Full PICSO Report on the SPOTLIGHT 2024 Election Chatbot

Sodi E. Kroehler Madeline I. Franz Nefriana
Dr. Elise Silva, PhD, MLIS Dr. Beth Schwanke, JD
Dr. Yuru Lin

2025-08-02

Table of contents

1 Full PICSO Report on the SPOTLIGHT 2024 Election Newsbot

2 Executive Summary

Between August and November 5th, 2024, the Spotlight News team deployed an experimental chatbot on their public-facing website to assist users in navigating the complexities of the 2024 U.S. presidential election.

Following the election, an anonymized dataset of user queries and chatbot responses was shared with researchers from the University of Pittsburgh's PICSO Lab and Pitt Cyber. The goal was to rigorously evaluate the system's performance, political neutrality, and factual reliability using a combination of human annotation and computational analysis.

This report presents the complete findings of that evaluation. It includes: - A detailed description of the data cleaning and annotation methodology - A breakdown of inter-annotator agreement and flagging patterns - Evaluation of the chatbot's response quality, including error types and topic coverage - Summary performance metrics for a follow-up fine-tuned model

While the analysis is extensive, this document is focused strictly on methodological and quantitative aspects. Interpretive commentary, qualitative insights, and broader discussion are deferred to separate publications.

Many associated model weights, preprocessing scripts, and code artifacts are available on GitHub:
https://github.com/SodiKroehler/spot-less-light

This document follows the structure of a traditional research report, including dedicated **Methodology**, **Results**, and **Appendix** sections.

3 Introduction

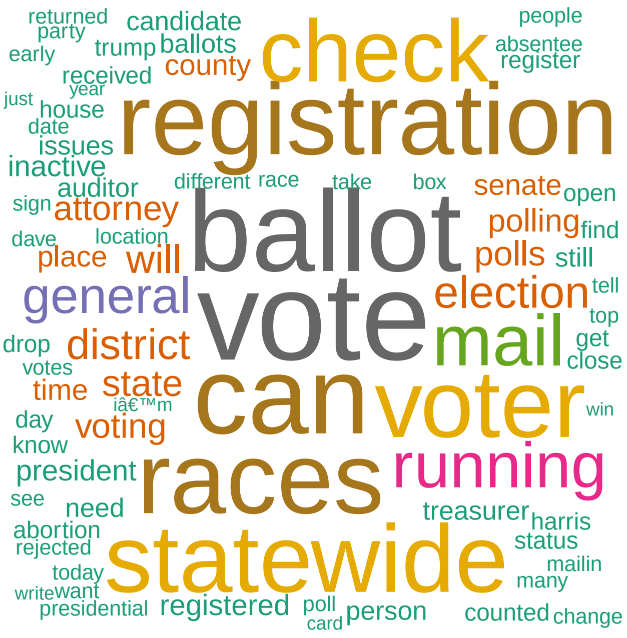
3.1 Conversations

Groups of user utterances were grouped based on the conversation they occurred in, where a conversation is defined as a single user engaging with the bot in a contiguous space of time. A total of 1660 conversations were recorded, with a total of 2 utterances. The average conversation length was 1.75 utterances, with the the longest at 15 utterances.

3.2 Utterances

Utterances had an average word count of 8.38 words, with the longest utterance being 295 words long. There were 1903 duplicate utterances, with 715 of those being exact duplicates within the same conversation.

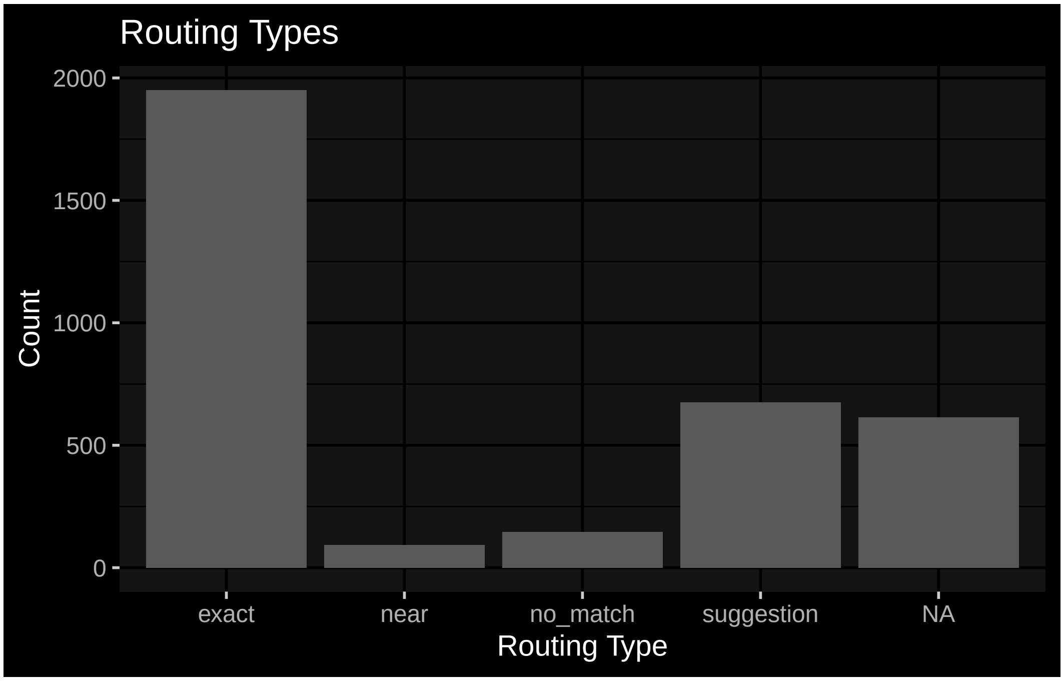
Here is a word cloud of the top 100 words used in the utterances:



3.3 Bot Routing

Due to concerns about hallucination, trust, and other ethical issues, the Spotlight team structured the bot only to match user questions to pre-defined question/answer pairs. This predefined list, while expansive, did not contain every question asked by users. The tool recorded routing information for each utterance, either “exact” if it directly matched a question, or something else if it didn’t. A histogram of these values is available below:

```
Inverted geom defaults of fill and color/colour.
To change them back, use invert_geom_defaults().
```



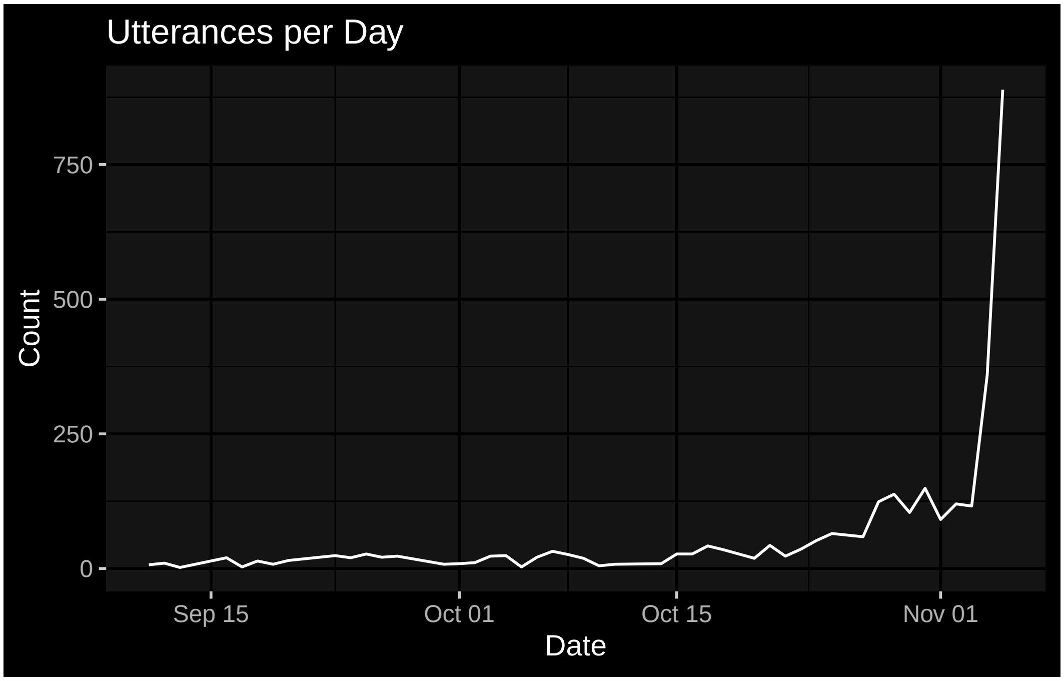
Human coders did not find these automatically created values to be very accurate. A further discussion of this is available in the human coding section.

3.4 Bot Answers

Answers seemed to be generatively reworded from the predefined master list, and gave very similar, but still slightly different responses to the same question. Overall, the answers had a mean length of 48.28 words, with the longest answer being 273 words long. There were 78 utterances which were exact duplicates but received different bot answers.

3.5 Timing

Conversations lasted an average of 0.8 seconds, with the longest conversation lasting 47 seconds. The earliest given utterance was at 2024-09-11 12:18:00 and the latest utterance was at 2024-11-05 19:58:00. Below is a line graph showing the frequency of utterances, per day, over the full time period.



4 Human Coding

Three human coders were employed to manualy code a subgroup of these utterance/answer pairs. Coding occurred across five rounds, and lasted for five months. Meetings were held after each round to review agreement and process. All codebooks and examples were refined iteratively. The rounds of coding, as well as details about them, are summarized in the table below:

Table 4.1: Summary of Annotation Rounds

Round	Coders	Utterances	Coding_Dates	Meeting_Date	Tooling
1	Sodi, Maddy, Nefriana	50 (shared)	Dec 3–5	Dec 7	Manual
2	Sodi, Maddy, Nefriana	50 (shared)	Dec 9–12	Dec 13	Manual
3	Sodi, Maddy, Nefriana	50 (shared)	Dec 9–Jan 16	Jan 17	App
4	Sodi, Maddy, Nefriana	50 (shared)	Jan 29–Feb 6	Feb 10	App
5	Maddy only	1005	Feb 26–May 4	N/A	App

The app was a tool created in Google Appsheets which helped organize and streamlined the coding process. Functionally, all codes were stored in a google sheets, and some coders preferred to edit directly in the sheet rather than use the app.

4.0.1 Sampling Weights

```r
sampling_notes <- data.frame(
  Round = 1:5,
  Sampling = c(
    "Weighted: recency (90%) + low-confidence (10%)",
    "Same weights; not in R1; conversation_size = 1",
    "Same weights; not in R1/R2; conversation size   6",
    "No weights; not in R1-R3",
    "Random from all uncoded utterances; no weights"
  ),
  stringsAsFactors = FALSE
)
```

```r
kable(
  sampling_notes,
  caption = "Sampling Strategies Used for Each Annotation Round",
  align = "l"
)
```

Table 4.2: Sampling Strategies Used for Each Annotation Round

Round	Sampling
1	Weighted: recency (90%) + low-confidence (10%)
2	Same weights; not in R1; conversation_size = 1
3	Same weights; not in R1/R2; conversation size 6
4	No weights; not in R1–R3
5	Random from all uncoded utterances; no weights

Each annotation round used a different sampling strategy, decided upon by all coders and advisors. Round 1 was sampled randomly, but weighted 90% to prefer more recent queries, and 10% to prefer utterances where the routing option was not "exact" (more on this later). Round 2 had the same weights (with utterances coded in R1 removed, and also restricted to conversations with only one utterance. Round 3 did the opposite, and only looked at longer conversations. Round 4 and 5 did not have any weighting.

These weighting options are summarized in the table below.

4.0.2 Coding Specifics

All utterance/answer pairs were coded for the following categories:
- **Primary Code**: The general category of the utterance.
- **Secondary Code**: A more specific sub-category of the utterance.
- **Sentiment**: Usually "NEUTRAL" but marked as "POSITIVE" or "NEGATIVE" in more noteworthy cases.
- **Answer Rating**: How good the bot's answer was, and whether this was acceptable given the context.
- **Flags**: A series of flags to indicate special characteristics: - **Trust Flag**: If the user seems dubious, there could be conspiratorial thinking, or even just the answer given was misleading.
- **Context**: If the context of the conversation is important to why we coded as we did. - **Repeated Question**: If the user is repeating their question exactly. This might indicate accidental misclick, or jailbreaking attempts.
- **Wack Answer**: Occasionally the bot gave very odd answers that didn't match the question at all.

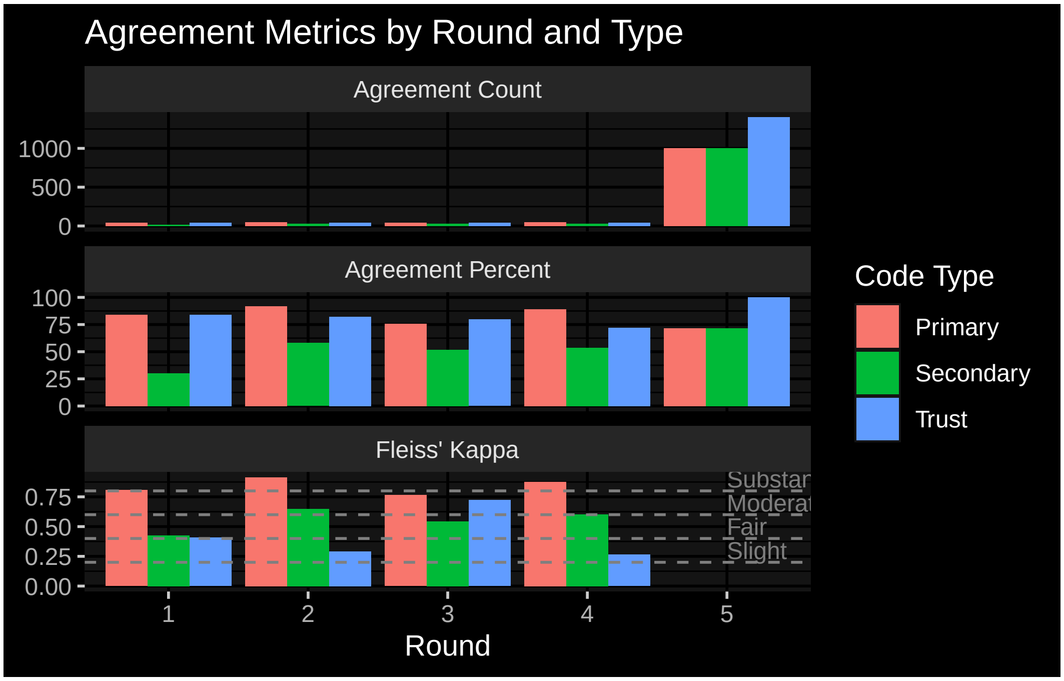
A full list of all values is available in the APPENDIX.

4.0.3 Inter-Rater Reliability

At each post-round meeting, we compared our answers, and tried to come to agreement on any differences. We focused our conversations on three principal aspects: primary code, secondary code, and trust flag, as these are the ones that we feel are most important for future work.

```
Inverted geom defaults of fill and color/colour.
To change them back, use invert_geom_defaults().
```

```
Warning: Removed 3 rows containing missing values or values outside the scale range
(`geom_col()`).
```



4.1 AI Classifier

After all these rounds, an LLM was fine-tuned based on the gold labels provided by the coders. Weights for this model are available on the GitHub, and can be easily extended to label new utterances, with only minimal preprocessing.

4.1.1 Classifier Details

We began with the TinyLlama/TinyLlama-1.1B-Chat-v1.0 model, using the same as an encoder. A good number of different parameter values and architectures were experimented with. The final model was trained over 3 epochs, with a per-batch size of 4, and a learning rate of 2e-5. The model was trained on two cores of a A100 GPU, and took approximately 2 hours to train.

4.1.2 Classifier Performance

Table 4.3: Overall Model Performance (Precision, Recall, F1)

Method	precision	recall	f_meas
Macro	0.853	0.844	0.849
Micro	0.865	0.865	0.865
Weighted	0.869	0.865	0.864

A more detailed discussion of model performance is available in the appendix.

5 Results

Using both the human and AI coding methods described above, we obtained a fully coded dataset. In this next section, we look at the codes and their interactions.

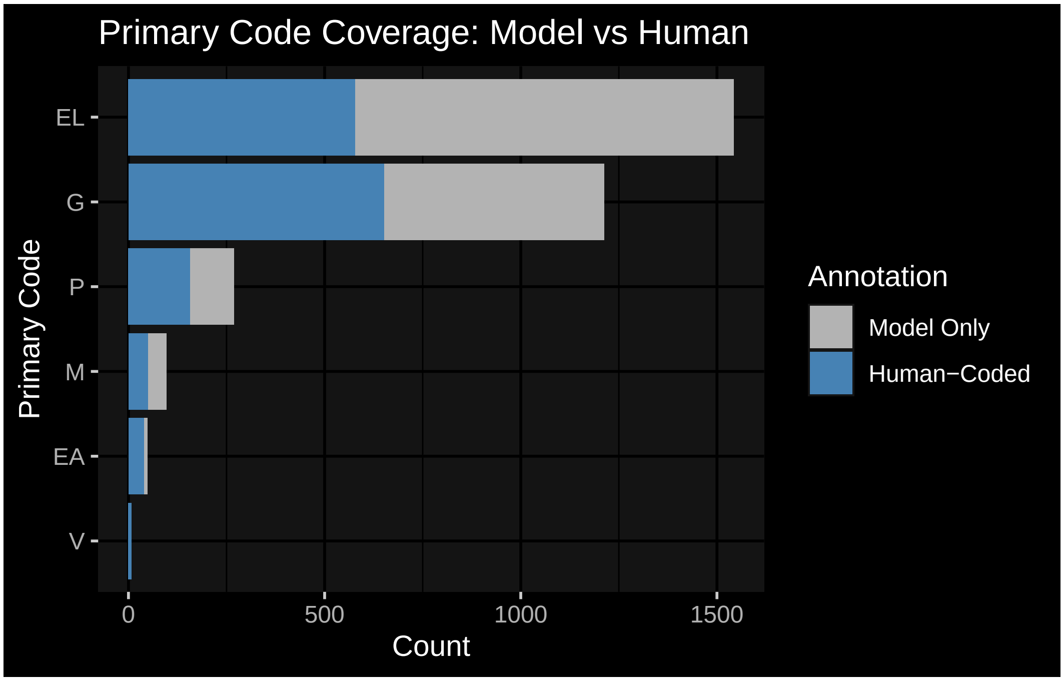
5.1 Distributions

The distribution information given below includes all utterances - both the ones which were human-coded as well as utterances which were not human coded but were coding with the classifer (if applicable). In instances where the human coders disagreed, the final value was decided on by all the coders, with ties being broken by one or more of the advisors.

In the graphs below, if there are classifier-created codes, they are shown in grey while the human-coded values are shown in blue.

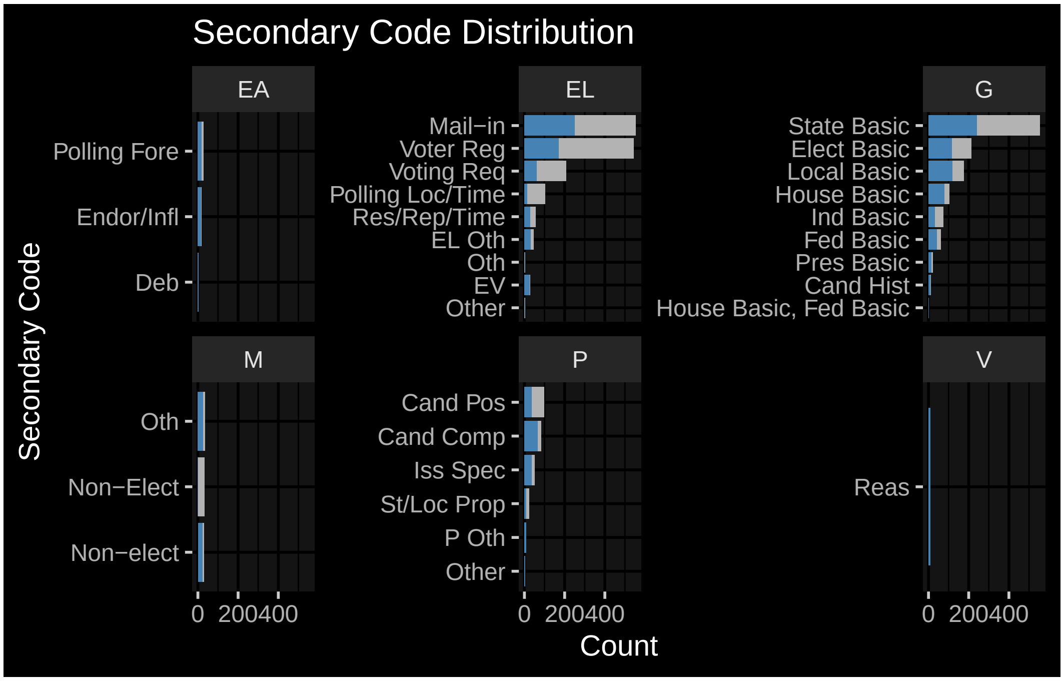
5.1.1 Primary Code

```
Inverted geom defaults of fill and color/colour.
To change them back, use invert_geom_defaults().
```

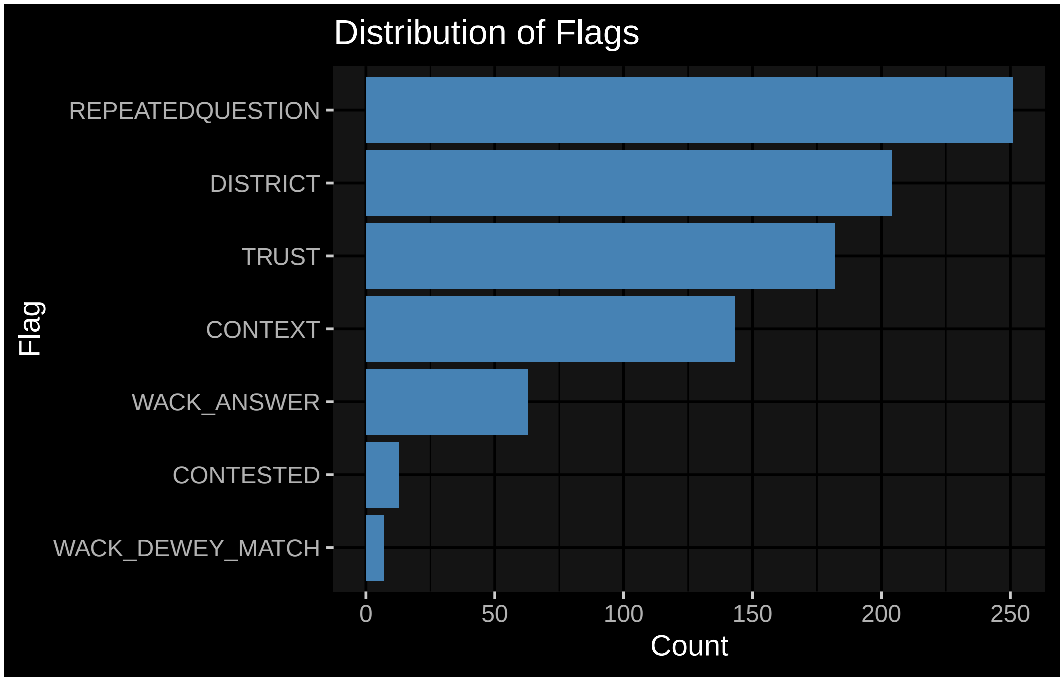


5.1.2 Secondary Codes

Note that some utterances were coded with multiple secondary codes.

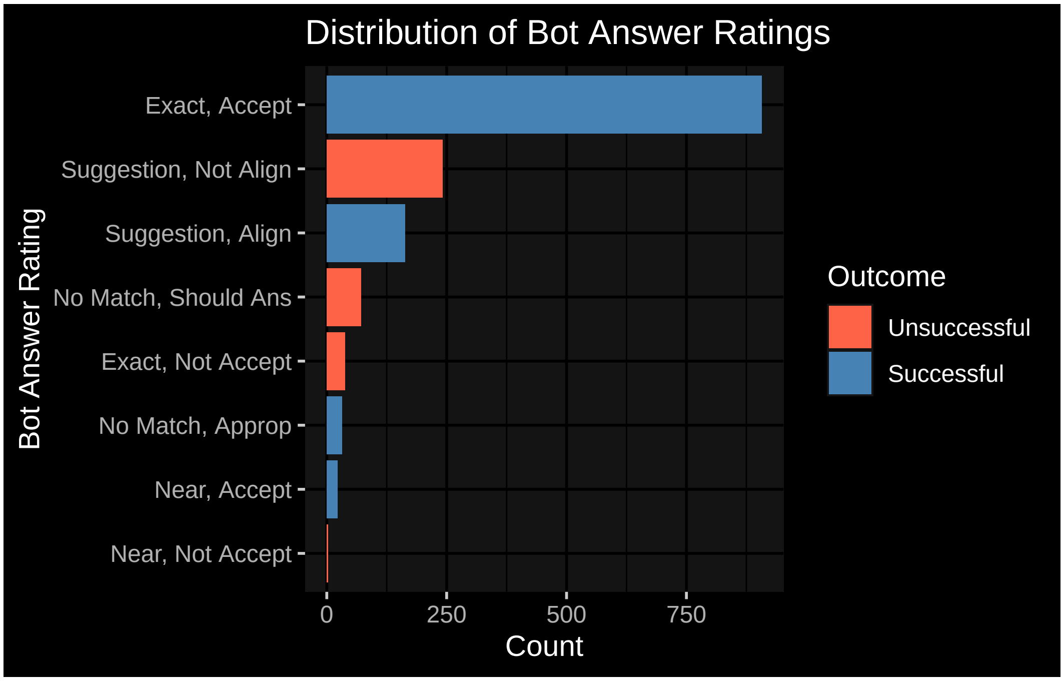


5.1.3 Flags



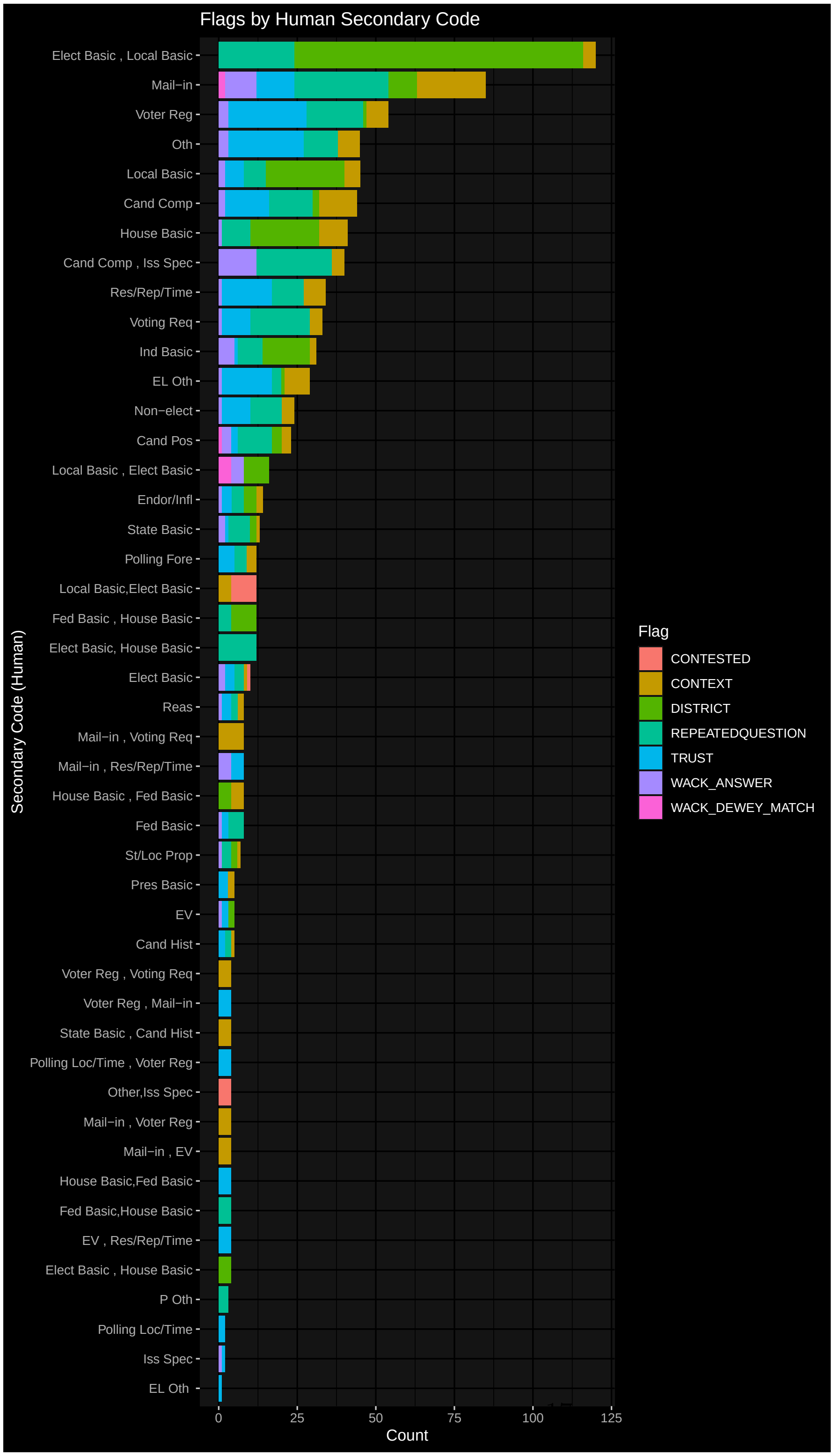
5.1.4 Accept/Not Accept

We wanted to record a wider scope of bot-generated acceptability, so we use a custom set of 8 possible values. A full breakdown of the meaning of these is available in the appendix; in this graph we show the distribution, color coded depending on whether it can be take as a "successful" or "unsuccessful" answer.

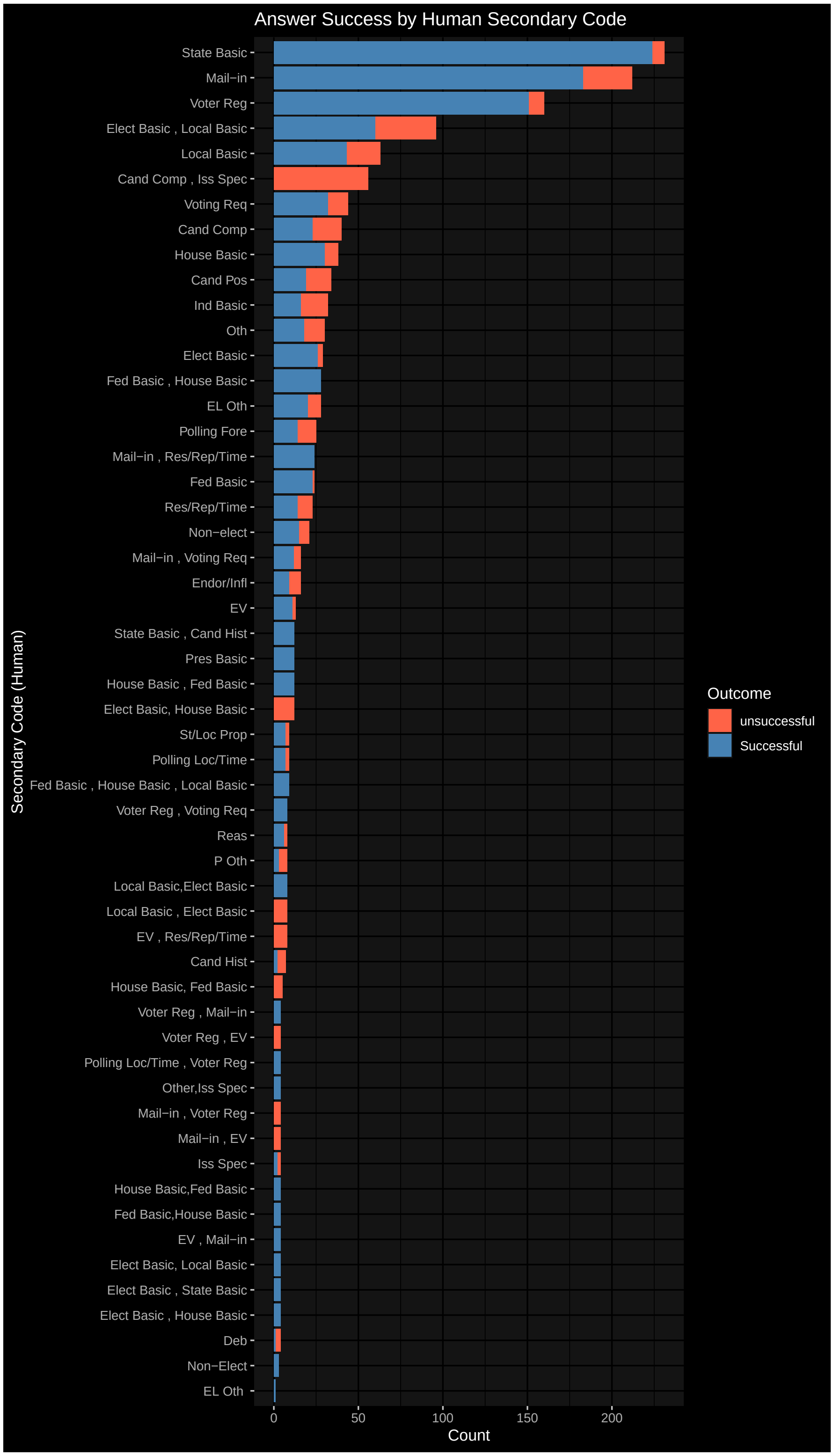

Distribution of Bot Answer Ratings
Bot Answer Rating
Exact, Accept
Suggestion, Not Align
Suggestion, Align
No Match, Should Ans
Exact, Not Accept
No Match, Approp
Near, Accept
Near, Not Accept
0
250
500
750
Count
Outcome
Unsuccessful
Successful

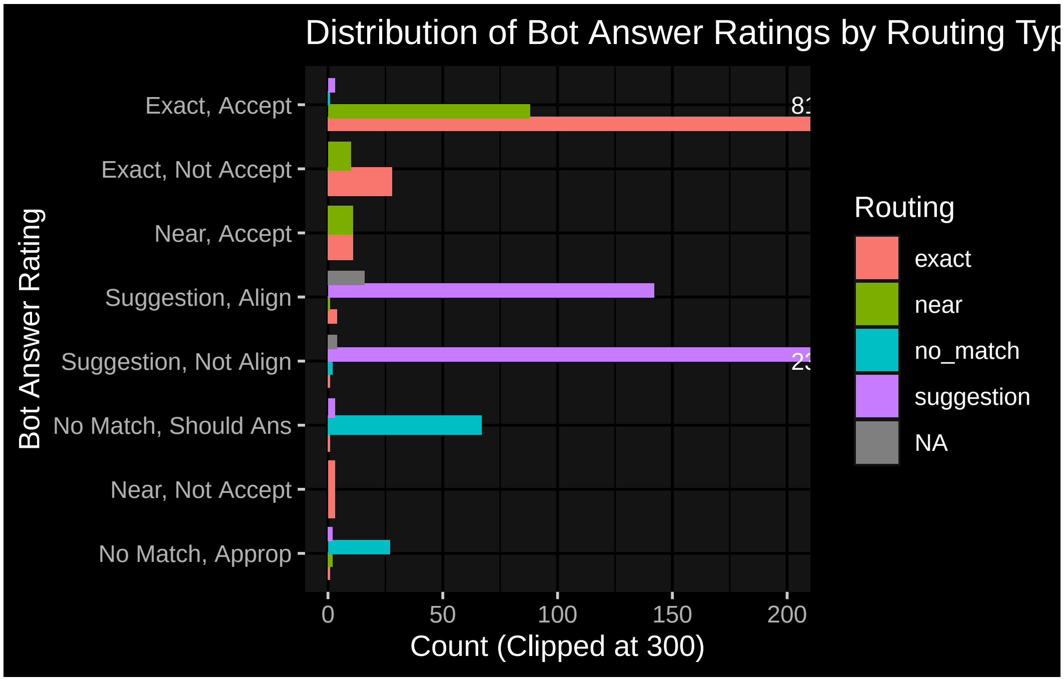
5.2 Interactions

5.2.1 Secondary Codes and Flags



5.2.2 Bot Answer Ratings




Distribution of Bot Answer Ratings by Routing Typ
Bot Answer Rating
Exact, Accept
Exact, Not Accept
Near, Accept
Suggestion, Align
Suggestion, Not Align
No Match, Should Ans
Near, Not Accept
No Match, Approp
81
23
0
50
100
150
200
Count (Clipped at 300)
Routing
exact
near
no_match
suggestion
NA

6 Appendix

6.1 Primary Codes

The primary codes were broad categories of the utterance, and were defined as follows:

Category	Code
Candidate and Campaign Information	[G]
Policies, Positions, and Propositions	[P]
Elections Logistics and Procedures	[EL]
Voter Motivation and Civic Engagement	[V]
Election Analysis and Insights	[EA]
Miscellaneous	[M]

6.2 Secondary Codes

The secondary codes were more specific sub-categories of the utterance, and were defined as follows:

Secondary Code	Description	Primary Code
[Ind Basic]	Basic Information, Individual Campaign	[G]
[Elect Basic]	Basic Information, Election	[G]
[Fed Basic]	Basic Information, Federal Races – Senate	[G]
[Pres Basic]	Basic Information, Federal Races – Presidential	[G]
[House Basic]	Basic Information, Federal Races – House	[G]
[State Basic]	Basic Information, Statewide Races	[G]
[Local Basic]	Basic Information, Local Races	[G]

Secondary Code	Description	Primary Code
[Cand Hist]	Candidate History	[G]
—	—	—
[Cand Comp]	Candidate Comparisons and Policy Differences	[P]
[Cand Pos]	Candidate Positions on Issues	[P]
[Iss Spec]	Issue-Specific	[P]
[St/Loc Prop]	State and Local Propositions	[P]
[P Oth]	Other	[P]
—	—	—
[Voter Reg]	Voter Registration	[EL]
[Mail-in]	Mail-in Ballots	[EL]
[EV]	Early Voting	[EL]
[Polling Loc/Time]	Polling Locations and Times	[EL]
[Voting Ac]	Voting Accessibility	[EL]
[Res/Rep/Time]	Results and Reporting Timelines	[EL]
[Rec/Res]	Recounts and Contested Results	[EL]
[Voting Req]	Voting/Polling Place Requirements and Rules	[EL]
[EL Oth]	Other	[EL]
—	—	—
[Reas]	Reasons for Voting, Civic Duty	[V]
—	—	—
[Polling Fore]	Polling and Election Forecasts	[EA]
[Endor/Infl]	Endorsements and Influences	[EA]
[Deb]	Debates	[EA]
—	—	—
[Non-elect]	Non-election Query	[M]
[Oth]	Other	[M]

6.3 Flags

There were 5 possible flags, as described below:

Flag Name	Description	Required?
[Trust Flag]	If user seems dubious, there could be conspiratorial thinking, or even just the answer given was misleading.	Yes (required for every row)
[Context]	If the context of the conversation is important to why we coded as we did.	Optional
[RepeatedQuestion]	If the user is repeating their question exactly. This might indicate accidental misclick, or jailbreaking attempts.	Optional
[WackAnswer]	Occasionally the bot gave very odd answers that didn't match the question at all.	Optional

6.4 Answer Ratings

The bot answer ratings were given as a two-part array, with interpretation and polarity as shown below:

Suggestion Level	Bot Answer Description	Codes	Polarity
Suggestion	Suggestions align with user query/intent	[Suggestion, Align]	Positive
	Suggestions seem to miss the point of the question	[Suggestion, Not Align]	Negative
Near	Acceptable Answer	[Near, Accept]	Positive
	Not Acceptable Answer	[Near, Not Accept]	Negative
Exact	Acceptable Answer	[Exact, Accept]	Positive
	Not Acceptable Answer	[Exact, Not Accept]	Negative
No Match	Bot should have been able to answer	[No Match, Should Ans]	Negative
	Appropriate that bot did not answer	[No Match, Approp]	Positive

6.5 AI Performance Breakdown

Table 6.5: Classification Report: Per-Class and Summary Averages

class	precision	recall	f_meas	support
EA_Deb	1.000	1.000	1.000	4
EA_Endor/Infl	1.000	1.000	1.000	16
EA_Polling Fore	0.200	0.760	0.864	25
EL_EL Oth	0.500	0.968	0.984	31
EL_EV	0.333	0.826	0.905	23
EL_Mail-in	0.200	0.954	0.977	240
EL_Oth	0.500	0.500	0.667	2
EL_Polling Loc/Time	0.500	0.909	0.952	11
EL_Res/Rep/Time	0.333	0.641	0.781	39
EL_Voter Reg	0.250	0.953	0.976	172
EL_Voting Req	0.333	0.911	0.953	56
G_Cand Hist	0.500	0.615	0.762	13
G_Elect Basic	0.333	0.753	0.859	97
G_Fed Basic	0.333	0.566	0.723	53
G_House Basic	0.200	0.760	0.864	75
G_House Basic, Fed Basic	1.000	1.000	1.000	1
G_Ind Basic	1.000	1.000	1.000	32
G_Local Basic	0.333	0.694	0.819	124
G_Pres Basic	1.000	1.000	1.000	13
G_State Basic	0.250	0.979	0.989	239
M_Non-elect	0.500	0.952	0.976	21
M_Non-Elect	0.500	0.333	0.500	3
M_Oth	0.333	0.929	0.963	28
P_Cand Comp	0.500	0.768	0.869	69
P_Cand Pos	1.000	1.000	1.000	34
P_Iss Spec	0.250	0.559	0.717	34
P_Other	1.000	1.000	1.000	2
P_P Oth	1.000	1.000	1.000	8
P_St/Loc Prop	1.000	1.000	1.000	9
V_Reas	1.000	1.000	1.000	8
macro avg	0.853	0.844	0.849	1482
micro avg	0.865	0.865	0.865	1482
weighted avg	0.869	0.865	0.864	1482