Steering AI for Classifying Open-Ended Text: An Evaluation of ANES 2020 Candidate Likes and Dislikes

George Quinn[†] and Richard Lau[*]

[†]PhD Candidate. Rutgers University. Department of Political Science.
george.quinn@rutgers.edu

[*]Distingushed Professor of Political Science. Rutgers University.
rick.lau@polisci.rutgers.edu

Abstract

Open-ended responses are a central but underused source of survey data because they preserve respondents' own language while requiring costly and time-intensive coding. This study evaluates a human-in-the-loop, API-based large language model workflow for classifying 16,416 open-ended responses from the 2020 American National Election Study candidate "likes" and "dislikes" questions. Using an adapted ANES codebook with 35 topic codes nested within nine major categories, we first establish a human-coder benchmark, then train the model through iterative batches of hand-coded examples, and finally validate the completed AI-coded database against a new set of 800 manually coded responses not used during training. The results show that the workflow performs best at the broader nine-category level, where micro-F1 ranges from .703 to .798 and mean respondent-level kappa ranges from .601 to .752. At the more granular 35-topic level, the model showed, a micro-F1 ranging from .564 to .700 and mean respondent-level kappa ranging from .488 to .699. Across conditions, the AI tends to assign more codes per response than the validation coder, producing higher recall than precision. These findings suggest that human informed API-based classification produces a scalable and substantively useful open-ended survey dataset, especially for broad thematic analysis, while fine-grained topic-level claims require more caution. The paper contributes a validated workflow for AI-assisted survey coding and releases a standardized 2020 ANES candidate likes/dislikes dataset for future research.

1 Introduction

The 2020 U.S. presidential election produced one of the richest collections of open-ended survey responses in the history of the American National Election Studies (ANES). Across four "likes" and "dislikes" questions about the two major-party candidates, respondents contributed more than 16,000 valid, text-based answers ranging from single-word descriptors to detailed policy critiques. Capturing the meaning in this volume of qualitative data is critical for understanding voter attitudes in a year shaped by the COVID-19 pandemic, nationwide racial justice protests, economic disruption, and record-high political polarization. Yet traditional human coding of such a dataset is both time-intensive and resource-demanding, limiting the speed and scalability of analysis.

Open-ended responses have been central to political science research for decades, offering unique insight into how citizens articulate preferences, priorities, and perceptions. Landmark works such as *The American Voter* (Campbell et al., 1960) and later studies by Kelley and Mirer (1974), Lau (1986), and Ridout and Franz (2011) have demonstrated the analytical value of ANES "likes/dislikes" data in measuring ideological sophistication, mapping policy priorities, and evaluating the role of campaign effects. More recent computational approaches, such as Structural Topic Modeling (STM) (Tingley et al, 2014), have automated parts of this process, but they are ill-suited for applying a fixed, theory-driven codebook like the ANES 35-topic scheme. STM discovers patterns in text but cannot consistently enforce predefined categories, making it difficult to compare results to historical ANES datasets.

Generative Artificial Intelligence offers a promising alternative. Unlike earlier statistical text-mining methods, large language models can be guided to apply a fixed codebook while incorporating contextual cues from the text. This capacity opens the possibility of achieving human-level coding reliability at scale, enabling both historical comparability and timely analysis of emergent political themes.

This study evaluates the use of GPT-5.5 to classify the 2020 ANES open-ended responses using an adapted 35-topic, nine-category codebook. Our research design proceeds in three distinct evaluation stages. First, we judge human–human agreement — measuring inter-coder reliability between two expert human coders on a shared training set to establish a benchmark for consistency. Then we test AI–human agreement (training phase), comparing the AI model's pre-feedback classifications to the agreed human codes across iterative batches to track learning over time. Lastly, we conduct post-training validation — after the AI coded the full dataset, comparing its classifications on a new sample to fresh human coding to test reliability on unseen cases. By first establishing reliability through these stages, we ensure that subsequent thematic analysis rests on a solid coding foundation. The results

provide both a methodological assessment of generative AI for political text classification and a substantive analysis of voter evaluations from the largest ANES respondent pool ever, collected during one of the most consequential U.S. elections in modern history.

Our results show that the AI model achieved substantial agreement with human coders (mean respondent-level Cohen's Kappa ranging from 0.60 to 0.75 across the nine major categories in the held-out validation data), in several cases exceeding inter-human agreement. Thematically, we find that personal qualities, leadership attributes, and comparative evaluations ("better than the alternative") dominate 2020 voter discourse, while explicit policy references occur less frequently. By restoring large-scale, standardized ANES coding for the first time since 2008, this project provides both a methodological demonstration and a substantive dataset for future research. All code, data, and training transcripts are publicly available to facilitate replication and further innovation in open-ended survey analysis.

2 Historical and Methodological Context: ANES Open-Ended Responses and the Evolution of Text Classification

Open-ended survey questions have long been a distinctive and valuable tool in political science. Unlike closed-ended items, which confine respondents to pre-specified options, open-ended questions invite respondents to describe their views in their own words. This allows researchers to capture subtleties in political reasoning, emotional tone, and issue salience that fixed-response formats may obscure. The American National Election Studies (ANES) has included such open-ended questions since 1948, most notably the "likes and dislikes" series about presidential candidates and parties. These responses have been central to many influential studies. Research drawing on ANES open-ended "likes and dislikes" responses has produced influential contributions to the study of political behavior. *The American Voter* (Campbell et al., 1960) used these data to develop the "levels of conceptualization" framework, a foundational measure of ideological sophistication in the electorate. Kelley and Mirer (1974) demonstrated that open-ended responses could be linked directly to vote choice, while Lau (1982) examined how negativity shapes political perceptions, and how political schemata influenced candidate choice in the 1956 and 1976 U.S. presidential elections (1986, 1989). More recently, Ridout and Franz (2011) used ANES open-ended data to assess how campaign advertising influences voter attitudes, illustrating the continued value of these responses for understanding electoral dynamics. Repass (2020) used them to provide a comprehensive overview of American presidential elections between 1960 and 2016. In the

space of methodologically examining generative AI approaches recent studies have looked to ANES open-ended responses to test various theories (Hobbs and Green, 2025).

From 1984 to 2004, ANES manually coded these responses into nine major thematic categories, adapting the code list to each election cycle (Repass, 2018). In 2008, the project undertook its largest-ever coding operation, classifying responses into 35 specific topics nested within the nine categories (Krosnick et al., 2015). Since then, however, ANES has released raw open-ended responses without standardized codes, creating a significant gap for longitudinal and comparative research. Without these codes, scholars face two unattractive options: (1) undertake expensive, time-consuming manual coding of thousands of responses, or (2) use automated text analysis methods that may not align with the ANES coding framework. This study addresses that gap by restoring comprehensive, standardized coding for the 2020 ANES candidate like/dislike responses.

2.1 Manual Coding

For decades, manual coding by trained human annotators was the standard for classifying open-ended survey responses. This approach yields high interpretive accuracy but is labor-intensive, slow, and subject to variation across coders, especially when fatigue or ambiguity is involved. The scale of the 2020 ANES dataset including over 16,000 valid responses makes manual coding prohibitively costly for most research teams (Krosnick, Lupia, & Berent, 2015).

2.2 Structured Topic Modeling (STM)

Structured Topic Modeling (STM) represented a major methodological advance by automating the discovery of themes in large text datasets (Tingley et al., 2014; Roberts et al., 2016). STM is probabilistic and unsupervised, grouping words into topics based on co-occurrence patterns, and it can incorporate covariates (e.g., party identification) to produce politically nuanced topics. STM has been applied to ANES data, uncovering important thematic structures without the need for manual annotation. However, STM has important limitations when the objective is to apply a fixed, theory-driven codebook such as the ANES 35-topic scheme. It generates probabilistic topic assignments rather than single, definitive codes, and the topics it discovers may not correspond to the predefined categories used in earlier ANES coding efforts. As a result, STM's outputs are not directly comparable to historical ANES datasets, limiting its utility for longitudinal research and replication. Replicability with these models indeed becomes a limitation, given that the process of topic sorting is unique to one session.

2.3 AI and NLP Applications

Other supervised and semi-supervised approaches, including keyword-based classifiers (Eshima, et al 2024), have been explored in survey research, but these also face trade-offs in accuracy, interpretability, and alignment with existing frameworks. More recently, large language models (LLMs) have been used for survey text classification. Pham et al. (2023) introduced "TopicGPT," a prompt-based approach for topic modeling using LLMs. Mellon et al. (2024) found that generative AI could match or exceed human coders in identifying salient topics. Argyle et al. (2023) applied GPT-3 and GPT-4 to subsets of ANES data to provide synthetic samples. While not utilizing open-ended text, this research provides a strong basis of AI usage for understanding political behavior. Most recently Hobbs and Green (2025) utilized ANES open-ended responses to validate their inferred attitudes model. Although not fully relying on generative AI, its utilization complimented proposed their "implied word method."

However, none of these studies produced a full, ANES codebook–aligned classification of the complete likes/dislike's dataset for a given election. Using a prompt-based, few-shot supervised training approach (Brown et al., 2020), GPT-5.5 can be instructed to apply the precise definitions in the ANES 35-topic codebook (adapted for the 2020 election) to large datasets. This produces deterministic, replicable outputs that directly align with historical ANES coding schemes—something STM and most older NLP approaches cannot do. The ANES likes/dislikes series is a uniquely rich source of information on voter sentiment, but the absence of standardized coding since 2008 has limited its utility. Advances in generative AI now make it possible to restore this coding at scale, aligning with historical frameworks while exceeding manual methods in efficiency and consistency. This study uses GPT-5.5, trained on expert-coded examples, to produce a full, replicable classification of the 2020 ANES candidate likes and dislikes open-ended responses, evaluate its performance against human coders, and make the resulting dataset available to the research community. Overall, new developments in automated text classification have improved upon the pioneering work of manual open-ended classification. Recent developments in Generative AI have allowed even more coherent and scalable methods for mass text classification projects. The work presented in this paper builds on the developments from large-scale text classification methods and leverages generative AI to tackle the issue of coding ANES open-ended responses.

3 Methods

The data for this study come from the 2020 American National Election Study (ANES) Time Series survey, which interviewed 8,200 respondents before Election Day. Among the survey items were four open-ended questions asking respondents what they liked and disliked about each of the two major-party presidential candidates.[1] Respondents could mention multiple themes in a single answer, and all valid responses were retained. Across the four questions, this yielded 16,416 individual open-ended responses (more than the number of respondents because each person could answer up to four questions).

3.1 Codebook Adaptation

We based our classification scheme on the 35-topic, nine-category codebook developed for the 2008 ANES open-ended likes/dislikes project (Krosnick et al., 2015). The 2008 codebook was selected for its comprehensive coverage of candidate attributes, policy issues, and general evaluative categories, as well as its historical role in ANES coding. To make it applicable to the 2020 election cycle, we introduced several new topics such as "Public Health / COVID" and "Better than the Alternative" clearly reflecting contemporary themes in the political environment. The final 2020 codebook retained the full 35-topic structure, with each topic nested under one of nine broader categories, enabling both fine-grained and aggregated analyses. Table 1 are the available 35 topic codes bound together by the 9 major categories.

3.2 Expert Human Coding for Training Data

A high-quality training data set was created through expert human coding. Three political science researchers with prior ANES coding experience were selected as primary annotators. All underwent a refresher training on the adapted 2020 codebook to ensure consistency in applying topical definitions. For each of the four candidate like/dislike questions, 200 responses were randomly sampled (N = 800 total).

To minimize coder fatigue and reduce bias, two coders assigned two of the four questions, with a supervising senior coder reviewing all four responses. Each coder independently assigned one or more topic codes to every response in their set, following the definitions in the adapted codebook. To create the agreed dataset, the coders met to reconcile disagreements, reviewing the relevant definitions until they reached consensus. A third reviewer examined the reconciled set to verify consistency between question types. The resulting set of 800

[1] More precisely, "Is there anything in particular about Joe Biden (Donald Trump) that might make you want to vote for him?" and "Is there a anything in particular about Joe Biden (Donald Trump) that might make you want to vote against him?"

Table 1: Mapping of ANES Major Categories to 2020 Valid Codes

ANES Major Categories	2020 Valid Codes
1: Experience	• Electability • Political Experience • Non-Political Experience • Running Mate / Endorsement
2: Leadership Qualities	• Better than the alternative • Personality – Ability • Personality – Leadership
3: Personal Qualities	• Personality – Honesty • Personality – Intelligence • Personality – Other • Candidate Health / Mental Fitness • Religion • Education • Physical Appearance • Demographics
4: Government Management	• Public Health / Covid • Democracy / Voting Rights • Judiciary
5: Government Activity / Philosophy	• Party • Ideology / Philosophy
6: Domestic Policies	• Policy – Economic • Policy – Poor People • Policy – Liberty • Policy – Other
7: Foreign Policies	• Policy – International Affairs
8: Group Connections	• Groups • Racial Issues
9: Miscellaneous	• General • Scandal / Cover-up • Other Past Activity • Emotions / Feelings • Candidate – Other • Non-Candidate • Don't Know • All Others

responses with agreed human-assigned codes served as the "gold standard" training data set for the AI model.

3.3 AI Model Training Procedure

We trained a generative AI model, GPT-5.5, (OpenAI, 2026), to apply the adapted 2020 ANES codebook using a few-shot supervised approach. The model was provided with the complete 35-topic code list and definitions, explicit instructions allowing multiple codes per response, and sample coded responses drawn from the training set. Training was conducted interactively in batches of 10 responses per question type. For each batch, the model classified the responses and returned its assigned codes. These outputs were compared against the gold-standard human codes, and any discrepancies were noted. For each mismatch, the model was supplied with the correct human-assigned code(s) along with a brief explanation referencing the relevant codebook definition.

This feedback loop was repeated for 20 batches per question type, covering all 200 training examples for each question. The iterative process allowed the model to refine its application of category definitions after each round, progressively aligning its classifications with the human coders' decisions ensuring longitudinal comparability. After this extensive training procedure, we maintained a curated learning log for each question type. Allowing us to move forward with classification of the full universe of ANES Like/Dislike responses in 2020. Figure 1 below captures our traning, classification, and independent validation workflow.

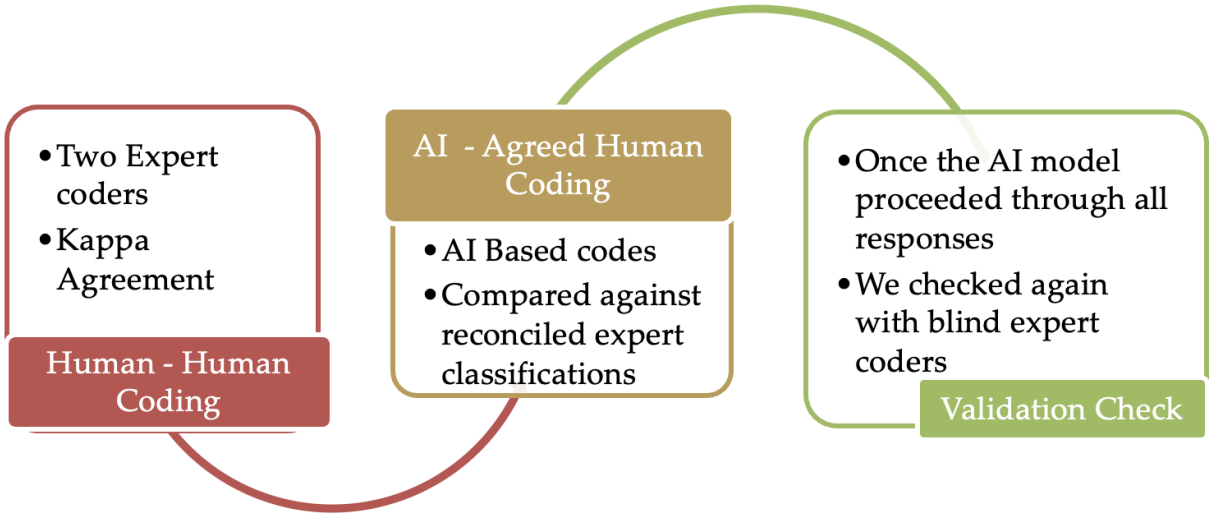

Figure 1: Flow Chart– Three Stages of Results

3.4 Validation and Full Classification

To assess model performance on unseen data, we created a new validation set of 800 responses (200 per question type), randomly selected from responses not included in the original training data. This set was manually coded by one of the expert coders, who was blind to the AI's classifications. The validation set thus provided an independent test of the model's generalization beyond the training data. We evaluated agreement between the AI's classifications and the human codes at both the detailed (35-topic) and aggregated (nine-category) levels. Inter-coder reliability was quantified using Cohen's Kappa (k), which corrects for chance agreement.

Values above 0.60 were interpreted as substantial agreement, following the benchmarks of Landis and Koch (1977). After the training procedure, we binded a learning log from all of the coding of these 200 responses to a GPT-5.5 model which then classify all uncoded responses from the 2020 ANES. Each response was assigned one or more of the 35 topic codes according to the adapted codebook. The final dataset includes codes at both the 35-topic and nine-category levels, enabling comparisons with earlier ANES-coded datasets as well as new analyses of 2020-specific themes. To ensure transparency and reproducibility, we have made all materials publicly available, including: the adapted 2020 ANES codebook; the full 800-response training set with human and AI codes; the 800-response validation set; the complete coded dataset of 16,416 responses; the training dialogue between the researcher and GPT-5.5; and the R scripts used for reliability analysis and data processing. These resources allow future researchers to replicate the classification procedure and apply similar methods to other open-ended survey datasets.

4 Data

The dataset for this study comprises all open-ended responses to the 2020 American National Election Study (ANES) "likes" and "dislikes" questions about the two major-party presidential candidates. The survey, conducted prior to the general election, included 8,200 respondents, each asked four questions fairly early in during the pre-election survey[2]: what they liked about Joe Biden, what they disliked about Joe Biden, what they liked about Donald Trump, and what they disliked about Donald Trump.[3] Most people provided responses for two of these four questions, most often likes for their preferred candidate and dislikes for their opposed candidate. Some people answered one question (disproportionately dislike

[2]See pages 48-52 of the ANES 2020 Questionare

[3]Unlike all previous ANES surveys, because the covid pandemic made conducting face-to-face interviews impossible, all interviews were conducted online in 2020.

Trump) and a very small number of respondents answered all four questions. Table 2 below highlights the distribution of valid responses from the 2020 ANES candidate like dislike questions. As noted earlier, our validation and training databases contain 800 responses each, equally distrbuted with 200 to each like and dislike question respectivley.

Table 2: Valid Open-Ended Responses Gathered from the 2020 ANES

Question Asked	N	%
What do you like about Joe Biden?	3,857	23.50
What do you dislike about Joe Biden?	3,900	23.76
What do you like about Donald Trump?	3,650	22.23
What do you dislike about Donald Trump?	5,009	30.51
Total Responses	**16,416**	**100.00**

We begin by evaluating the reliability of the classification process before turning to the substantive patterns in the coded data. The first step compares agreement between human coders in the training dataset, establishing a benchmark for what can reasonably be achieved through manual annotation. We then assess the AI model's performance against human-coded validation data at both the 35-topic and nine-category levels, reporting Cohen's Kappa as a measure of inter-coder reliability. Once classification accuracy is established, we present descriptive findings on the thematic distribution of responses across the four question types, identifying both consistent patterns with prior ANES cycles and shifts unique to the 2020 election context.

4.1 Agreement Statistics for Open-Ended Coding

Because ANES candidate like/dislike responses can contain multiple distinct considerations, the coding task is best understood as a multi-label classification problem. A single response may mention, for example, a candidate's honesty, policy positions, age, and comparison to the opposing candidate. Conventional single-label reliability statistics are therefore insufficient on their own, because they assume that each response receives one mutually exclusive category. To evaluate agreement in a way that reflects the structure of the coding task, we report several complementary measures: respondent-level percent agreement, respondent-level Cohen's κ, exact-set agreement, Jaccard similarity, response-level precision, recall, and F1, and micro-averaged precision, recall, and F1.

Let $C = \{1, \ldots, K\}$ denote the set of valid codes, where $K = 35$ for the full topic-code scheme and $K = 9$ for the major-category scheme. For each respondent i, let $A_i \subseteq C$ denote the set of codes assigned by the AI model and let $H_i \subseteq C$ denote the set of codes assigned by the human coder or agreed human-coded standard. The comparison is order-insensitive: if one coder lists codes as $\{5, 13\}$ and another lists them as $\{13, 5\}$, the two classifications are treated as identical.

We report a whole battery of inter-coder reliability scoring[4]. Most relevent for survey researchers we capture respondent-level percent agreement and Cohen's κ. For each response, the set of assigned codes is converted into a binary vector of length K, where each element indicates whether a given code is present or absent. Let a_{ik} equal 1 if the AI assigns code k to response i and 0 otherwise. Let h_{ik} equal 1 if the human coder assigns code k to response i and 0 otherwise. Respondent-level percent agreement is:

$$P_{o,i} = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}(a_{ik} = h_{ik}).$$

[4]Please See Appendix A for relevant explanations of reliability scoring

Cohen's κ adjusts this observed agreement for chance agreement implied by the marginal distribution of assigned and unassigned codes. For respondent i, expected agreement is:

$$P_{e,i} = \bar{a}_i \bar{h}_i + (1 - \bar{a}_i)(1 - \bar{h}_i),$$

where

$$\bar{a}_i = \frac{1}{K} \sum_{k=1}^{K} a_{ik} \quad \text{and} \quad \bar{h}_i = \frac{1}{K} \sum_{k=1}^{K} h_{ik}.$$

Respondent-level Cohen's κ is then:

$$\kappa_i = \frac{P_{o,i} - P_{e,i}}{1 - P_{e,i}}, \quad \text{for } P_{e,i} < 1.$$

Given that this statistic is assigned to reach respondent. We report the mean and median respondent-level κ across validation responses.

4.2 Human-to-Human Training Data Results

Before evaluating the AI model, we first assessed the level of agreement achievable through expert human coding. This human-to-human comparison provides the baseline against which the AI-assisted classification should be judged. Because each open-ended response could receive more than one code, we evaluate agreement as a multi-label classification task rather than as a single-label assignment problem. Accordingly, Table 3 reports exact-set agreement, Jaccard similarity, response-level precision, recall, and F1, micro-averaged precision, recall, and F1, respondent-level percent agreement, and respondent-level Cohen's kappa.

The results show that human coders achieved moderate to substantial agreement when applying the full 35-topic codebook. Exact-set agreement ranged from 0.330 for Trump Dislikes to 0.420 for Biden Dislikes. This measure is intentionally strict: it requires both coders to assign the same complete set of topic codes to a response. More forgiving set-overlap measures indicate stronger consistency. Mean Jaccard similarity ranged from 0.511 for Trump Dislikes to 0.611 for Biden Dislikes. Mean response-level F1 ranged from 0.587 to 0.677, suggesting that even when coders did not produce identical code sets, they often identified overlapping substantive content.

Respondent-level Cohen's kappa ranged from 0.515 for Trump Dislikes to 0.657 for Biden Dislikes. These results demonstrate two important features of the coding task. First, the 35-topic coding scheme contains meaningful ambiguity even for trained human coders, especially when respondents provide brief, vague, or multi-dimensional answers. Second, human agreement is strongest for the two Biden conditions and weakest for Trump Dislikes,

suggesting that some response domains are more difficult to classify consistently than others. These benchmarks are important because they establish the level of reliability that can reasonably be expected from expert human coding before evaluating AI-generated classifications.

Table 3: **Human-to-Human Reliability for 35-Topic Open-Ended Coding.** Entries compare Coder 1 and Coder 2 across 200 responses per condition. Percent agreement and Cohen's κ are calculated at the respondent-code level across the 35 valid topic codes. Exact set agreement requires both coders to assign the same complete set of topic codes to a response. Jaccard similarity measures set overlap. Response-level F1 averages precision and recall across respondents, while micro-averaged precision, recall, and F1 pool all respondent-code decisions.

Condition	**% Agree.**	**Cohen's κ**	**N**	**Exact Set**	**Jaccard**	**Resp. F1**	**Micro P**	**Micro R**	**Micro F1**
Biden Likes	0.961	0.656	200	0.410	0.606	0.677	0.687	0.633	0.659
Biden Dislikes	0.966	0.657	200	0.420	0.611	0.675	0.675	0.711	0.692
Trump Likes	0.960	0.592	200	0.410	0.574	0.629	0.614	0.636	0.625
Trump Dislikes	0.940	0.515	200	0.330	0.511	0.587	0.514	0.551	0.532

Table 4: **Human-to-Human Reliability for 9-Major-Category Open-Ended Coding.** Entries compare Coder 1 and Coder 2 across 200 responses per condition after the 35 topic codes are collapsed into nine major categories. Percent agreement and Cohen's κ are calculated at the respondent-category level. Exact set agreement requires both coders to assign the same complete set of major categories to a response. Jaccard similarity measures set overlap, while response-level and micro-averaged F1 summarize agreement across category assignments.

Condition	**% Agree.**	**Cohen's κ**	**N**	**Exact Set**	**Jaccard**	**Resp. F1**	**Micro P**	**Micro R**	**Micro F1**
Biden Likes	0.907	0.718	200	0.510	0.714	0.777	0.789	0.742	0.765
Biden Dislikes	0.916	0.721	200	0.540	0.721	0.775	0.768	0.806	0.786
Trump Likes	0.885	0.604	200	0.465	0.638	0.691	0.670	0.722	0.695
Trump Dislikes	0.854	0.544	200	0.370	0.600	0.680	0.612	0.697	0.652

As indicated utilizing this manual method of open-ended classification provides substantial agreement in two of the four like/dislike questions, with the two Trump conditions falling into the moderate range. The validity and scalability comes into question particularly given the volume of open-ended responses. For example, the 2020 universe of open-ended responses totals more than 16,000 valid responses and this project would be very hard to complete manually at this level. After working with both coders, we determined an agreed set of standard codes for all training responses. This gold standard agreed set of open-ended codes will then be processed using the AI model.

4.3 AI Model Training Performance Batch by Batch

After constructing the hand-coded training data, we evaluated whether an API-based large language model could learn to apply the adapted ANES codebook in a manner consistent with the human-coded standard. For each of the four candidate evaluation conditions, the model classified 200 training responses in 20 batches of 10 responses. In each batch, the model first assigned topic codes without seeing the correct answers. The hand-coded classifications were then revealed, and the model was prompted to update its coding rules by identifying missed codes, extra codes, and boundary cases in relation to the codebook. The updated learning log was carried forward into the next batch.

Figure 2 reports mean respondent-level Cohen's kappa by batch for each of the four conditions. The figure should be interpreted as evidence of internal model calibration rather than as a final test of database validity. Across the training process, the model generally maintained moderate to strong respondent-level agreement with the hand-coded classifications. Biden Likes and Trump Likes show the clearest upward trends, suggesting that iterative feedback improved the model's ability to apply the codebook consistently across batches. Biden Dislikes began at a relatively high level of agreement and remained comparatively stable across the training process. Trump Dislikes was the most difficult condition, with more uneven batch-level performance and a weaker upward trend, reflecting the greater ambiguity and heterogeneity of negative evaluations of Trump.

These results indicate that the feedback loop helped regularize the model's classifications over time. However, they should not be interpreted as proof that the model would perform equally well on unseen responses. Because the model received batch-level corrective feedback during training, these results primarily show that the model could be calibrated to the human coding standard within the training data. The stronger test of the completed AI-coded database is the held-out validation exercise reported below, where manually coded validation responses are compared against the full AI-generated classifications.

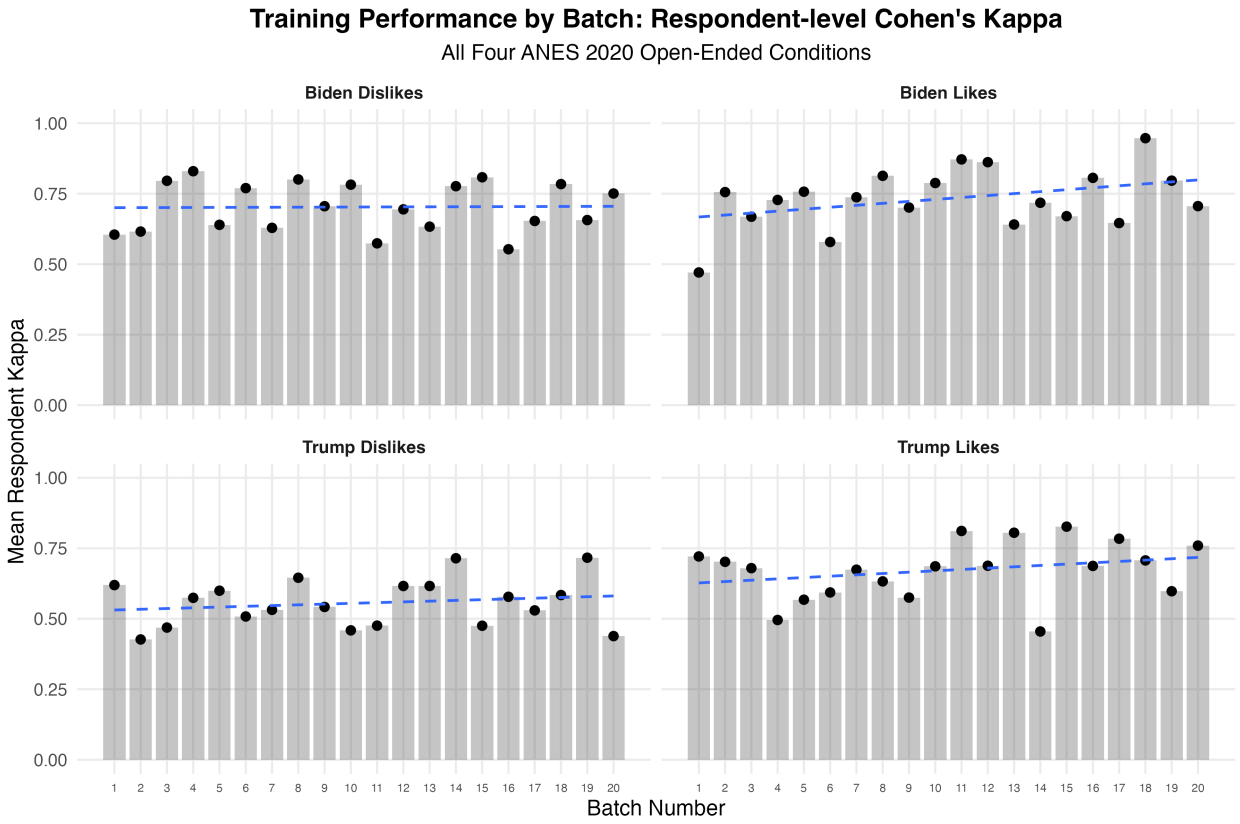

Figure 2: **Model Training Performance by Batch.** Points report mean respondent-level Cohen's kappa for each batch of 10 training responses. Dashed lines show linear trends across the 20 training batches for each candidate like/dislike condition.

In aggregate batch over batch, the generative classification outputs indicate that the model moves closer to mimicking human coherence. In aggregate, later batches obtain the highest levels of Cohen's Kappa (k) inter coder agreement. The iterative nature of this model suggests that overtime the model is learning and getting better at modeling the nuances within these open-ended responses. Furthermore, the trends from each batch's performance follow an improving direction for all conditions. See the Appendix B table for full outputs from the training data evaluation.

4.4 Overall AI Model Performance on the Training Data

After evaluating batch-by-batch learning, we next summarize the model's overall agreement with the agreed human-coded standard across the full training set. Tables 5 and 6 reports mean respondent-level Cohen's kappa for the AI classifications compared to the agreed hand codes. This statistic treats each respondent's answer as a set of topic codes and evaluates whether the AI recovered the same substantive codes assigned by human coders, regardless of the order in which codes were listed.

The results show that the API-based classification procedure achieved moderate to substantial agreement with the agreed human-coded classifications across the four candidate evaluation conditions. Agreement was strongest for Biden Likes, where the model achieved a mean respondent-level kappa of 0.733, followed by Biden Dislikes at 0.703 and Trump Likes at 0.672. Trump Dislikes was the most difficult condition, with a mean respondent-level kappa of 0.556. This pattern is consistent with the batch-level results: the model performed best on the Biden evaluations and weakest on negative evaluations of Trump, which often contain more heterogeneous and overlapping criticisms.

These results indicate that the iterative training procedure was able to calibrate the model to the human coding standard for most response conditions. Importantly, however, this table should be interpreted as evidence of training-set performance rather than external validity. Because the model received corrective feedback during the training process, the more demanding test is whether the completed AI-coded database performs well against manually coded validation responses that were not used during model calibration. That held-out validation exercise is reported in the following section.

Table 5: **Human-to-AI Reliability for 35-Topic Open-Ended Coding.** Entries compare the AI-generated classifications to the agreed human-coded benchmark across 200 responses per condition. Percent agreement and Cohen's κ are calculated at the respondent-code level across the 35 valid topic codes. Exact set agreement requires the AI and human benchmark to assign the same complete set of topic codes to a response. Jaccard similarity measures set overlap. Response-level F1 averages precision and recall across respondents, while micro-averaged precision, recall, and F1 pool all respondent-code decisions. **Bolded** entries indicate Human-to-AI scores that exceed the corresponding gold standard Human-to-Human reliability score.

Condition	% Agree.	Cohen's κ	N	Exact Set	Jaccard	Resp. F1	Micro P	Micro R	Micro F1
Biden Likes	**0.966**	**0.733**	200	**0.460**	**0.681**	**0.751**	0.678	**0.792**	**0.731**
Biden Dislikes	0.965	**0.703**	200	0.420	**0.651**	**0.721**	0.670	**0.771**	**0.717**
Trump Likes	0.960	**0.672**	200	0.385	**0.620**	**0.693**	**0.646**	**0.755**	**0.697**
Trump Dislikes	0.939	**0.556**	200	0.225	0.496	**0.588**	**0.568**	**0.680**	**0.619**

Table 6: **Human-to-AI Reliability for 9-Major-Category Open-Ended Coding.** Entries compare the AI-generated classifications to the agreed human-coded benchmark after the 35 topic codes are collapsed into nine major categories. Percent agreement and Cohen's κ are calculated at the respondent-category level. Exact set agreement requires the AI and human benchmark to assign the same complete set of major categories to a response. Jaccard similarity measures set overlap, while response-level and micro-averaged F1 summarize agreement across category assignments. **Bolded** entries indicate Human-to-AI scores that exceed the corresponding gold standard Human-to-Human reliability score.

Condition	% Agree.	Cohen's κ	N	Exact Set	Jaccard	Resp. F1	Micro P	Micro R	Micro F1
Biden Likes	**0.931**	**0.803**	200	**0.625**	**0.799**	**0.849**	**0.806**	**0.879**	**0.841**
Biden Dislikes	**0.927**	**0.783**	200	**0.595**	**0.778**	**0.830**	**0.793**	**0.861**	**0.826**
Trump Likes	**0.896**	**0.681**	200	0.460	**0.687**	**0.752**	**0.724**	**0.816**	**0.767**
Trump Dislikes	**0.893**	**0.688**	200	**0.425**	**0.695**	**0.763**	**0.757**	**0.830**	**0.792**

4.5 Validation Data

Following completion of the training process and full classification of the 2020 ANES open-ended responses, we conducted a final validation exercise to evaluate the quality of the AI-generated database on responses not used during iterative model calibration. For each of the four candidate like/dislike questions, 200 responses were selected from outside the training set, producing a validation sample of 800 responses. These responses had already been classified by the AI model during the full-scale coding process. A human coder, blind to the AI classifications, then independently assigned codes to each response using the adapted 2020 ANES codebook. The resulting human-coded validation data served as the reference standard for evaluating the completed AI-coded database.

Because each response may receive multiple codes, validation is evaluated as a multi-label classification task. We therefore report several complementary statistics rather than relying on a single reliability coefficient. Exact-set agreement measures whether the AI and validation coder assigned the identical complete set of codes to a response. Jaccard similarity measures partial overlap between the AI and human code sets. Micro-averaged precision, recall, and F1 pool all respondent-code decisions across the validation set. Respondent-level Cohen's κ compares the presence or absence of each valid code within a response, adjusting for chance agreement.

Table 7 reports the main validation results. At the 35-topic level, performance varies across question conditions. Biden Likes and Biden Dislikes show the strongest validation results, with micro-F1 scores of 0.692 and 0.700 and mean respondent-level κ values of 0.699 and 0.685, respectively. Trump Likes also shows moderate agreement, with a micro-F1 of 0.643 and mean respondent-level κ of 0.615. Trump Dislikes is the most difficult condition, with a lower exact-set agreement rate of 0.145, micro-F1 of 0.564, and mean respondent-level κ of 0.488. This pattern suggests that the AI model performs best on Biden-related evaluations and faces greater difficulty classifying negative evaluations of Trump.

As expected, agreement improves when the 35 topic codes are collapsed into nine major categories. At the major-category level, micro-F1 ranges from 0.703 for Trump Likes to 0.798 for both Biden Likes and Biden Dislikes. Mean respondent-level κ ranges from 0.601 to 0.752. These results indicate that the AI-generated database is especially reliable for broad thematic classification, even when exact fine-grained topic agreement is more uneven. Across all four conditions, the AI also assigns more codes per response than the validation coder. At the 35-topic level, the AI averages between 2.24 and 2.97 codes per response, compared with 1.79 to 2.17 codes in the validation data. This pattern helps explain why recall is consistently higher than precision: the AI frequently recovers human-coded topics but also adds additional plausible codes.

Table 7: **Validation of the Full AI-Coded Database Against Hand-Coded Validation Data**

Condition	**Level**	**% Agree.**	**Cohen's κ**	**Exact**	**Jaccard**	**P**	**R**	**F1**	**AI Codes**	**Val. Codes**
Biden Likes	35 Topics	**0.965**	**0.699**	**0.435**	**0.650**	0.623	**0.779**	**0.692**	2.24	1.79
	9 Major	**0.919**	**0.750**	**0.575**	**0.751**	0.744	**0.860**	**0.798**	1.94	1.68
Biden Dislikes	35 Topics	0.962	**0.685**	0.390	**0.631**	0.625	**0.796**	**0.700**	2.47	1.94
	9 Major	0.913	**0.752**	**0.555**	**0.757**	0.727	**0.885**	**0.798**	2.13	1.75
Trump Likes	35 Topics	0.957	**0.615**	0.335	0.564	0.580	**0.720**	**0.643**	2.31	1.86
	9 Major	0.872	0.601	0.405	0.620	0.646	**0.771**	**0.703**	2.11	1.77
Trump Dislikes	35 Topics	0.936	0.488	0.145	0.426	0.488	**0.668**	**0.564**	2.97	2.17
	9 Major	**0.871**	**0.620**	0.340	**0.633**	**0.669**	**0.821**	**0.737**	2.43	1.98

Note: Each condition includes 200 hand-coded validation responses matched to the full AI-coded dataset by respondent identifier. Percent agreement and Cohen's κ are calculated at the respondent-code level for the 35-topic scheme and at the respondent-category level for the 9-major-category scheme. Exact agreement requires identical code sets. Jaccard measures set overlap. P, R, and F1 report micro-averaged precision, recall, and F1 across respondent-code decisions. AI Codes and Val. Codes report the average number of codes assigned per response by the AI and validation coder, respectively. **Bolded** entries indicate validation scores that exceed the corresponding Human-to-Human gold-standard reliability score.

To evaluate where the model performs well or poorly by substantive domain, Table 8 reports F1 scores for each of the nine major categories. These category-level diagnostics show that the AI model performs especially well in several substantively central categories. Personal Qualities is consistently strong across conditions, with F1 scores of 0.886 for Biden Likes, 0.905 for Biden Dislikes, 0.718 for Trump Likes, and 0.879 for Trump Dislikes. Domestic Policies also shows strong performance, with F1 scores ranging from 0.784 to 0.892. Foreign Policies performs especially well for Trump Likes, where the F1 score is 0.905, but is weaker for Biden Likes, where the validation sample contains only four human-coded cases in that category.

Performance is more uneven in categories with small validation counts or broader conceptual boundaries. For example, Miscellaneous is consistently weaker, with F1 scores of 0.300 for Biden Likes, 0.700 for Biden Dislikes, 0.400 for Trump Likes, and 0.492 for Trump Dislikes. Experience is also weak for Trump Dislikes because the validation data include only four human-coded cases in that category. These patterns suggest that the AI-coded database is strongest for clear and substantively coherent categories, while more diffuse or low-prevalence categories require greater caution[5].

Table 8: **Validation F1 Scores by Major Category.** Entries report category-level F1 scores comparing the full AI-coded database to hand-coded validation data. Parentheses report the number of validation responses assigned to each category by the human validation coder.

Major Category	**Biden Likes**	**Biden Dislikes**	**Trump Likes**	**Trump Dislikes**
Experience	0.872 (46)	0.809 (47)	0.651 (28)	0.444 (4)
Leadership Qualities	0.844 (97)	0.646 (38)	0.644 (91)	0.695 (90)
Personal Qualities	0.886 (75)	0.905 (98)	0.718 (45)	0.879 (129)
Government Management	0.711 (23)	0.667 (6)	0.645 (17)	0.847 (29)
Government Activity / Philosophy	0.642 (27)	0.814 (54)	0.523 (19)	0.625 (7)
Domestic Policies	0.784 (40)	0.845 (50)	0.892 (64)	0.789 (37)
Foreign Policies	0.471 (4)	0.696 (8)	0.905 (41)	0.638 (16)
Group Connections	0.769 (15)	0.737 (14)	0.625 (26)	0.741 (48)
Miscellaneous	0.300 (8)	0.700 (34)	0.400 (22)	0.492 (36)

The validation results support a measured interpretation of the AI-generated database. AI classifications show significant agreement with independent human validation coding, especially at the level of nine-major categories. The database is therefore best suited for analyses of broad thematic patterns in open-ended candidate evaluations, while fine-grained

[5]Please see Appendix D for all ICR ratings by Major Category for Human to Human, AI to Human, and Validated Data

35-topic analyses should be interpreted with attention to the validation diagnostics reported here. With these reliability results established, we next turn to the substantive distribution of candidate likes and dislikes in the 2020 ANES data.

5 Like/Dislike Results for 2020 ANES Respondents

Having established the reliability of the AI-coded database, we next examine the substantive distribution of open-ended candidate likes and dislikes in the 2020 ANES. The analyses below use the full AI-classified dataset and summarize responses at two levels of aggregation: the nine major ANES-style categories and the full 35-topic codebook. Because responses could receive more than one code, the figures report the number of coded mentions rather than mutually exclusive respondent counts. Thus, a single respondent may contribute to multiple categories if their answer included more than one distinct reason.

Figure 3 presents the distribution of responses across the nine major categories for all four candidate evaluation questions. Several broad patterns emerge. First, evaluations of both Biden and Trump were heavily structured around candidate traits. Personal Qualities and Leadership Qualities are among the most frequently assigned categories across all four conditions, indicating that respondents often evaluated candidates in terms of honesty, competence, leadership, temperament, age, and other personal characteristics. This pattern is especially pronounced in the dislike conditions, where negative evaluations frequently centered on perceived personal deficiencies, trustworthiness, and leadership style.

Second, the two candidates' positive evaluations differ in their thematic emphasis. Biden Likes responses are especially concentrated in Leadership Qualities and Personal Qualities, suggesting that favorable evaluations of Biden were often framed around steadiness, decency, experience, and perceived suitability for office. Trump Likes responses, by contrast, show a stronger concentration in Domestic Policies, along with substantial mentions of Leadership Qualities, Personal Qualities, Experience, Foreign Policies, and Group Connections. This indicates that favorable evaluations of Trump were more likely to combine policy-based approval, perceived effectiveness, outsider or business experience, and group or ideological alignment.

Third, the dislike conditions reveal different structures of opposition. Biden Dislikes responses are strongly concentrated in Personal Qualities, followed by Government Activity / Philosophy, Domestic Policies, Leadership Qualities, and Experience. This suggests that criticism of Biden combined concerns about personal capacity, age or fitness, ideological direction, and policy disagreement. Trump Dislikes responses are even more concentrated in Personal Qualities, but also show substantial counts in Miscellaneous, Leadership Qualities,

and Group Connections. This indicates that negative evaluations of Trump were often expressed through broad affective or character-based criticisms, concerns about leadership, and references to groups or social conflict.

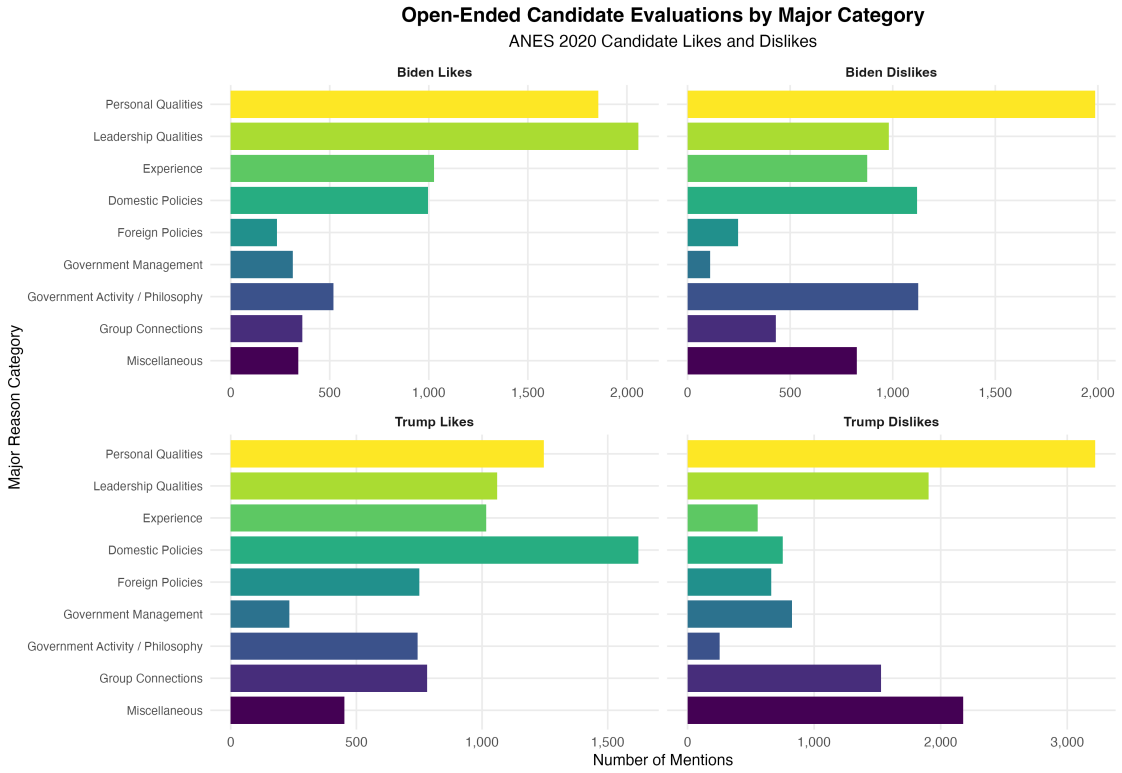

Figure 3: **Open-Ended Candidate Evaluations by Major Category.** Bars report the number of coded mentions in each major category for the four ANES 2020 candidate like/dislike questions. Because responses could receive multiple codes, counts represent coded mentions rather than mutually exclusive respondents.

Figure 4 disaggregates these patterns using the full 35-topic classification. The topic-level results clarify what drives the broader major-category patterns. Among Biden Likes responses, the most common specific topics include Personality–Honesty, Personality–Leadership, Political Experience, Policy–Other, and Better than Alternative. This suggests that favorable evaluations of Biden were rooted both in affirmative perceptions of character and competence and in comparative evaluations against Trump. The presence of Better than Alternative among the most frequent codes is particularly important: for many respondents, liking Biden was not only about positive attachment to Biden but also about viewing him as preferable to the opposing candidate.

Biden Dislikes responses are dominated by Candidate Health / Mental Fitness, Ideology / Philosophy / Views, Policy–Other, Personality–Honesty, Personality–Leadership, and Personality–Ability. This pattern is consistent with the major-category results showing a

heavy concentration in Personal Qualities and Government Activity / Philosophy. The prominence of Candidate Health / Mental Fitness indicates that age, cognitive capacity, and perceived readiness for office were central to many negative evaluations of Biden. At the same time, the frequency of ideology and policy-related codes shows that criticism of Biden was not exclusively personal; it also reflected opposition to his perceived political direction and policy commitments.

Trump Likes responses show a different topic structure. Policy–Economic is one of the most frequently mentioned codes, alongside Personality–Honesty, Personality–Leadership, Policy–International Affairs, Ideology / Philosophy / Views, Personality–Ability, and Non-Political Experience. These results suggest that positive evaluations of Trump were frequently grounded in perceptions of economic performance, toughness, outsider status, ideological alignment, and effectiveness. The relatively high count for Non-Political Experience is substantively important because it captures one of the distinctive features of pro-Trump evaluations: many respondents valued his business background or outsider status rather than conventional political experience.

Trump Dislikes responses are highly concentrated in Personality–Honesty, Personality–Leadership, Personality–Other, Groups, General, Scandal / Cover-up, Candidate Health / Mental Fitness, and Policy–International Affairs. This indicates that negative evaluations of Trump were especially centered on perceptions of dishonesty, temperament, leadership style, and the social or group-based consequences associated with his candidacy. The large number of mentions in Groups and Miscellaneous-related topics also suggests that opposition to Trump was often expressed in broader social, emotional, or conflict-oriented language, rather than only through discrete policy disagreement.

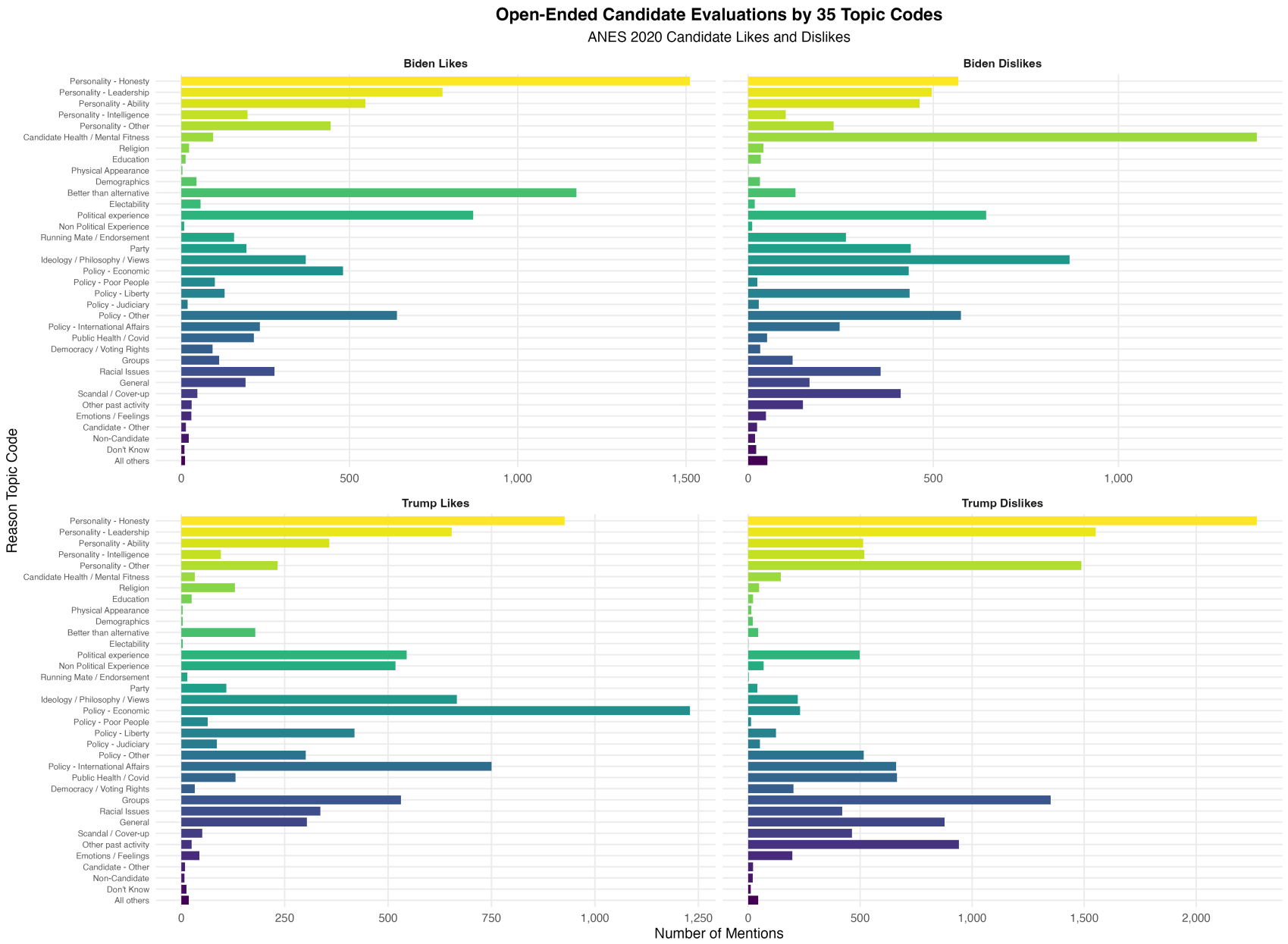

Figure 4: **Open-Ended Candidate Evaluations by 35 Topic Codes.** Counts represent coded mentions across the four 2020 ANES candidate like/dislike questions.

Overall, the results show that candidate evaluations in 2020 were much more complicated than one single variable. Instead, respondents combined personal, comparative, policy-based, ideological, and affective considerations. Still, the balance of these considerations differed by candidate and by whether respondents were explaining likes or dislikes. Biden Likes were especially anchored in character, leadership, experience, and contrast with Trump. Biden Dislikes centered on health and fitness, ideology, leadership, and policy concerns. Trump Likes were more strongly associated with economic policy, leadership, ideology, foreign affairs, and non-political experience. Trump Dislikes were dominated by personal qualities, leadership, group conflict, and broader affective or miscellaneous criticisms.

These descriptive patterns also reinforce the value of the multi-label coding approach. A single response could contain both a policy evaluation and a character judgment, or both an ideological statement and a comparison to the opposing candidate. By preserving multiple codes per response, the AI-coded database captures the layered structure of open-ended candidate evaluations more effectively than a single-label classification would.

6 Conclusion and Implications

This paper demonstrates that API-based large language model classification can provide a scalable and transparent method for coding open-ended survey responses, particularly when embedded within a human-in-the-loop research design. Rather than relying on unsupervised topic discovery or a purely automated zero-shot classification procedure, this project used human coding to establish a benchmark, iterative batch training to calibrate the model to an adapted ANES codebook, and held-out validation data to evaluate the completed AI-coded database. The result is a systematically coded set of 16,416 open-ended 2020 ANES candidate like/dislike responses classified into 35 topic codes and nine broader major categories.

The evidence supports a measured but encouraging conclusion. In the training data, the model achieved levels of agreement with the agreed human-coded standard that were often comparable to human-to-human reliability, especially for Biden Likes, Biden Dislikes, and Trump Likes. The batch-by-batch training results further suggest that iterative feedback helped regularize the model's classifications and improve its ability to apply the codebook consistently. However, the held-out validation exercise provides the more important test of the completed AI-coded database. In that validation sample, the model performed best at the nine-major-category level, where micro-F1 ranged from 0.703 to 0.798 and mean respondent-level Cohen's κ ranged from 0.601 to 0.752. At the more granular 35-topic level, performance was more uneven, with micro-F1 ranging from 0.564 to 0.700 and mean respondent-level κ ranging from 0.488 to 0.699.

These results suggest that AI-assisted coding is especially useful for broad thematic analysis of open-ended responses, while fine-grained topic-level estimates should be interpreted with more caution. The model did not perfectly reproduce human coding at the full 35-topic level, and performance varied by question condition. Biden-related responses were classified more reliably than Trump Dislikes, which proved to be the most difficult condition. The validation results also show that the model tended to assign more codes per response than the human validation coder. This produced relatively high recall but lower precision, indicating that the AI often recovered human-coded topics while also adding additional plausible codes. For survey researchers, this pattern is substantively important: AI classification can expand the usability of open-ended data at scale, but it should be accompanied by validation diagnostics, clear reporting of error patterns, and caution when making highly granular topic-level claims.

The broader implication is that AI-based classification should be treated as a measurement strategy rather than simply a computational shortcut. The value of this approach lies not only in its speed or cost efficiency, but in its ability to apply a predefined codebook to large volumes of open-ended text while preserving a transparent link to human coding standards. Given the lack of standardized quantifiable classifications from the ANES since 2008 (Krosnick et al, 2015), these codes provide political behavior and survey researchers with a new empirical resource for studying contemporary candidate evaluations. At the same time, the findings underscore the importance of validation. AI-generated codes should not be assumed to be equivalent to human coding simply because they are coherent or scalable. They should be evaluated against human benchmarks using agreement statistics appropriate for multi-label classification, including exact-set agreement, Jaccard similarity, precision, recall, F1, and respondent-level κ.

Substantively, the newly coded data reveal that 2020 candidate evaluations were multi-dimensional. Respondents did not simply like or dislike candidates on the basis of policy, personality, or party alone. Instead, open-ended evaluations frequently combined personal traits, leadership judgments, ideological considerations, policy preferences, comparative evaluations, and affective reactions. Biden Likes were especially anchored in character, leadership, experience, and contrast with Trump. Biden Dislikes centered on health and fitness, ideology, leadership, and policy concerns. Trump Likes were more strongly associated with economic policy, leadership, ideology, foreign affairs, and non-political experience. Trump Dislikes were dominated by personal qualities, leadership, group conflict, and broader affective or miscellaneous criticisms. These patterns suggest that open-ended survey responses continue to offer distinctive insight into the considerations voters use when explaining candidate support and opposition.

The prominence of comparative and negative reasoning is especially important. The

"Better than the Alternative" code appeared frequently in Biden Likes, illustrating that support for Biden often reflected comparative evaluation rather than only affirmative attachment. Likewise, the dislike responses for both candidates frequently invoked perceived deficiencies in leadership, integrity, competence, health, or temperament. These findings align with broader theories of affective polarization in American politics (Iyaengar, 2015), particularly where disapproval of the opposing candidate appears to structure political evaluation alongside support for one's own preferred candidate. The results also suggest that long-standing questions about levels of conceptualization and the structure of political reasoning remain highly relevant in the contemporary electorate (Campbell et al, 1960). By making open-ended candidate evaluations available in a standardized, analyzable format, this project creates new opportunities to study how voters articulate ideology, issue positions, group attachments, and affective judgments in their own words.

Future research can build on this project in several ways. First, scholars can use the nine-major-category classifications for broad analyses of candidate evaluation, negativity, and voter reasoning in 2020. Second, researchers can use the 35-topic codes for more detailed analysis while incorporating the validation diagnostics reported here to assess which topics are most reliable. Third, the workflow itself can be extended to other years of ANES open-ended responses, allowing scholars to revisit historical questions about change in political reasoning, affective polarization, and campaign evaluation. Finally, the approach can be adapted for other open-ended survey instruments where researchers need to apply a fixed codebook at scale. In this sense, the contribution of the paper is both substantive and methodological: it provides a new coded dataset of 2020 candidate evaluations and offers a replicable model for human-supervised, API-based classification of open-ended survey responses.

Ultimately, the evidence does not suggest that AI coding should replace human judgment. Rather, it suggests that human-supervised AI classification can extend the reach of expert coding by making large-scale open-ended survey data more usable, transparent, and replicable. When paired with careful validation, API-based classification provides a practical way to recover the richness of respondents' own language while producing structured data suitable for quantitative analysis. This makes open-ended responses a more accessible resource for studying political behavior, campaign effects, and the evolving nature of candidate evaluation in American elections.

Funding Statement No outside funding was provided for this project

Competing Interests No competing interest to report on this project

Ethical Standards The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Author Contributions

References

[1] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. "Out of one, many: Using language models to simulate human samples." *Political Analysis*, 31(3): 337–351. doi:10.1017/pan.2023.2.

[2] Jacob Beck, Stephanie Eckman, Bolei Ma, Rob Chew, and Frauke Kreuter. 2024. "Order effects in annotation tasks: Further evidence of annotation sensitivity." pp. 81–86, March. URL: https://huggingface.co/datasets/soda-lmu/tweet-annotation-sensitivity-2.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language models are few-shot learners." In *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[4] Angus Campbell, Philip E. Converse, Warren E. Miller, and Donald E. Stokes. 1960. *The American Voter*. Chicago: University of Chicago Press. ISBN 9780226092546. URL: https://books.google.com/books?id=JeYUrs_GOcMC.

[5] Matthew DeBell. 2013. "Harder than it looks: Coding political knowledge on the ANES." *Political Analysis*, 21(4).

[6] Justin Grimmer. 2022. *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton: Princeton University Press. ISBN 978-0-691-20754-4.

[7] William Hobbs and Jared Green. 2025. "Categorizing topics versus inferring attitudes: A theory and method for analyzing open-ended survey responses." *Political Analysis*, 33(3): 231–251.

[8] Shohei Eshima, Kosuke Imai, and Takuya Sasaki. 2024. "Keyword-assisted topic models." *American Journal of Political Science*, 68(2): 730–750.

[9] Stanley Kelley and Thad W. Mirer. 1974. "The simple act of voting." *The American Political Science Review*, 68(2): 572–591. doi:10.2307/1959506.

[10] Kathleen Knight. 1984. "The dimensionality of partisan and ideological affect: The influence of positivity." *American Politics Quarterly*, 12(3): 305–334. doi:10.1177/1532673X8401200304.

[11] Jon A. Krosnick, Arthur Lupia, and Matthew Berent. 2015. "Coding open-ended answers to questions measuring reasons to vote for or against candidates from the 2008 ANES time series interviews." *American National Election Studies*.

[12] Richard R. Lau. 1982. "Negativity in political perception." *Political Behavior*, 4: 353–377.

[13] Richard R. Lau. 1986. "Political schemata, candidate evaluations, and voting behavior." In Richard R. Lau and David O. Sears (eds.), *Political Cognition: The 19th Annual Carnegie Symposium on Cognition*, pp. 95–125. Hillsdale, NJ: Erlbaum.

[14] Richard R. Lau. 1989. "Construct accessibility and electoral choice." *Political Behavior*, 11(1): 5–32.

[15] Arthur Lupia. 2018. "How to improve coding for open-ended survey data: Lessons from the ANES." In *The Palgrave Handbook of Survey Research*, pp. 121–127.

[16] Jonathan Mellon, Jack Bailey, Ralph Scott, James Breckwoldt, Marta Miori, and Phillip Schmedeman. 2024. "Do AIs know what the most important issue is? Using language models to code open-text social survey responses at scale." *Research & Politics*, 11(1). doi:10.1177/20531680241231468.

[17] Chau Minh Pham, Alexander Hoyle, Simeng Sun, and Mohit Iyyer. 2023. "TopicGPT: A prompt-based topic modeling framework."

[18] Julie A. Phillips, Thomas R. Davidson, and Marilyn S. Baffoe-Bonnie. 2023. "Identifying latent themes in suicide among Black and White adolescents and young adults using the National Violent Death Reporting System, 2013–2019." *Social Science & Medicine*, 334: 116144. doi:10.1016/j.socscimed.2023.116144.

[19] Matthew T. Pietryka and Randall C. MacIntosh. 2013. "An analysis of ANES items and their use in the construction of political knowledge scales." *Political Analysis*, 21(4): 407–429.

[20] David Repass. 2018. "Problems with open-ended ANES questions measuring what respondents like and dislike about candidates and political parties." *ANES*. URL: `https://electionstudies.org/wp-content/uploads/2018/03/RePass_transcript.pdf`.

[21] David E. Repass. 2020. *Listening to the American Voter: What Was on Voters' Minds in Presidential Elections, 1960 to 2016*. New York: Routledge.

[22] Travis N. Ridout and Michael M. Franz. 2011. *The Persuasive Power of Campaign Advertising*. Philadelphia: Temple University Press.

[23] Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. 2016. "A model of text for experimentation in the social sciences." *Journal of the American Statistical Association*, 111(515): 988–1003. doi:10.1080/01621459.2016.1141684.

[24] Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, David G. Rand, and Margaret E. Roberts. 2014. "Structural topic models for open-ended survey responses." *American Journal of Political Science*, 58(4): 1064–1082. doi:10.1111/ajps.12103.

[25] Simone Zhang, Janet Xu, and AJ Alvero. 2023. "Generative AI meets open-ended survey responses: Participant use of AI and homogenization."

7 Appendix A: Usage of Inter-coder reliability scoring metrics

We first report *exact-set agreement*, the strictest measure of agreement. Exact-set agreement equals one only when the AI and human coder assign the same complete set of codes to a response:

$$\text{Exact}_i = \begin{cases} 1, & \text{if } A_i = H_i, \\ 0, & \text{otherwise.} \end{cases}$$

The overall exact-set agreement rate is:

$$\text{Exact} = \frac{1}{N} \sum_{i=1}^{N} \text{Exact}_i \, .$$

To capture partial agreement, we report *Jaccard similarity*, which measures the proportion of shared assigned codes among all unique codes assigned by either source:

$$J_i = \frac{|A_i \cap H_i|}{|A_i \cup H_i|}.$$

When both sets are empty, J_i is defined as 1, although such cases are rare in the present application.

We also report response-level precision, recall, and F1. For each respondent, true positives are the codes assigned by both the AI and the human coder, false positives are codes assigned by the AI but not by the human coder, and false negatives are codes assigned by the human coder but missed by the AI:

$$\begin{aligned} TP_i &= |A_i \cap H_i|, \\ FP_i &= |A_i \setminus H_i|, \\ FN_i &= |H_i \setminus A_i|. \end{aligned}$$

Response-level precision is:

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i},$$

and response-level recall is:

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i}.$$

The response-level F1 score is the harmonic mean of precision and recall:

$$F1_i = 2 \times \frac{\text{Precision}_i \times \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}.$$

In addition to response-level averages, we report micro-averaged precision, recall, and F1. Micro-averaging pools all code decisions across all respondents before calculating the statistic:

$$TP = \sum_{i=1}^{N} TP_i,$$
$$FP = \sum_{i=1}^{N} FP_i,$$
$$FN = \sum_{i=1}^{N} FN_i.$$

The micro-averaged statistics are then:

$$\text{Precision}_\mu = \frac{TP}{TP + FP},$$
$$\text{Recall}_\mu = \frac{TP}{TP + FN},$$
$$F1_\mu = 2 \times \frac{\text{Precision}_\mu \times \text{Recall}_\mu}{\text{Precision}_\mu + \text{Recall}_\mu}.$$

8 Appendix B: Batching Data

Batch	Biden Likes	Biden Dislikes	Trump Likes	Trump Dislikes
1	0.471	0.605	0.721	0.619
2	0.756	0.616	0.702	0.427
3	0.668	0.796	0.680	0.469
4	0.728	0.830	0.496	0.574
5	0.757	0.640	0.568	0.599
6	0.579	0.770	0.593	0.508
7	0.737	0.629	0.674	0.532
8	0.814	0.801	0.633	0.645
9	0.701	0.706	0.575	0.542
10	0.788	0.782	0.686	0.459
11	0.872	0.574	0.811	0.476
12	0.862	0.695	0.688	0.616
13	0.641	0.633	0.805	0.616
14	0.718	0.777	0.455	0.715
15	0.670	0.808	0.827	0.475
16	0.806	0.553	0.687	0.578
17	0.646	0.653	0.784	0.530
18	0.947	0.784	0.707	0.584
19	0.796	0.657	0.598	0.717
20	0.706	0.751	0.759	0.438
Mean	**0.733**	**0.703**	**0.672**	**0.556**

Table 9: **Mean Respondent-Level Cohen's κ by Training Batch.** Each batch contains 10 training responses. Entries report the mean respondent-level Cohen's κ comparing AI-assigned codes to the agreed human-coded standard for each candidate like/dislike condition.

9 Appendix C: 2020 ANES Likes/Dislikes Codebook

The following coding guidelines were provided to the human coding team and were also incorporated into the AI prompt instructions. The codebook was adapted from Krosnick and Lupia's ANES open-ended coding framework and updated for the 2020 presidential election context (Krosnick et al., 2015).

9.1 Project Summary

The objective of this coding task is to classify open-ended responses from the 2020 presidential election. For this component of the survey, respondents were asked four candidate evaluation questions:

1. What is it that you **like** about **Joe Biden**?
2. What is it that you **dislike** about **Joe Biden**?
3. What is it that you **like** about **Donald Trump**?
4. What is it that you **dislike** about **Donald Trump**?

For each response, coders were asked to identify which topics were mentioned among 35 agreed-upon topic codes. The same 35 codes apply across all four questions. Because responses may contain multiple distinct considerations, each response can receive one or several codes depending on the content of the answer. Coders were instructed to read each response carefully and assign all topic codes that were substantively present.

9.2 Thirty-Five Topic Codes

Table 10: 2020 ANES Candidate Likes/Dislikes Codebook: 35 Topic Codes

#	Label	Annotation Criteria
1	General	Good person, bad person, likable, not likable, "I like him," "I do not like him," or any general feeling toward a candidate that is not more specific.
2	Party	Mentions that the candidate is or is not a Democrat, is or is not a member of the Democratic Party, is or is not a Republican, or is or is not a member of the Republican Party.

Continued on next page

#	Label	Annotation Criteria
3	Ideology / Philosophy / Views	Mentions conservative, liberal, socialist, Marxist, communist, fascist, libertarian, moderate, America First, MAGA, shares my views, or opposes my views.
4	Electability / Electoral Contest	Mentions that the candidate can win or cannot win.
5	Better than the Alternative	Mentions "Not Trump," "Not Biden," lesser of two evils, or similar comparative reasoning.
6	Political Experience	Mentions political experience, work the candidate has done as a politician or elected official, being experienced or inexperienced in government, insider status, incumbent job performance, or political record.
7	Non-Political Experience	Mentions business experience, work in business or industry, educational experience, outsider status, or support rooted in non-political background.
8	Scandal / Cover-up	Mentions that the candidate was involved in a scandal, cover-up, or controversy.
9	Other Past Activity	Mentions something the candidate did in the past that does not match another code.
10	Personality – Ability	Mentions the candidate's ability to accomplish a task, get the job done, keep promises, or fulfill objectives.
11	Personality – Honesty	Mentions honesty, integrity, consistency, predictability, sincerity, empathy, truthfulness, character, or caring about people like the respondent.
12	Personality – Intelligence	Mentions the candidate's intelligence, knowledge, understanding, or lack thereof.
13	Personality – Leadership	Mentions the candidate's ability to lead, get people to work together, inspire people, motivate people, serve as a strong or weak leader, fight for people like the respondent, or express patriotic sentiments.
14	Personality – Other	Mentions anything about the candidate's personality that does not fit codes 10, 11, 12, or 13.
15	Religion	Mentions that the candidate does or does not belong to a specific religious group, is religious or not religious, or advances a religious agenda.
16	Education	Mentions the candidate's education, where the candidate went to school or college, or the candidate's diploma or degree.
17	Physical Appearance	Mentions the candidate's physical appearance, attractiveness, or how well the candidate looks.

Continued on next page

#	Label	Annotation Criteria
18	Demographics	Mentions the candidate's age, sex, sexual orientation, race, ethnicity, height, weight, marital status, or where the candidate is from.
19	Policy – Economic	Mentions the economy, taxes, government spending, the budget, deficit, national debt, economic policy, fiscal concerns, or Covid stimulus.
20	Policy – Poor People	Mentions government programs to help poor people, such as welfare, food stamps, Medicaid, Aid for Families with Dependent Children, or public housing. This code should be assigned only to ideas that mention helping poor people or government programs to help poor people. General references to caring about poor people are assigned to Groups.
21	Policy – Liberty	Mentions what the candidate will allow people to do, prevent people from doing, or make legal or illegal. Examples include abortion, LGBTQ policy, marijuana, guns, or similar liberty-related issues.
22	Policy – International Affairs	Mentions how the candidate will deal with other countries, foreign policy, the reputation of the United States, relations with other nations, or support/opposition to the military.
23	Policy – Other	Mentions what the candidate will do about a policy issue that does not match codes 19–22 or 31–35. Also includes general references to change or policy positions that are not more specifically classified.
24	Groups	Mentions the candidate's support for or opposition to ideological or other groups, feelings about a specific group of people, or attacks on candidates' supporters.
25	Emotions / Feelings	Mentions that the candidate makes the respondent feel happy, sad, angry, proud, afraid, scared, or otherwise evokes a generic emotional reaction.
26	Candidate – Other	Mentions something about Trump or Biden that does not match another candidate-related code.
27	Non-Candidate	Mentions something other than the candidate and not about the respondent.
28	Running Mate / Endorsement	Mentions the running mate, such as Kamala Harris or Mike Pence, or key endorsements from respected political or non-political figures, party luminaries, or public figures.
29	Don't Know	Mentions "I don't know," "don't know," "DK," "I'm unsure," "I'm not sure," "unsure," "you got me," "I can't remember," "no clue," "no idea," "no guess," "no," or "nothing."
30	All Others	Any statement that does not fit the other codes.

Continued on next page

#	Label	Annotation Criteria
31	Judiciary	Mentions the Supreme Court, judicial appointments, court packing, or other references to judicial policy.
32	Democracy / Voting Rights	Mentions election fraud, voter suppression, ballot access, democratic insecurities, mail-in voting, voting rights, or democracy-related concerns.
33	Racial Issues	Mentions racial issues, civil equality, defund the police, law and order, pro-police sentiments, racial tensions, racial divisions, or concerns over race relations.
34	Candidate Health / Mental Fitness	Mentions the candidate's health, mental sharpness, competency, cognitive sharpness, or readiness to serve. This is distinct from general ability and should be used for health, age, cognitive fitness, or readiness concerns.
35	Public Health / Covid	Mentions pro- or anti-Covid mitigation, masks, lockdowns, school closures, public health infrastructure, pandemic planning, or Covid-related public health concerns.

9.3 Steps for Coding

Coders were instructed to use the following process:

1. Read each response in detail.
2. Identify every distinct topic mentioned in the response.
3. Assign all applicable topic codes. Responses are not restricted to a single code.
4. If a topic has been selected, mark the corresponding code for that response.
5. Leave notes when necessary to clarify ambiguous coding decisions.

For example, consider the following response to the Joe Biden Likes question:

> "He isn't Donald Trump. He shares my view of what America stands for. I want to save our democracy."

This response contains several distinct topics:

- **5: Better than the Alternative** because the respondent says Biden "isn't Donald Trump."
- **3: Ideology / Philosophy / Views** because the respondent says Biden shares their view of what America stands for.

- **32: Democracy / Voting Rights** because the respondent says they want to save democracy.

Coders were reminded that the task involves judgment and that there is not always one mechanically correct way to code a response. The goal was to fully vet each answer and assign all codes that were substantively justified by the respondent's language.

10 Appendix D: Major Category Inter Coder Reliability Scores

Table 11: **Major Category ICR Scores: Human Coder 1 vs. Human Coder 2.** Entries report category-level F1 scores for the 9-major-category coding scheme. Parentheses report the number of responses assigned to each category by the reference coder.

Major Category	**Biden Likes**	**Biden Dislikes**	**Trump Likes**	**Trump Dislikes**
Experience	0.815 (55)	0.884 (49)	0.763 (57)	0.222 (9)
Leadership Qualities	0.827 (105)	0.255 (19)	0.648 (69)	0.689 (86)
Personal Qualities	0.849 (91)	0.883 (102)	0.617 (37)	0.858 (126)
Government Management	0.824 (8)	0.353 (3)	0.696 (9)	0.771 (35)
Government Activity / Philosophy	0.786 (32)	0.894 (66)	0.468 (17)	0.364 (5)
Domestic Policies	0.646 (41)	0.843 (52)	0.897 (87)	0.681 (29)
Foreign Policies	0.778 (11)	0.842 (9)	0.765 (29)	0.688 (17)
Group Connections	0.480 (11)	0.538 (9)	0.545 (7)	0.185 (10)
Miscellaneous	0.333 (14)	0.639 (36)	0.379 (15)	0.439 (36)

Table 12: **Major Category ICR Scores: AI vs. Agreed Hand Code.** Entries report category-level F1 scores comparing AI classifications to the agreed human-coded benchmark for the 9-major-category coding scheme. Parentheses report the number of responses assigned to each category in the agreed hand-coded benchmark.

Major Category	**Biden Likes**	**Biden Dislikes**	**Trump Likes**	**Trump Dislikes**
Experience	0.920 (57)	0.926 (46)	0.811 (53)	0.424 (15)
Leadership Qualities	0.892 (101)	0.618 (27)	0.677 (55)	0.730 (81)
Personal Qualities	0.905 (97)	0.909 (104)	0.718 (55)	0.861 (122)
Government Management	0.800 (9)	0.235 (13)	1.000 (10)	0.921 (38)
Government Activity / Philosophy	0.862 (31)	0.950 (58)	0.694 (31)	0.762 (9)
Domestic Policies	0.725 (37)	0.893 (49)	0.940 (84)	0.766 (23)
Foreign Policies	0.857 (9)	0.762 (10)	0.862 (32)	0.773 (18)
Group Connections	0.585 (15)	0.579 (18)	0.704 (25)	0.814 (48)
Miscellaneous	0.486 (17)	0.620 (36)	0.464 (35)	0.750 (88)

Table 13: **Major Category ICR Scores: Full AI Output vs. Hand-Coded Validation Data.** Entries report category-level F1 scores comparing the full AI-coded database to the hand-coded validation data for the 9-major-category coding scheme. Parentheses report the number of validation responses assigned to each category by the human validation coder.

Major Category	**Biden Likes**	**Biden Dislikes**	**Trump Likes**	**Trump Dislikes**
Experience	0.872 (46)	0.809 (47)	0.651 (28)	0.444 (4)
Leadership Qualities	0.844 (97)	0.646 (38)	0.644 (91)	0.695 (90)
Personal Qualities	0.886 (75)	0.905 (98)	0.718 (45)	0.879 (129)
Government Management	0.711 (23)	0.667 (6)	0.645 (17)	0.847 (29)
Government Activity / Philosophy	0.642 (27)	0.814 (54)	0.523 (19)	0.625 (7)
Domestic Policies	0.784 (40)	0.845 (50)	0.892 (64)	0.789 (37)
Foreign Policies	0.471 (4)	0.696 (8)	0.905 (41)	0.638 (16)
Group Connections	0.769 (15)	0.737 (14)	0.625 (26)	0.741 (48)
Miscellaneous	0.300 (8)	0.700 (34)	0.400 (22)	0.492 (36)