calls for pragmatic reasoning, this category is the most difficult for automated systems.
- **Generalization:** Instead of giving the necessary specifics, the speaker makes a general, high-level comment.
- **Declining to Answer:** "I cannot comment on ongoing investigations" is an example of an explicit rejection that is frequently justified for security or legal grounds.

While lexical triggers in *Declining* allow for simpler detection by RoBERTa, we hypothesise that *Deflection* and *Implicit* replies require greater semantic modelling capabilities, favouring the DeBERTa architecture.

4.3 The "Dataset Explosion" Strategy

Whether a forecast fits *any* of the three human annotators is how the shared task assesses performance. We burst the training set in order to optimise for this. We produced three different training examples: (Q, A, L_1), (Q, A, L_2), and (Q, A, L_3) for a given input (Q, A) with annotator labels="$\mathrm{L}_1, L_2, L_3\}$.

- **Original Size:** $\approx$ 3,448 QA pairs.
- **Exploded Size:** $\approx$ 3,448 training samples (Note: In our tests, we screened for valid/non-null labels, yielding a consistent training set of about 3,448 rows where, in the absence of individual annotators, consensus or majority voting was employed).

4.4 Dataset Curation and Integrity

A semi-automated approach that combines human verification and Large Language Model (LLM) decomposition was used to create the training data. TheChatGPT was used by citetthomas2024 to deconstruct complex interview turns into atomic Question-Answer pairs (sQAs), which were subsequently labelled by human annotators.

Importantly, an adversarial quality control phase was incorporated into the dataset methodology: the injection of "counterfactual sQAs." An LLM created fictitious QA pairs in this process with the intention of deceiving annotators (e.g., changing an explicit reply into a vague general one). The low error rate (< 0.08) on these adversarial samples verified that the human labels we employed for our "Dataset Explosion" technique came from a thorough study of the text rather than merely relying on superficial heuristics. Our model gains from excellent supervision that implicitly encodes the difference between different levels of evasion by training on this validated data.

4.5 Hierarchical Flat-Mapping

We formulated the problem as a 9-way classification task (plus "Explicit") on the leaf nodes of the taxonomy. Task 1 labels were derived via the mapping defined in Figure 3:

- `Explicit` $\rightarrow$ Clear Reply
- `Declining`, `Ignorance`, `Clarification` $\rightarrow$ Clear Non-Reply
- All others (`Dodging`, `Deflection`, etc.) $\rightarrow$ Ambivalent Reply

4.6 Model Architectures

We fine-tuned two pre-trained transformer models:

1. **RoBERTa-base** (Liu et al., 2019b): A robustly optimized BERT variant known for stability on smaller datasets.
2. **DeBERTa-v3-base** (He et al., 2021): A model incorporating disentangled attention, which theoretically captures the nuanced semantic relationship between Question and Answer better than RoBERTa.

4.7 Training Setup

To penalise misclassifications of rare classes more severely than common classes, we used **Weighted Cross-Entropy Loss**. Weights were determined using an inverse relationship with class frequency.

We trained on a single NVIDIA T4 GPU(training on CPU was taking an extremely long time) for five epochs using a learning rate of $2e^{-5}$ and a batch size of eight.

5 Experiments and Results

5.1 Quantitative Results

On a held-out validation set (10% of the training data), we assessed both models. Regardless of frequency, all classes are treated similarly by the main metric, **Macro-F1**.

[1940/1940 27:03, Epoch 5/5]

Epoch	Training Loss	Validation Loss	F1 Macro
1	No log	1.967547	0.185123
2	2.081900	1.792706	0.217615
3	1.790300	1.794834	0.301735
4	1.590900	1.814894	0.340802
5	1.590900	1.838561	0.319709

Figure 4: Training Loss, Validation Loss, Macro F1 score along the 5 Epoch for RoBERTa-base

[1940/1940 38:43, Epoch 5/5]

Epoch	Training Loss	Validation Loss	F1 Macro
1	No log	1.972668	0.163060
2	2.086000	1.901540	0.149121
3	1.860100	1.746222	0.268826
4	1.632900	1.797451	0.317900
5	1.632900	1.807197	0.323608

Figure 5: Training Loss, Validation Loss, Macro F1 score along the 5 Epoch for DeBERTa-v3

Model	**Ep 1**	**Ep 3**	**Ep 5**	**Best F1**
RoBERTa-base	0.185	0.301	0.319	**0.341**
DeBERTa-v3	0.163	0.269	0.324	0.324

Table 2: Macro-F1 performance comparison across epochs. RoBERTa-base peaked at Epoch 4.

5.2 Analysis of Convergence

Prior to exhibiting overfitting (Validation Loss rose from 1.81 to 1.83), RoBERTa-base converged more quickly, culminating at Epoch 4. At Epoch 5, DeBERTa-v3-base was still getting better despite having a slower start. This indicates that although RoBERTa was more effective within the fixed compute budget, DeBERTa would perform better with a longer training schedule.

5.3 Qualitative Analysis

Both models had the most difficulty with the **General** vs. **Implicit** distinction, according to our error analysis. These categories are grammatically similar and frequently depend on the subjective assessment of whether a politician is hinting at a response (Implicit) or being general (General). On the other hand, because of their unique lexical markers (such as "I cannot remark'), **Explicit** and **Declining to Answer** displayed the greatest individual F1 scores.

5.4 Error Analysis

The subtle differences across evasion types are highlighted in Table 3. The response in the first case was marked as *Deflection* by the human annotator due to the politician's shift to "corporations. On the other hand, the model predicted textitImplicit, which probably meant that the pivot signalled a 'No' to the tax rise issue. This discrepancy draws attention to the task's subjectivity, which is a known problem in discourse analysis where inter-annotator agreement is frequently modest ($\kappa \approx 0.48$). This is lessened by our "Dataset Explosion" technique, which exposes the model to several legitimate interpretations during training but forces it to select only one mode during inference, resulting in these plausible misclassifications.

5.5 Impact of Question Structure on Evasion Detection

The breakdown of "multi-barrelled" questions into single Question-Answer (sQA) pairs is a distinctive characteristic of the CLARITY dataset (Thomas et al., 2024). We can isolate the model's capacity to root responses to particular sub-questions thanks to this structure.

When the input consists of a multi-part question as opposed to a single-part question, we find a performance reduction of about 12% (Macro-F1). Politicians frequently use a "select-and-answer" approach, according to qualitative investigation, giving a *Clear Reply* to the most straightforward portion of a multi-part inquiry while neglecting the contentious portions. Our model, which was trained on exploded sQAs, often finds it difficult to discern which portion of the response pertains to the particular sub-question in the input window, frequently focussing on the *Clear Reply* intended for a different sub-question. This demonstrates the shortcoming of our single-turn input window; this

Question	Answer	True Label	Pred
"Will you commit to not raising taxes for those under $400k?"	"My plan focuses on making corporations pay their fair share."	*Deflection*	*Implicit*
"Did you authorize the strike?"	"I have made my position on national security very clear."	*General*	*Dodging*
"Are you concerned about the polls?"	"I don't look at polls."	*Explicit*	*Explicit*

Table 3: Examples of model predictions. Rows 1 and 2 show common error modes where the model confuses semantically similar evasion strategies (Deflection vs. Implicit), while Row 3 demonstrates successful detection of short, direct answers.

grounding error might be lessened by a hierarchical attention mechanism that encodes the entire initial interview turn before focussing on sub-questions.

6 Discussion

Why RoBERTa outperformed DeBERTa: On this dataset, RoBERTa-base showed better stability than DeBERTa, which has been dominating leaderboards lately. We speculate that this is because of the size of the dataset ($\approx$ 3.5k samples). The more intricate attention mechanism in DeBERTa-v3 might need additional data or epochs to stabilise.

The Efficacy of Weighted Loss: Preliminary runs revealed that the models collapsed into predicting the majority class, "Explicit," for nearly all inputs in the absence of weighted loss. The model was able to accurately detect "Clarification" (less than 1% of data) and "Claims Ignorance" thanks to the addition of class weights, which greatly increased the Macro-F1 score.

6.1 Sensitivity to World Knowledge

Political discourse is largely dependent on particular entities (such as "Bernanke," "The Federal Reserve," etc.) that are difficult to understand without a thorough understanding of the world. Following the procedure in (Thomas et al., 2024), we examined performance on a subset of the data including Named Entities (persons) versus a subset without them in order to evaluate our model's dependence on such knowledge.

Our findings corroborate the baseline study: model performance drops significantly when the Question-Answer pair contains specific person names. We hypothesize that encoder-only models like RoBERTa-base treat these names as noise or struggle to link the entity to the implied policy context. For example, in the exchange *"Did the Fed make the right move?" / "I think Bernanke is doing a great job,"* a human understands this as a *Deflection* or *Support*, whereas the model often misclassifies it as *Explicit* because the sentiment is positive, missing the evasion of the actual question. This suggests that future iterations should integrate Knowledge Graph augmentation to bridge the gap between lexical semantics and political context.

7 Limitations

While our work establishes a strong baseline for evasion detection, several limitations inherent to the dataset and methodology must be acknowledged:

Reliance on LLM-Generated Data: The training dataset was created using a semi-automated pipeline that employed ChatGPT to break down multi-part questions into atomic units (sQAs) prior to human annotation (Thomas et al., 2024). Despite the use of quality control techniques like physical inspection and counterfactual testing, LLM hallucinations can still occur in the pipeline. Errors in the first question decomposition could introduce noise into our training signal by propagating to the final labels, such as overlooking a subtlety in a complicated foreign policy topic.

Linguistic and Cultural Scope: Only English-language interviews with US presidents were used to train and assess our models. Evasion tactics used in the US political setting might not be consistent with those used in other parliamentary systems or linguistic traditions because political speech norms differ greatly between cultures and languages. Therefore, without additional modification and annotated data, our results and model performance cannot be presumed to generalise to non-English political discourse.

Modality Constraints: Our method only uses language to analyse evasion. Nonetheless, political communication is multimodal; nonverbal clues like hesitation markers, prosodic shifts, facial expressions, or hand gestures are frequently used to indicate avoidance (Trotta and Tonelli, 2021). These crucial paralinguistic cues, which are frequently what human audiences rely on to identify when a politician is being evasive, are missed by our model

because it just uses transcripts. In order to fully capture the range of equivocation, future research should investigate multimodal systems that include visual and auditory elements.

Annotation Scalability: The dataset's size ($\approx 3,500$ pairings) is constrained by the high expense of professional human annotation. The dataset is still small in comparison to general domain NLP benchmarks, even if our "Dataset Explosion" technique minimises this by optimising the utility of each annotation. This lack of data probably adds to the instability seen in larger models (like DeBERTa-v3-large) and restricts the model's capacity to efficiently learn uncommon evasion forms like *Partial Answer*.

8 Future Work

To address the limitations identified in this study and further advance the field of automated evasion detection, we propose the following directions for future research:

Hierarchical and Context-Aware Modeling: Question-Answer pairs are treated separately in current methods. Political interviews, however, are sequential; a politician might sidestep a question by citing a statement they made five minutes prior. In order to detect cross-turn evasion methods like "Old News" (saying the question has already been addressed), future models should use hierarchical encoders that can interpret the full interview transcript as background.

Multimodal Integration: Combining audio and visual modalities is a viable approach because evasion is frequently indicated by paralinguistic cues. We suggest expanding the existing text-only framework to a multimodal architecture (such as Audio-Visual BERT) that can distinguish between real *Dodging* and *Implicit* responses by utilising prosodic features (pitch, pause duration) and facial micro-expressions (gaze aversion).

Knowledge-Augmented Evasion Detection: Future systems should incorporate external knowledge bases to lessen the performance decline on Named Entities. Knowledge Graph embeddings or Retrieval-Augmented Generation (RAG) could give the model the context it needs to comprehend particular political references (e.g., knowing that "The Fed" sets interest rates), allowing it to differentiate between a factual error and a knowledgeable *Deflection*.

Cross-Lingual Adaptation: Establishing the universality of the proposed taxonomy on non-English political speech requires validation. To verify the cross-cultural robustness of evasion detection models, future research should gather and annotate interview datasets in various languages and political systems (e.g., legislative debates in Europe or Asia).

9 Conclusion

We used hierarchical mapping and data explosion to offer a robust baseline strategy for the CLARITY shared task. Our tests demonstrate that a competitive Macro-F1 of 0.34 is attained by a refined RoBERTa-base model trained with class-weighted loss. These results emphasise how crucial it is to manage class imbalance and annotator disagreement in subjective NLP tasks.

10 Ethical Considerations

There are serious ethical concerns when using automated evasion detection systems in the political sphere. False positives, which classify a straightforward response as evasive, could be used as a weapon to discredit political rivals or erode public confidence in legitimate discourse. On the other hand, if clever evasion is not detected (False Negatives), dishonest actors may appear transparent.

Additionally, only English-language US presidential interviews are used to train our models. Due to cultural differences in communication techniques, these models would probably perform poorly if they were applied to different political contexts or languages without modification. Before being used in the actual world, future research must address these prejudices.

11 Broader Impact

The expanding discipline of computational journalism benefits from this work. We make it possible for political scientists to examine broad patterns in discourse by automating the identification of evasion, such as determining whether evasion rates are correlated with particular subjects (such as foreign policy versus the economics) or election cycles. Additionally, our "Flat-Mapping" method provides a framework that may be applied to various hierarchical categorisation tasks where low-level events are observable but high-level labels are subjective.

References

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

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

Elisa Ferracane, Greg Durrett, Junyi Jessy Li, and Katrin Erk. 2021. Did they answer? subjective acts and intents in conversational discourse. In *Proceedings of NAACL-HLT*, pages 1626–1644.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019a. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, and 1 others. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of EMNLP*, pages 10671–10682.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of ACL*, pages 784–789.

Konstantinos Thomas, Giorgos Filandrianos, Maria Lymperaiou, Chrysoula Zerva, and Giorgos Stamou. 2024. "i never said that": A dataset, taxonomy and baselines on response clarity classification. *arXiv preprint arXiv:2409.13879*.

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