HU at SemEval-2026 Task 6: A Hybrid Discriminative Modeling of Political Clarity and Evasion

Taha Munawar[†], **Basil Ali Khan**[†], **Arsal Jangda**[†],
Sarfaraz Baig[†], **Sandesh Kumar**[†], and **Abdul Samad**[†]
[†]Habib University, Dhanani School of Science & Engineering, Pakistan
{tm08122, bk08221, aj08514, sb08112}@st.habib.edu.pk
{sandesh.kumar, abdul.samad}@sse.habib.edu.pk

Abstract

We describe our submission to SemEval-2026 Task 6: CLARITY, which aims to classify political question–answer pairs by response clarity and evasive technique. We investigate several approaches, including long-context transformers, multiple instance learning, hierarchical multi-task models, and a natural language inference (NLI) formulation. On the development set, our best-performing NLI model achieves a macro-F1 of 0.79 for Subtask 1, while our best attention-based MIL model achieves a macro-F1 of 0.43 for Subtask 2. On the hidden evaluation set, our official submission obtains macro-F1 scores of 0.81 for Subtask 1 and 0.45 for Subtask 2. Our findings demonstrate the benefits of entailment-based modeling for clarity prediction and localized reasoning for evasion detection under limited computational resources.

1 Introduction

In high-stakes public discourse, particularly political interviews, the clarity of a response can be as important as its content. Speakers often employ equivocation or evasion, responding without directly answering, to deflect scrutiny or manage public perception. While such behavior has been studied extensively in political science and linguistics, its automatic detection remains a challenging Natural Language Processing (NLP) problem.

This challenge was formalized by Thomas et al. (2024) in "I Never Said That": A Dataset, Taxonomy and Baselines on Response Clarity Classification, which asks: **How can we automatically evaluate the clarity of a response with respect to its corresponding question?** To address this, the authors introduced a two-level hierarchical taxonomy (Fig. 1). The first level, Clarity, captures the overall directness of a response, while the second level, Evasion, identifies the specific linguistic strategy employed.

Our work addresses SemEval-2026 Task 6: CLARITY (Thomas et al., 2026), tackling both subtasks by predicting labels across this hierarchical taxonomy.

To support reproducibility, our code and experiment scripts are publicly available at https://github.com/tahamunawar/semeval-task6. The repository includes model-specific training configurations, including learning rates, batch sizes, maximum sequence lengths, warmup settings, epoch counts, loss functions, and data-processing scripts.

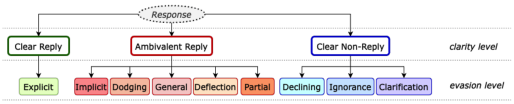

Figure 1: Two-level hierarchical taxonomy for evaluating response clarity, adapted from Thomas et al. (2024).

2 Literature Review

Research on ambiguity detection in question answering (QA) provides useful foundations for modeling political evasion. Shi et al. (2025) show that large language models often generate coherent yet multi-faceted responses, a pattern similar to evasive political communication. Their self-consistency prompting framework combined with a lightweight Random Forest classifier outperforms larger transformer baselines, particularly in low-resource settings. Similarly, Liu et al. (2023) introduce the AMBIENT benchmark, a linguistically annotated dataset covering lexical, syntactic, and pragmatic ambiguity, including political claims. They demonstrate that even advanced models struggle to resolve ambiguity at human-level performance, while multi-label NLI models perform more effectively, highlighting the need for domain-specific approaches. More broadly, Chatsiou and Mikhaylov (2020) survey deep learning

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 235–241

methods for political text analysis, emphasizing the role of contextual transformers and transfer learning in capturing nuanced political language.

Complementary work from linguistics and political science provides theoretical grounding for evasion taxonomies such as the SemEval-2026 framework. Gabrielsen et al. (2020) define "shifting" as an evasive strategy in which speakers appear to address a question while subtly refocusing it, corresponding closely to the "Deflection" category. They identify systematic patterns such as temporal or agent shifts, abstraction changes, and non-answers. These pragmatic patterns align with categories such as Dodging, Deflection, General, and Partial replies. Additional linguistic analyses provide fine-grained examples that map to these categories, including "non-type-conforming answers" and divergence-minimization strategies (Vázquez Carranza, 2018).

Unlike traditional transformers limited to 512 tokens, ModernBERT supports sequences up to 8192 tokens (Warner et al., 2025). This addresses the risk of omitting critical evasive content due to truncation, enabling full-context modeling of lengthy political responses.

Finally, Alvarez and Morrier (2026) introduce a self-supervised approach for assessing answer quality in political question-answer sessions by training language models to retrieve correct answers from candidate sets using only question text. Using 58,343 parliamentary exchanges from the Canadian House of Commons, they show that meaningful discourse representations can be learned without additional human annotation, suggesting that large-scale political QA corpora may provide useful domain signals for downstream question-response modeling tasks.

3 Dataset

Sourced from the White House (2006–2023), the dataset includes 287 presidential interview transcripts decomposed via ChatGPT-3.5 Turbo into singular QA pairs (sQAs). The final processed data comprises 3,450 training and 308 test rows, both characterized by significant class imbalance in Clarity and Evasion labels.

Due to high inter-annotator variance, Subtask 2 (Evasion) uses three ground-truth labels per test instance. Per task guidelines, a prediction is correct if it matches any of the three.

4 Methodology

We propose six architectural approaches for this task and explore data augmentation strategies to mitigate class imbalance.

4.1 Task Formulation and Evaluation

Given a question Q and an answer A, the objective is to predict the high-level clarity label for Subtask 1 and the fine-grained evasion label for Subtask 2. We use macro-F1 as the primary evaluation metric.

For Subtask 1, each instance has a single gold clarity label. For Subtask 2, each instance has three annotator-provided evasion labels, which are treated as a set of acceptable gold labels G_i. A prediction $\hat{y}_i$ is considered correct for class y if $\hat{y}_i = y$ and $y \in G_i$. Following the official scorer for the shared task (Thomas et al., 2026), per-class precision, recall, and F1 are computed using this set-valued gold criterion, and the final score is the unweighted mean across evasion classes:

$$F1_{\text{macro}} = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} F1_y.$$

To mitigate class imbalance, we experimented with Focal Loss, Weighted Cross-Entropy, and Dice Loss across architectural variants.

4.2 Direct Classification (ModernBERT)

To preserve the full context of sQA pairs, we fine-tune `ModernBERT-base` and `large`, leveraging their native 8192-token sequence length to capture the complete interaction within a single classification head (Warner et al., 2025), without having to truncate long political responses.

4.3 Multiple Instance Learning (MIL)

To address the input length constraint without relying solely on newer architectures, we implemented a Multiple Instance Learning (MIL) framework with Attention Pooling. This approach is particularly relevant for Subtask 2, where the specific "evasion technique" (e.g., a deflection) may occur in a small segment of a long answer.

We split the long answers into overlapping segments, or instances, such that for an input pair (Q, A), the answer A forms a bag of instances $\mathcal{I} = \{I_1, I_2, \ldots, I_k\}$.

- **Input Formatting:** Each instance I_i was constructed by concatenating the full question Q and an answer segment A_{segment}

using the standard BERT format: I_i = [CLS] Q [SEP] A_{segment} [SEP].

- **Chunking Parameters:** Maximum length 256 tokens with 160-token stride, limiting instances to $k \leq 16$.
- **Attention Pooling:** Each instance is encoded independently (frozen encoder) to obtain $\mathbf{h}_i$ from the final [CLS] state, followed by learnable attention weights α_i to produce $\mathbf{c} = \sum_{i=1}^{k} \alpha_i \mathbf{h}_i$ for classification.

We experimented with multiple encoders, including BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021), and ALBERT (Lan et al., 2020).

4.4 Natural Language Inference (NLI)

For Subtask 1 (Clarity Prediction), we hypothesized that the concept of "clarity" is fundamentally related to logical entailment. We effectively reframed the classification task into a three-way Natural Language Inference (NLI) problem. For every Question-Answer pair (Q, A), we constructed an NLI input consisting of a Premise and a Hypothesis:

- **Premise (P):** The politician's full answer $\{A\}$.
- **Hypothesis (H):** "The speaker explicitly answers the Question: $\{Q\}$."

We used the specialized NLI model `MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli`, pretrained on multiple NLI and factual consistency tasks. NLI has been widely used for modeling entailment relationships between premise–hypothesis pairs (Bowman et al., 2015). While Thomas et al. (2024) noted that traditional encoder-only classifiers are often constrained by 512-token inputs, we set the maximum sequence length to 2048 tokens for this model. This allowed substantially more of the answer context to be retained, although it remained bounded by available GPU memory. We treat this as an implementation choice rather than an architectural guarantee, since the model was not originally introduced as a long-context encoder.

We map NLI labels to clarity categories as follows: **Entailment** → **Clear Reply**, **Contradiction** → **Clear Non-Reply**, and **Neutral** → **Ambivalent Reply**.

This formulation also reduces the need to learn the semantics of each clarity label from scratch. Instead, the model can reuse entailment-oriented representations learned during NLI pre-training, which is particularly useful given the limited size and class imbalance of the CLARITY training set.

4.4.1 Adversarial Weight Perturbation (AWP)

We applied AWP (Wu et al., 2020) to improve robustness by encouraging flatter minima. During training, weights θ are perturbed along the gradient direction:

$$\theta_{adv} = \theta + \epsilon \frac{\nabla_\theta \mathcal{L}(\theta)}{\|\nabla_\theta \mathcal{L}(\theta)\|_2}$$

with $\epsilon = 10^{-3}$ starting from the second epoch.

4.4.2 Hybrid Probability Calibration and Blending

To reconcile biases between different calibration methods, we implemented a Weighted Probability Blend (Soft Voting). We utilized Isotonic Regression (P_{iso}) to act as a conservative estimator and Temperature Scaling (P_{ts}) to serve as a liberal estimator by softening probability distributions. The final probability distribution P_{final} is the weighted average of the two calibrated distributions:

$$P_{final} = w_{iso} P_{iso} + (1 - w_{iso}) P_{ts}$$

This NLI formulation was effective for Subtask 1 because its three outputs aligned directly with the clarity labels. We did not extend it to Subtask 2 in the final system because the nine evasion categories do not admit the same direct one-to-one mapping from entailment, contradiction, and neutral. A possible extension would require testing multiple label-specific hypotheses per instance and then designing an aggregation rule over the resulting NLI scores. We considered this design informally, but running it properly was computationally expensive and difficult to validate under our resource constraints.

4.5 Generative Approach

We explored the capabilities of generative LLMs by fine-tuning the `Llama-3.2-3B` instruct model (Meta AI, 2024). Due to computational constraints, we used 4-bit quantization with QLoRA (Dettmers et al., 2023). The inputs consisted of each question–answer pair provided directly to the model, and it was prompted to generate the classification label. For training stability and efficiency, we applied

LoRA (Hu et al., 2022) with an **r** value of **16** and an **alpha** of **32**.

We report this approach only for Subtask 1 in our main results because preliminary Subtask 2 experiments were less reliable than the discriminative MIL setup under the same compute constraints. We therefore did not use the generative model for the final Subtask 2 submission.

4.6 Multi-Task Hierarchical Gating (MTL)

We implemented a Multi-Task Learning (MTL) architecture using ModernBERT-large to jointly optimize Clarity and Evasion labels. The model utilizes a hierarchical gating mechanism where the Clarity Head first computes the probability distribution $P(c|h)$. This distribution is projected into a gating vector $g = \sigma(W_g P(c|h) + b_g)$, which modulates the shared representation via $h_{gated} = h \odot g$ to isolate features relevant to the predicted clarity level. To enforce taxonomic logic, we introduced a Consistency Loss, $\mathcal{L}_{cons} = \sum_i \text{softplus}(P(e_i) - P(c_{M(e_i)}))$, which penalizes child label probabilities that exceed those of their parent. The final objective is a weighted sum: $\mathcal{L}_{total} = \alpha\mathcal{L}_{evasion} + \beta\mathcal{L}_{clarity} + \gamma\mathcal{L}_{cons}$.

4.7 Hierarchy-Aware Global Model (HiAGM)

To capture taxonomic dependencies, we implemented a Hierarchy-Aware Global Model (HiAGM) (Zhou et al., 2020) that treats labels as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. We utilize a single-layer Graph Convolutional Network (GCN) to learn hierarchy-aware embeddings, $W_{labels} = \text{ReLU}(\tilde{A} \cdot E \cdot W_g)$, allowing feature propagation from parent clarity labels to child evasion techniques. Using ModernBERT-base as a text encoder (1024-token sequence), final logits are computed via dot-product compatibility: $\text{Logits} = h \cdot (W_{labels})^\top$. The model is optimized via Focal Loss ($\gamma = 3.0$) with class weights. To ensure 100% hierarchical consistency, Subtask 1 labels are deterministically inferred post-hoc from the predicted Subtask 2 leaf nodes.

4.8 Data Augmentation

We experimented with several augmentation strategies to improve robustness:

- **Random Word Deletion:** `nlpaug`-based deletion applied to answer text, generating augmented samples equal to 25% of the original data per clarity class.
- **Oversampling:** Weighted random sampling using inverse class frequency to mitigate imbalance without explicit duplication.
- **Contextual Word Replacement:** Masked language model–based substitutions using `nlpaug` (BERT).
- **Back-Translation:** English → German → English via Gemini 2.5 Flash, applied selectively to misclassified samples identified through 5-fold cross-validation.
- **Few-Shot Synthetic Generation:** Gemini 2.5 Flash prompted with minority-class examples to generate additional question–answer pairs.

These techniques were applied across both subtasks to address class imbalance.

5 Results and Discussion

5.1 Model Selection and Internal Benchmarking

Table 1 compares our best architectural approaches on the **development set**, against the baseline benchmarks established by the task organizers (Thomas et al., 2024).

Model	T1	T2
Baseline: Llama-2-70B	0.68	0.51
Gen: Llama-3.2-3B	0.57	-
Direct: ModernBERT-L	0.69	0.38
MIL: DeBERTa-v3-B	0.65	**0.43**
NLI: DeBERTa-v3-L	**0.79**	-
MTL: ModernBERT-L	0.64	0.40
HiAGM: ModernBERT-L	0.52	0.40

Table 1: Comparison with the organizers' development-set baseline. T1 = Subtask 1, T2 = Subtask 2.

5.1.1 Final Evaluation Results

After selecting the best-performing models based on development-set performance, we evaluated them on the hidden evaluation set. Our official SemEval-2026 Task 6 submission achieved a final leaderboard rank of **8** for Subtask 1 and **17** for Subtask 2 in the shared task (Thomas et al., 2026). The corresponding macro-F1 scores are shown in Table 2.

5.2 Comparison with Benchmark and Shared-Task Systems

A key finding of our work is the efficiency–performance trade-off relative to the organizers' Llama-2-70B baseline. While the baseline uses

Task	Model	Dev	Eval
Subtask 1	Llama-2-70B baseline	0.68	0.82
Subtask 1	NLI: DeBERTa-v3-L	**0.79**	0.81
Subtask 2	Llama-2-70B baseline	0.51	0.57
Subtask 2	DeBERTa-v3-B + MIL	0.43	0.45

Table 2: Development and hidden evaluation performance. Baseline scores are shown separately for the development and hidden evaluation splits to avoid cross-split comparison.

Class	Support	Prec.	Rec.	F1
Claims ignorance	14	0.83	0.71	0.77
Clarification	4	0.67	0.50	0.57
Declining to answer	13	0.29	0.54	0.38
Deflection	29	0.31	0.17	0.22
Dodging	72	0.54	0.29	0.38
Explicit	88	0.77	0.23	0.35
General	92	0.56	0.38	0.45
Implicit	82	0.41	0.57	0.48
Partial/half-answer	12	0.27	0.25	0.26

Table 3: Per-class development-set performance of the MIL model for Subtask 2. Support is computed as TP+FN under the official set-valued gold-label criterion.

a 70B-parameter generative model, our best Subtask 1 model uses DeBERTa-v3-large, with roughly 400M parameters. On the development set, our NLI formulation improves Subtask 1 macro-F1 from 0.68 to 0.79; on the hidden evaluation set, it remains competitive with the baseline (0.81 vs. 0.82), though not superior.

The broader shared-task results further contextualize this trade-off. The strongest Subtask 1 systems relied mainly on LLM prompting, multi-stage reasoning, and hierarchical mappings, with the top system reaching 0.89 macro-F1 (Thomas et al., 2026). Our result suggests that a smaller NLI-based encoder can remain competitive when clarity prediction is framed as entailment. For Subtask 2, however, the baseline and top shared-task systems remain substantially stronger, indicating that fine-grained evasion detection benefits from taxonomy-aware routing, constrained label spaces, and LLM-based verification beyond segment-level MIL.

5.3 Performance Analysis

5.3.1 Context vs. Segmentation in Evasion

For Subtask 2, the Multiple Instance Learning (MIL) approach outperformed ModernBERT (0.43 vs. 0.38). While ModernBERT captures the entire context (8192 tokens), the "evasion" label is often localized. The global pooling of ModernBERT may dilute these specific signals. In contrast, the Attention Pooling mechanism in MIL explicitly assigns high weights to the most discriminative segments of the answer, effectively isolating the evasion tactic.

The MIL model performs best on relatively explicit refusal-like categories such as *Claims ignorance* and *Clarification*, while performance is weakest for more pragmatically subtle categories such as *Deflection*, *Dodging*, and *Partial/half-answer*. This broadly aligns with the evasion-level annotation difficulty reported by Thomas et al. (2024), where annotators showed high disagreement among closely related categories such as *General*, *Implicit*, *Deflection*, and *Dodging*. In particular, Thomas et al. (2024) report low pairwise agreement for *General* vs. *Implicit* ($\kappa = 0.43$), *General* vs. *Deflection* ($\kappa = 0.56$), *General* vs. *Dodging* ($\kappa = 0.57$), and *Dodging* vs. *Deflection* ($\kappa = 0.62$). Since we do not compute a model-level confusion matrix, we interpret this only as supporting evidence that several low-F1 categories correspond to inherently difficult evasion distinctions.

5.3.2 The Trade-off of Hierarchical Constraints

While the Multi-Task Hierarchical Gating (MTL) and Hierarchy-Aware Global Model (HiAGM) were designed to explicitly model the taxonomic dependencies described in Sections 4.6 and 4.7, they achieved lower macro-F1 scores (0.64 and 0.52 respectively) on Subtask 1 compared to the NLI approach. We observe that the Hierarchical Consistency Loss ($\mathcal{L}_{cons}$) and the post-hoc deterministic inference used in HiAGM may have been overly restrictive. By forcing the model to strictly adhere to the parent-child relationship, we likely introduced a bottleneck where errors in fine-grained evasion detection (Subtask 2) propagated upward, penalizing the clarity prediction (Subtask 1). In contrast, the NLI model's flexibility to treat clarity as a logical entailment problem allowed for better generalization across varied presidential rhetorical styles. We note, however, that this explanation remains interpretive: we did not run a dedicated ablation removing only the consistency loss or post-hoc hierarchy constraint. Future work should isolate these components to determine whether the performance drop comes from hierarchy modeling itself or from the specific enforcement mechanism used here.

5.3.3 The Generative Lag

Our generative approach (`Llama-3.2-3B`) underperformed, achieving only 0.57 F1. We observed early overfitting, with performance dropping from 0.57 to 0.50 between epochs. This reinforces our finding that without the massive scale of the researchers' 70B model, generative decoding is less stable than discriminative classification for this task.

5.3.4 Inefficacy of Augmentation

The data augmentation strategies in Section 4.8 did not improve macro-F1 for either Subtask 1 or 2.

We attribute this to the *fragility of political nuance*. Political evasion depends on subtle rhetorical cues (e.g., "shifting" or "flouting Gricean Maxims"), making class boundaries highly sensitive to perturbations. This is consistent with Thomas et al. (2024), who reported low inter-annotator agreement ($\kappa = 0.48$) and difficulty distinguishing semantically proximate categories such as *General* and *Implicit*. Such inherent ambiguity likely caused both lexical augmentation and LLM-generated data to introduce noise rather than useful diversity, explaining the lack of performance gains.

We also experimented with intermediate domain pre-training on external political question–answer data (as described in Section 2). However, this did not yield measurable improvements, likely because the source domain (parliamentary proceedings) differs substantially from the target domain (presidential interviews), limiting transfer effectiveness given the small downstream dataset.

Task	Augmentation Strategy	Macro-F1
Subtask 1	*None (Best Model)*	**0.79**
	Random Word Deletion	0.72
	Weighted Oversampling	0.69
	Contextual Word Replacement	0.73
	Back-Translation	0.74
	Few-Shot Synthetic Generation	0.70
Subtask 2	*None (Best Model)*	**0.43**
	Random Word Deletion	0.41
	Weighted Oversampling	0.42
	Contextual Word Replacement	0.39
	Back-Translation	0.40
	Few-Shot Synthetic Generation	0.43

Table 4: Detailed impact of five data augmentation techniques on top-performing models for Subtask 1 and Subtask 2.

5.3.5 NLI Calibration and Error Analysis

Initial experiments revealed significant semantic overlap between the *Clear Reply* and *Ambivalent* classes, leading to a performance bottleneck. This aligns with Thomas et al. (2024), who reported that annotators struggled most to distinguish between the parent categories *Clear Reply* and *Ambivalent* in the clarity dimension, with inter-annotator agreement dropping to $\kappa = 0.65$, compared to much higher agreement for other high-level distinctions. The introduction of AWP proved vital, seeking a flatter loss landscape that improved the macro-F1 from 0.70 to 0.74 by enhancing model robustness against varied political phrasing. Building on this, our hybrid post-training probability calibration further decoupled the remaining semantic overlap, resulting in our final development-set macro-F1 score of 0.79. Our analysis showed that the distinct calibration methods yielded specific performance biases: Isotonic Regression (P_{iso}) provided high precision (0.84) for clear replies, while Temperature Scaling (P_{ts}) recovered recall (0.78). Through a grid search on the validation set, we determined the optimal temperature to be $T = 1.23$ and the optimal weighting factor to be $w_{iso} = 0.59$.

6 Limitations

A primary challenge in our experimentation was the significant computational disparity between our infrastructure and the task's 70B-parameter baseline. Our generative fine-tuning was restricted by Kaggle and Google Colab runtime limits, where the 8-hour-per-epoch training cycle frequently triggered platform timeouts and session disconnections. These compute constraints prevented us from performing exhaustive hyperparameter tuning or completing longer training runs, which likely contributed to the "generative lag" observed in our results.

A further limitation is that most results are based on single development-set runs rather than multi-seed averages. This is especially relevant for Subtask 2, where several evasion categories have low support and may therefore be sensitive to random initialization and sampling effects. In addition, although the final NLI system used a validation-selected calibration blend with $w_{iso} = 0.59$, we did not perform an extensive sensitivity analysis of this weight. Future work should evaluate whether the calibration gains remain stable across seeds and whether they improve all clarity classes uniformly or mainly benefit the difficult *Clear Reply*–*Ambivalent Reply* boundary.

References

R. Michael Alvarez and Jacob Morrier. 2026. Measuring the quality of answers in political q&as with large language models. *Political Analysis*, 34(1):78–95.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Katerina Chatsiou and Slava Jankin Mikhaylov. 2020. Deep learning for political science. In *The SAGE Handbook of Research Methods in Political Science and International Relations*. SAGE Publications, London, UK.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jonas Gabrielsen, Heidi Jønch-Clausen, and Christina Pontoppidan. 2020. Answering without answering: Shifting as an evasive rhetorical strategy. *Journalism*, 21(9):1355–1370.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. ICLR 2020 OpenReview submission.

Meta AI. 2024. Llama 3.2 3B Instruct Model Card. `https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/`.

Zhengyan Shi, Giuseppe Castellucci, Simone Filice, Saar Kuzi, Elad Kravi, Eugene Agichtein, Oleg Rokhlenko, and Shervin Malmasi. 2025. Ambiguity detection and uncertainty calibration for question answering with large language models. In *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 41–55, Albuquerque, New Mexico. Association for Computational Linguistics.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5204–5233, Miami, Florida, USA. Association for Computational Linguistics.

Konstantinos Thomas, Giorgos Filandrianos, Maria Lymperaiou, Chrysoula Zerva, and Giorgos Stamou. 2026. SemEval-2026 Task 6: CLARITY – unmasking political question evasions. *Preprint*, arXiv:2603.14027.

Ariel Vázquez Carranza. 2018. Evading and resisting answering: An analysis of mexican spanish news interviews. In Maj-Britt Mosegaard Hansen and Rosina Márquez Reiter, editors, *The Pragmatics of Sensitive Activities in Institutional Discourse*, pages 65–89. John Benjamins Publishing Company.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. 2020. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems*.

Jie Zhou, Chunping Ma, Dingkun Long, Guangwei Xu, Ning Ding, Haoyu Zhang, Pengjun Xie, and Gongshen Liu. 2020. Hierarchy-aware global model for hierarchical text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1106–1117, Online. Association for Computational Linguistics.