

# Towards Fair and Accurate Short VAAs: Adaptive Ordering and Imputation in Democratic Recommendation Systems

Thilo I. Dieing<sup>1,2\*</sup>, Lea Cohausz<sup>2\*</sup>

<sup>1</sup>Technical University Darmstadt

<sup>2</sup>University of Mannheim

thilo.dieing@tu-darmstadt.de, lea.cohausz@uni-mannheim.de

## Abstract

In increasingly fragmented party systems, voters face growing difficulties in making informed electoral decisions. Voting Advice Applications (VAAs), questionnaires consisting of political statements, have become central democratic tools by helping citizens compare their political preferences with those of competing parties. Yet, modern VAAs face a fundamental trade-off: while shorter questionnaires could reduce user fatigue and dropout, they also suffer from severe advice degradation because they consider fewer statements. Existing solutions using adaptive questionnaires and imputation show promise, but remain largely focused on candidate-based systems, which only make up a small fraction of the overall electoral systems, and rarely address fairness across parties and opinion groups. This paper introduces a novel adaptive framework for party-based VAAs that combines active learning with imputation trained solely on previous user responses. Rather than relying on a fixed questionnaire, our system dynamically personalizes which statements users receive, in what order, and how many are needed before reliable voting advice can be generated. Evaluating the approach on two German municipal VAA datasets from Saarbrücken and Halle (Saale), we show that imputation and adaptive item selection strategies improve recommendation quality by more than 12 percentage points compared to conventional short VAAs. Beyond performance gains, our findings reveal systematic inequalities in how short VAAs represent minority opinions and smaller parties, highlighting the importance of cross-party fairness in the design of democratic digital tools.

## 1 Introduction

Our democracies cannot survive without people making informed choices and expressing their opinions. Elections are the main participatory event for citizens and, accordingly, the decision of which political party to vote for is very important. In many countries, especially those with proportional representation, the number of parties can be large, making decision-making challenging (Riker 1986; Anckar

2000; Janega and Zymová 2025). For example, in Germany’s latest federal election, 29 parties competed; in Belgium’s latest federal election, 26 parties appeared on the ballots. Especially in countries with many parties (Dassonneville, Blais, and Dejaeghere 2015), voters often show shifting alliances, voting for different parties in different elections (Dalton 2014), reacting to parties’ political programs as well as current economic, social, and environmental events (Dassonneville 2015; Ferland and Dassonneville 2021; Dieing and Stecker 2025). Further raising the cost of voting (Blais et al. 2019), voters have to forage through large collections of information sources, such as lengthy party programs in specialized language, news reports, or debates, to make an informed decision (Metag and Gurr 2023).

For this reason, so-called Voting Advice Applications (VAAs) have emerged as heuristic tools to indicate the extent to which a party’s political program for a given election matches a person’s opinions (Garzia and Marschall 2019). VAAs consist of politically relevant statements, and respondents express their stance on these statements. The respondents’ answers are then compared to the political parties’ responses to the same items. In the end, the respondent receives a ranked list of political parties, with the party at the top having the highest agreement with the respondent’s answers and the party at the bottom having the lowest. VAAs have great popularity, especially in European countries with the proportional representation system (Marschall, Hagemann, and Bongartz 2026; Garzia and Marschall 2019).<sup>1</sup> These self-selective online tools (van de Pol et al. 2014, 2019; Dieing 2025a) not only make politics accessible to millions of users but have also been shown to have a positive impact on users’ political knowledge (Westle, Bege-mann, and Rütter 2014; Schultze 2013; Kamoen et al. 2015), increase turnout (Gemenis and Rosema 2014; Munzert and Ramirez-Ruiz 2021), and possibly influence the voting behavior of (undecided) voters (Pianzola et al. 2019; Tromborg and Albertsen 2023).

One problem developers face when constructing VAAs is that, on the one hand, they should be short enough that people do not get bored or overwhelmed with information

\*These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

<sup>0</sup>The GitHub repository with our code and results, as well as the appendix, can be found in our online repository ([https://github.com/lea-cohausz/adaptive\\_short.vaas](https://github.com/lea-cohausz/adaptive_short.vaas)).

<sup>1</sup>Germany’s most popular VAA, *Wahl-O-Mat*, registered 26 million uses for Germany’s latest federal election VAA, with Germany having 59.2 million eligible voters for this election (Bundeszentrale für politische Bildung 2025).

and drop out, which would potentially even negatively affect voter turnout. On the other hand, they should be covering enough topics to differentiate between all parties.

Consequently, recent VAA designs increasingly feature so-called rapid versions, or short VAAs, in which users can skip a large portion of statements (e.g., 50%) and go directly to the voting advice. While this design might counter user fatigue, research has shown that due to so many missing user statement answers, short VAAs present users with 60 percentage points less accurate advice (Bachmann, Sarasua, and Bernstein 2024), which could bias users' voting decisions and misinform them.

To address this inherent advice degradation problem in short VAAs, caused by missing information from unanswered statements, initial studies have proposed adaptive systems that select items based on their relevance (i.e., the accuracy with which the top party is correctly recommended) (Sigfrid 2024). Others have expanded this by adding imputation of the skipped statements to the adaptive design (Bachmann, Sarasua, and Bernstein 2024; Bachmann et al. 2026).

Despite initial promising results, several gaps remain to be addressed to overcome advice degradation in short VAAs. The majority of approaches have been tested on VAAs constructed for a completely different electoral system than prevalent in most European countries, meaning that the developed methods are not directly applicable to most VAAs (Bachmann, Sarasua, and Bernstein 2024; Bachmann et al. 2026). Methods developed for the type of VAA we are interested in were not tested on real VAA data but a combination of other questionnaires and synthetic data, limiting the meaningfulness of the results (Sigfrid 2024).

Moreover, the current evaluation of short or adaptive VAAs, which is heavily focused on the accuracy of recovering the Top-1 party, neglects the severe imbalance in how often specific parties emerge as the top-recommended party: some parties are hardly ever on top, whereas others are so frequently. This imbalance then affects the ability of ML methods to learn patterns, and, combined with the evaluation focus on Top-1 accuracy, may lead us to overestimate the performance of short and adaptive VAAs for parties who are less frequently at the top. Such a result would be problematic, as it can disadvantage more niche parties and topics that are less often represented in the VAA user population, hinder differentiation among programs, and yield less accurate results for respondents with less common opinions. Consequently, models that do not account for what we now call cross-party fairness can yield results that appear overly simplistic, omitting minority opinions. Therefore, users could receive politically biased results, thereby undermining the established political equality of electoral tools and exposing them to possible lawsuits from smaller, less represented parties (heise online 2009).

In this paper, we address and evaluate the degradation problem. Using data from two German municipal VAAs, Saarbrücken (SB) and Halle (Saale) (HA), we investigate to what extent short VAAs lead to degraded advice while considering three evaluation metrics (Top-1, Top-3 accuracy, and MAE). We explore the effect of imputation (i.e., pre-

dicting unseen item responses) on performance and evaluate cross-party fairness. Furthermore, we investigate whether full personalization of the questionnaire via active learning improves results across both overall advice quality and cross-party fairness. The core idea of our active learning approach is that the items, the order in which they are asked, and possibly even the number of items, depend entirely on respondents' answering patterns. Lastly, acknowledging that our models, because they are trained on previous user data, are affected by the cold-start problem, we conduct an additional test to determine how many initial user cases are needed for robust performance, while again accounting for possible cross-party differences.

Correspondingly, we have the following research questions:

**RQ1:** Does imputation reduce advice degradation?

**RQ2:** How do short VAA approaches compare to our active learning strategy?

**RQ3:** Do political parties that appear less frequently as the top matching party get predicted as accurately as parties that appear more frequently on top?

**RQ4:** How many fully answered questionnaires do we need as training data to predict the top parties accurately with our active learning approach?

To answer our research questions, we continue as follows:

- We start by providing additional background on VAAs, short VAAs, and existing approaches to adapting VAAs in Section 2. We also explain how we define the requirements of short VAAs and formalize the setup.
- In Section 3, we explain how we evaluate our research questions, describing the experiments, our active learning strategy, and the used machine learning (ML) methods.
- Subsequently, in Section 5, we discuss our results. We observe that short VAAs without any additional imputation lead to a significant advice degradation, and that by imputing statements and adapting them for each user, we can significantly reduce degradation while still leaving some users with incorrect advice. Our analysis of cross-party fairness further demonstrated that short VAAs also treat minority opinions unequally to majority opinions. In comparison, our proposed model is able to weigh more equally, although inequality between majority and minority opinions remains. Considering the cold start problem, similar to other studies that use ML techniques trained on VAA statement data (Katakis et al. 2014; Bachmann et al. 2025), we observe that after the first couple of hundred users, our models reach their peak performance, while minority opinions are impacted significantly more by decreasing numbers in the train set, something researchers in future investigations should be aware of.
- In Section 6, we derive recommendations for researchers and developers and also highlight limitations of our study.

This study contributes to the ongoing improvement of Voting Advice Applications by introducing models that can counteract advice degradation in short VAAs aimed at addressing user fatigue, making VAAs more accessible. The

investigation shows that adaptive sorting of statements and imputation of the remaining statements can significantly outperform the regular short VAA. Furthermore, we demonstrate that tools are greatly affected by imbalanced opinion distributions, leading to short VAAs that treat minority opinions unequally to majority opinions, thereby damaging the neutrality of electoral tools. As such, we not only showcase a proof of concept for improving short VAAs but also demonstrate that researchers should be aware of the imbalance when creating their models to account for cross-party fairness, so that VAAs remain the essential, neutral, and fair tools used by millions of citizens.

## 2 Background

### VAAs and their Architecture

First introduced as pen-and-paper tests in schools (De Graaf 2010), VAAs have evolved into worldwide heuristic tools to inform users about the political landscape, with millions of users each election (Marschall, Hagemann, and Bongartz 2026; Garzia and da Silva 2026). In their basic structure, VAAs compare user and party answers to politically relevant statements to derive a final voting advice, ranking parties by their agreement with a user. As self-selective online tools, VAAs cater to specific user types that are mostly politically interested and use VAAs to check whether their preexisting party choices are reaffirmed (van de Pol et al. 2014, 2019; Dieing 2025a). Aside from such users, there are also so-called “Seekers” who use the tools, driven by high uncertainty and lesser political sophistication, to find out whom to vote for and are therefore especially influenced by the VAA’s voting advice (Tromborg and Albertsen 2023). Overall, studies show that VAAs have a positive impact on users’ political knowledge (Westle, Begemann, and Rütter 2014; Schultze 2013; Kamoen et al. 2015), while tool usage can also increase overall interest in politics (Fivaz and Nadig 2010; Manavopoulos et al. 2018) and the campaign (Heinsohn et al. 2019). Other studies show that VAAs can increase turnout and users’ willingness to cast a vote (Gemenis and Rosema 2014; Munzert and Ramirez-Ruiz 2021) and can potentially (Munzert and Ramirez-Ruiz 2021; Enyedi 2015) influence the voting behavior of (undecided) voters (Pianzola et al. 2019; Tromborg and Albertsen 2023; Germann, Mendez, and Gemenis 2023).

Based on the premise that AI and machine learning “will help ensure that VAAs will continue to contribute positively to the democratic processes” (Gemenis 2024, p. 12), research has utilized, computational and ML approaches in the VAA domain to predict actual voting intentions from VAA data (Mendez 2017; Tsapatsoulis and Agathokleous 2017), analyze party positions in depth (Dieing and Siefken 2025), create dynamic party profiles (Terán and Mancera 2019), guardrail (Dieing 2025b) and implement conversational VAAs through chatbots (Kamoen and Liebrecht 2022; Liebrecht, Kamoen, and Aerts 2023; Dieing 2024).

From a design perspective, developers use different methods to match users to parties/candidates. Traditionally, VAAs compare user and party positions on relevant statements and derive overlapping matches via either low-

dimensional (e.g., Euclidean distance) or high-dimensional scoring techniques (Mendez 2017). The more commonly used high-dimensional methods, building on Manhattan distance, are Proximity, Scalar/Directional, or Hybrid (Moreno, Padilla, and Chueca 2022; Mendez 2017).

### Short VAAs and User Fatigue

In general, VAAs can be considered as long political questionnaires (van Camp, Lefevere, and Walgrave 2014) or democratic recommender systems (Berdoz et al. 2025) and include roughly 30 statements (in some cases, it can go above 70 (Bachmann, Sarasua, and Bernstein 2024; van Camp, Lefevere, and Walgrave 2014)), which are by default sorted either by the developers or randomly. Similarly to questionnaires, VAA users may experience time constraints or user fatigue (Bachmann et al. 2026) and may not finish the VAA, meaning they will not receive voting advice, thereby reducing the positive effects the VAA can have. While some VAAs already allow skipping a single statement, recent developments, as mentioned in the introduction, introduced a more drastic design change called the rapid version, or short VAA. As a reminder, these short VAAs allow users to skip the remaining statements, shortening the VAA and arriving faster at the voting advice. Examples using this design are the Swiss Smartvote with normally 75 statements and a short VAA possibility to skip after 31 statements, the STANDARD-Wahlkompass, where users can skip after 40% of the statements ( $N = 12$ ), or the VAA development tool VOTO that allows users to skip after answering half of the statements.<sup>2</sup>

While this feature may help mitigate user fatigue, recent studies show that the short VAA design comes at the cost of voting advice degradation, with short VAAs achieving only 40% accuracy (Bachmann, Sarasua, and Bernstein 2024) compared to the voting advice that would be given if all questions were answered. This advice degradation due to the reduced number of statements answered by a user could have negative implications for the otherwise positive effect of the VAA, by misinforming users or even changing their vote choice. One idea proposed by scholars to counter the advice degradation is to make the VAA statement ordering adaptive, and in doing so, ask users the statements first that are most impactful to their final voting advice (Sigfrid 2024; Bachmann, Sarasua, and Bernstein 2024). With this methodology, users could use short VAAs while still receiving less degraded advice, since the most important items could be included in the advice calculation. Building such an adaptive user-responsive design, Sigfrid (2024) demonstrated that the applied method improved the top-party-match accuracy of the short VAA by roughly 10 percentage points. However, the test was performed on simulated and survey data, not on real data from a party-based VAA.

Extending the approach Bachmann, Sarasua, and Bernstein (2024) demonstrated that combining adaptive question ordering with imputation of skipped statements could fur-

<sup>2</sup>VOTO: <https://www.voto.vote/>; STANDARD-Wahlkompass: <https://www.derstandard.de/story/3000000227558/der-standard-wahlkompass-zur-nationalratswahl> (last visited 22.04.2026)

ther improve the voting advice provided by a candidate-based short VAA. Further studies focusing on candidate-based VAAs demonstrate how large language models can be useful for generating additional training cases when few candidates are available (Bachmann et al. 2025) and how to incorporate advice certainty into the VAA design (Bachmann et al. 2026).

## Contribution

Despite significant improvements in short VAAs reported by the discussed studies, several gaps remain to be addressed. Most methods have been applied to candidate-based VAAs (cVAAs). cVAAs compare users to candidate positions in electoral systems such as Switzerland’s open-list proportional representation, with a large number of candidates up for vote (Bachmann, Sarasua, and Bernstein 2024; Bachmann et al. 2026). These VAAs, however, differ significantly from the more often used party-based VAAs (pVAAs) while at the same time being relevant for a much smaller number of democracies. As such, while models can be trained on a large number of candidates in cVAAs, pVAAs often include far fewer parties (e.g., 29 in the 2025 Wahl-O-Mat), making the generated data too small for substantive ML training. Additionally, scoring is also affected by the choice between candidates and parties, as Bachmann et al. (2026, p. 17) have noted: “there are no ‘sets’ of candidates but ranked lists with recommended parties”. While some of these points have already been adapted by Sigfrid (2024), but not on real pVAA data, reliable imputation methods of the remaining statements and their performance with adaptive statement ordering have so far not been tested on pVAAs. Additionally, there is no established standard to reliably evaluate trained models in a pVAA context. As such, Sigfrid (2024) only reports the accuracy of recovering the top party match results, thereby offering a limited assessment of performance.

Lastly, while VAA users are distinct in their characteristics from the general population (van de Pol et al. 2014, 2019; Dieing 2025a)(even though some characteristics are diffusing (Albertsen 2022)), the users as a group are not uniform and instead demonstrate heterogeneity in many characteristics, including their voting behavior (Stecker et al. 2025). This means that during training, ML developers must keep in mind that opinions and party preferences are not evenly distributed. Hence, as already highlighted in the introduction, some parties may appear more often in the higher ranks of the voting advice because more people align with their political positions, while others may appear less frequently in higher ranks because VAA users align less with their political views. As such, ML models may ultimately overfit to the majority of user opinions and become better at predicting parties in a short VAA setting that are more often ranked highly. Therefore, applying only averaged metrics, such as the overall Top-1 accuracy, can mask whether models are fit for all opinions or only the majority. We call this requirement for equal representation cross-party fairness, meaning that ML tools trained on VAA data must account for political heterogeneity and treat all opinions, majority and minority alike, to remain politically neutral. Should short VAAs utilize models that do not account for

cross-party fairness, minority opinions would be at a disadvantage, and VAAs could become politically biased and lose trustworthiness. Despite this importance, the current literature on short VAAs has yet to address this issue.

## Setting

In this paper, in contrast to most previously discussed works, we focus on the classical VAA setting of party VAAs rather than candidate VAAs. We define the components of this setting as follows:

- We have  $n$  respondents  $R = \{r_1, \dots, r_n\}$  and  $m$  items (questions/statements)  $Q = \{q_1, \dots, q_m\}$ .
- For each respondent  $r_i \in R$ , we denote their answer to item  $q_k$  as  $a_{i,k} \in \mathcal{A}$ , where  $\mathcal{A}$  is the set of admissible answers. The full answer vector of respondent  $r_i$  is  $\mathbf{a}_i = (a_{i,1}, \dots, a_{i,m}) \in \mathcal{A}^m$ .
- We have  $l$  parties  $P = \{p_1, \dots, p_l\}$ . Analogously, for each party  $p_j \in P$ , we denote their answer to item  $q_k$  as  $pa_{j,k} \in \mathcal{A}$ , with full answer vector  $\mathbf{pa}_j = (pa_{j,1}, \dots, pa_{j,m}) \in \mathcal{A}^m$ .
- The matching score between respondent  $r_i$  and party  $p_j$  is determined by a deterministic scoring function  $f: \mathcal{A}^m \times \mathcal{A}^m \rightarrow R$ , yielding  $s_{i,j} = f(\mathbf{a}_i, \mathbf{pa}_j)$ .
- For each respondent  $r_i$ , the full vector of matching scores across all parties is  $\mathbf{s}_i = (s_{i,1}, \dots, s_{i,l})$ . Based on  $\mathbf{s}_i$ , the respondent receives an ordered ranking of parties  $\sigma_i: \{1, \dots, l\} \rightarrow P$ , where  $\sigma_i(1)$  is the best-matching party, i.e.,  $s_{i,\sigma_i(1)} \geq s_{i,\sigma_i(k)}$  for all  $k \in \{1, \dots, l\}$ , and parties are ranked in descending order of their matching score.

## Requirements

What follows from existing research and the core ideas of VAAs is that a shortening of VAAs must meet certain requirements. First, shortening the VAA should still lead to the same results, i.e., the same ordering of parties, as if a person had answered the full questionnaire. This is a requirement that was recognized by all research on short VAAs (Sigfrid 2024; Bachmann, Sarasua, and Bernstein 2024). We therefore consider voting advice given based on the full set of statements as ground truth, while acknowledging that this voting advice might not be ideal for all users. Second, all political parties and all users should be represented by the short VAA as well as by the full VAA, meaning that it should not be the case, e.g., that parties that are only rarely the best-matching ones are predicted much less accurately. We argue that, especially, this cross-party fairness is not covered or discussed by the other strategies presented so far and has not yet been evaluated. Beyond these requirements regarding quality, we also need to ensure that our strategy does not imply a large number of completed full VAAs for training our models. While we can assume that some respondents willingly complete the full set of items, we will still encounter problems if our strategy requires substantial training data from other respondents.

### 3 Data and Evaluation

#### Data

The proposed strategy for a personalized adaptive short VAA was tested based on an original dataset of German VAAs for municipal elections in June of 2024 (Stecker et al. 2025). For this study, two VAAs from this dataset were selected: the VAAs for the elections in Halle (Saale) and Saarbrücken, hereafter referred to as HA and SB. The VAAs were built by political scientists using the VOTO software,<sup>3</sup> offering users a 5-point Likert scale (disagreement–agreement) to answer randomly ordered statements. Party match scores were calculated based on the Hybrid scoring technique (Moreno, Padilla, and Chueca 2022; Stecker et al. 2025). Both HA and SB were online from mid-May 2024. HA included  $m = 37$  statements and had  $n = 24,289$  users, while SB also had  $m = 37$  statements and  $n = 5,224$  users. Both VAAs included parties from all ideological sides, with HA including  $l = 11$  parties (AfD (31.7%) appearing most often in the advice for first place, while Die Partei (0.6%) appearing the least often among all first-place recommendations) and SB including  $l = 6$  parties (FDP (27.6%) being most often in the first rank and the Linke (6.0%) being the least often). The statements’ topics in both VAAs addressed issues on “traffic”, “immigration”, “equality”, “education”, and others, covering the most relevant topics for each election, with additional saliency checks (for relevance and party-disagreeing statements) before going live. Both tools were chosen because they have similar numbers of statements, enabling more valid feature comparisons. Additionally, HA and SB differ strongly based on user and party numbers as well as the party advice distribution, making them ideal for cross-party fairness testing. Hence, testing our ML strategy on both VAA datasets can assess generalizability and potential biases from user or party counts, which short VAA developers should be aware of.

#### Evaluation Metrics

To evaluate our proposed model configurations, we use three metrics to provide more robust performance reporting. Acknowledging that users are often most interested in the top party match and that, due to the VAA’s interface emphasizing a ranked list, the top-ranked party becomes the focal point of attention by design, we evaluate models based on Top-1 accuracy, i.e., whether the party that had the top rank in a users voting advice can be successfully recovered, which has also been utilized by Sigfrid (2024). Using Top-1 accuracy alone, however, can be misleading, since parties can share the same rank or be very close to each other based on their matching score, limiting the usefulness of exact ranks. We also include the Top-3 accuracy, which we consider a harder metric, since it measures whether the mod-

<sup>3</sup>Note that at the time, VOTO did not offer a short VAA design. For this reason, we simulate the short VAA as VOTO implemented it in 2025, where users can skip half of the statements. In this version, statements are also randomly ordered. Additionally, acknowledging that cut-off can vary from tool to tool and also be based on developer decisions, we test different cut-off dates as well, which is reported in the appendix.

els lead to the correct party matches for the first three ranks (however, regardless of the exact position), not just the first rank as Top-1 accuracy.<sup>4</sup> As a third metric, we move beyond rank-based evaluation to matching score evaluation. As a result, we use the mean absolute error (MAE) to measure how much a party’s matching score based on our models deviates from the actual score, and we take the mean across all parties. Including this metric enables more fine-grained evaluation, free from rank-sharing bias. We also considered including Spearman’s correlation, as recommended by Bachmann et al. (2026), for party VAAs. However, we concluded that this measure cannot be reliably used because it cannot account for parties with the same rank.

### 4 Strategies

#### Regular Short VAAs: Random Item Selection

The core strategy we evaluate is the regular short VAA, as it is used by tools such as VOTO (recall that our data is from VOTO). In accordance with VOTO’s implementation, we randomly select 50% of the overall items and reveal the answers to these items. Based on the answers, we compute the party match scores  $s_{i,j} = f(\mathbf{a}_i, \mathbf{p}_{a_j})$  for all parties  $p_j \in P$  and then evaluate the performance metrics. Crucially, the 50% subset is drawn independently and at random for each respondent  $r_i$ .

#### Active Learning

We compare the results of the regular short VAA to a fully personalized strategy based on active learning. The core idea is as follows: for each respondent  $r_i$ , the items to be asked are selected in a completely personalized way based on that respondent’s answers so far. The underlying assumption is that an answer to one item, or a collection of items, is informative about the likely answers to other items. Hence, rather than asking items whose answers we can already predict with reasonable confidence, we prioritize items for which we are most uncertain. Observing such items is then also expected to be most informative about the remaining unanswered items.

Concretely, we implement the active learning strategy as follows. We assume a training dataset of fully completed VAAs. For each new respondent  $r_i$ , we begin by revealing one randomly selected item. At each subsequent step, let  $O \subset Q$  denote the set of items already answered by  $r_i$ , with observed answer vector  $\mathbf{a}_i^O = (a_{i,k})_{k \in O}$ . For each unasked item  $q_k \in Q \setminus O$ , we fit a separate supervised model on the training data, using  $O$  as the feature set and  $q_k$  as the target variable. The model thus predicts the answer  $a_{i,k}$  from  $\mathbf{a}_i^O$ . Based on these per-item predictions, we select the next item to ask according to one of several selection criteria (see below). We iterate—fitting models for all unasked items, selecting the next item—until a stopping criterion is reached.

Although our active learning approach serves more as a proof of concept than as a fully optimized short VAA, we

<sup>4</sup>A party can only be the correct Top-1 match if it is the only party ranked on top. Top-3 is handled equivalently.

explore several design choices: the ML model used for prediction, the item selection criterion, and the stopping criterion.

**ML Models** Since a separate model must be fitted for each unasked item at each step, we require methods that are computationally efficient. We compare Ridge Regression, Naïve Bayes (both using default hyperparameters), and a relatively shallow Random Forest. Fitted models are cached across respondents whenever the observed feature set  $O$  is identical, which substantially reduces redundant computation in practice. Still, we can only use methods that allow for fast training of models, limiting candidate ML methods.

**Item Selection Criteria** Let  $\hat{p}_{i,k}(v)$  denote the predicted probability that respondent  $r_i$  answers item  $q_k$  with response value  $v \in \mathcal{A}$ , as given by the fitted model. We evaluate the following selection criteria:

- **Random.** The next item is selected uniformly at random from  $Q \setminus O$ . This serves as a baseline and is, in the case when 50% of items are answered, identical to the regular short VAA.
- **Uncertainty.** We select the item  $q_k \in Q \setminus O$  for which the model is least certain about its prediction, measured as  $q^* = \arg\max_{q_k \in Q \setminus O} (1 - \max_{v \in \mathcal{A}} \hat{p}_{i,k}(v))$ .
- **Entropy.** We select the item with the highest Shannon entropy of the predicted class distribution:  $q^* = \arg\max_{q_k \in Q \setminus O} H(\hat{p}_{i,k}) = \arg\max_{q_k \in Q \setminus O} (-\sum_{v \in \mathcal{A}} \hat{p}_{i,k}(v) \log \hat{p}_{i,k}(v))$ . Note that uncertainty and entropy both treat all items as equally predictable; they differ only in how they aggregate the predicted class probabilities.
- **Variance Reduction.** We select the item expected to most reduce the variance of the observed-only party match scores across parties. Let  $\hat{s}_{i,j}^{O \cup \{q_k=v\}}$  denote the match score for party  $p_j$  computed from the observed items plus the hypothetical answer  $v$  to item  $q_k$ . We select  $q^* = \arg\max_{q_k \in Q \setminus O} (\text{Var}_j(\hat{s}_{i,j}^{O \cup \{q_k=v\}}) - \sum_{v \in \mathcal{A}} \hat{p}_{i,k}(v) \text{Var}_j(\hat{s}_{i,j}^{O \cup \{q_k=v\}}))$ , where  $\text{Var}_j$  denotes the variance over parties  $p_j \in P$ .
- **Top-1 Change.** We select the item for which observing the answer is most likely to change the currently top-ranked party. Let  $\hat{p}_1^O = p_j$  denote the current top-1 party under the observed scores. We select  $q^* = \arg\max_{q_k \in Q \setminus O} \sum_{v \in \mathcal{A}} \hat{p}_{i,k}(v) \mathbf{1}_{[p_j \hat{s}_{i,j}^{O \cup \{q_k=v\}} \neq \hat{p}_1^O]}$ .
- **Top-3 Change.** Analogous to top-1 change, but targeting the set of the three highest-ranked parties. Let  $T_3^O$  denote the current top-3 set under the observed scores. We select  $q^* = \arg\max_{q_k \in Q \setminus O} \sum_{v \in \mathcal{A}} \hat{p}_{i,k}(v) \mathbf{1}_{[T_3^{O \cup \{q_k=v\}} \neq T_3^O]}$ .

In all cases, ties in item scores are broken uniformly at random.

**Stopping Criteria** We evaluate two stopping criteria. First, a fixed 50% cut-off, consistent with regular short VAAs, which stops after  $\lceil m/2 \rceil$  items and allows direct comparison. Second, adaptive stopping conditions in

which we continue asking items until either the observed-only top-1 party  $\sigma_i(1)$  or the observed-only top-3 set  $\{\sigma_i(1), \sigma_i(2), \sigma_i(3)\}$  matches the ground truth (i.e., the ranking derived from all  $m$  items). If neither condition is reached, the simulation is capped at all  $m$  items. These adaptive conditions are evaluated separately. Note that in practice, such stopping criteria would require knowing the ground-truth ranking, which is not available; we include them to assess how quickly the correct recommendation can, in principle, be recovered.

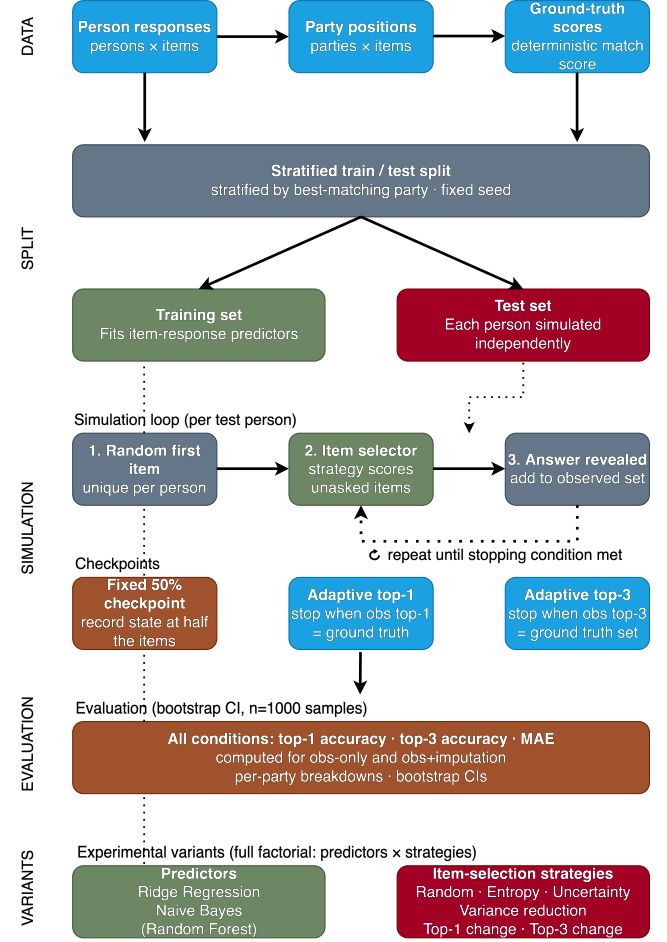


Figure 1: The pipeline for our experiments.

## Imputation

As stated in **RQ1**, we are interested in the effect of imputation on recommendation quality. Both the regular and the adaptive short VAAs are therefore evaluated in two variants: without imputation, using only the observed answers  $(a_{i,k})_{k \in O}$  to compute match scores, and with imputation, where the predicted answers  $\hat{a}_{i,k}$  for all  $q_k \in Q \setminus O$  are substituted to obtain a complete answer vector before computing scores. We use the same ML models as for item selection, i.e., Ridge Regression, Naïve Bayes, and Random Forest, to generate the imputed values.

## Pipeline

The final pipeline for our experiments can be seen in Figure 1. The pipeline is applied once to SB and once to HA, respectively. The user data is split into a train and test set (80/20 for HA; 60/40 for SB).<sup>5</sup> To account for cross-party fairness, the split is stratified by the top party match (if rank sharing occurred, a random selection was used). Doing so ensures that the models view proportional opinions during training, making them more aware of majority and minority opinions, whereas a non-stratified split biases the models from the start by failing to account for cross-party fairness. In the training, a random first item is selected for each person. For the regular short VAA, all other items until the stopping criteria are reached are also selected randomly. Based on this, we trained the ML models.<sup>6</sup> To select the next item, we tested the different selection strategies described above. After the stopping criterion was reached, performance was immediately recorded in the non-imputed (observed-only) setting. For the combination with imputation, the three ML models were first trained on the seen items to predict the unseen items, and performance was recorded at that time. We use bootstrap confidence intervals ( $n = 1000$ ) for better evaluation, accounting for sampling variability of the test set.

## 5 Results

### Performance Comparison

**Importance of Imputation** The initial comparison between imputing missing statements and not doing so (**RQ1**) shows that imputing the missing 50% of VAA statements can significantly improve model performance. Table 1 shows that, averaged over all settings, the imputation method increases Top-1 and Top-3 accuracy and decreases MAE significantly. Especially for the Top-3 Accuracy rating, i.e., whether the first three parties are correctly recovered, the observation-only results exhibit a big drop, showcasing that this harder problem benefits the most from imputation. The varying behaviors of the different performance metrics further showcase the importance of reporting more than one metric to get a more reliable performance estimation. In general, for both SB and HA, **RQ1** can be answered by stating that imputation clearly improves results. This result reconfirms the finding by Bachmann, Sarasua, and Bernstein (2024), demonstrating that imputations not only improve short VAAs in the candidate context but also in the party context when trained on user data.

**Adaptive Strategies** Our comparison of ML methods (see Appendix) showed that Ridge Regression generally performs better than both Naïve Bayes and Random Forest,

<sup>5</sup>We chose the larger test set percentage for SB to get more precise estimates of the results. The HA dataset is so large that 20% is sufficient to obtain accurate estimates. As our experiments also show, the decrease in training data for SB does not negatively affect results.

<sup>6</sup>Note the Random Forest was only tested on SB, as the large test set in HA increases the computational effort and time too much, and as the results of SB already indicated a better performance of Ridge Regression at least for the SB dataset.

although confidence intervals sometimes overlap. Because of this, Ridge Regression is chosen as the best-performing model, and Table 2 shows only its results. Generally, Table 2 shows the different selection strategies for adaptive systems in detail (answering **RQ2**), again comparing based on the 50% cut-offs to make the comparison fair. Due to our answer to **RQ1**, we only report the results of the different non-random selection strategies with imputation. The table additionally reports the baseline performance of just using the regular short VAA (i.e., only using the randomly selected 50% of items), to still have the comparison against the actual implementation of short VAAs developed by, e.g., VOTO, as well as adding imputation of missing items to the regular short VAA (short VAA + imp). Looking first at the baseline, the performance confirms the existence of the advice-degradation problem in short VAAs: 45.6% of users in SB and 64.4% of users in HA would not receive accurate voting advice for the Top-3 ranked parties. For both datasets, the baseline has an average MAE above 3.6, indicating that, on average, the advice deviates from the advice based on all statements by 3.6 units. By adding Ridge Regression-based imputation to the baseline, the performance can be increased to an MAE of around 2.7. Despite that, the adaptive plus imputation system proposed in this paper can significantly outperform the baseline and, except for the variance reduction strategy, also the baseline with imputation, while some strategies display overlapping confidence intervals. In SB, using entropy as an item selection criterion in combination with imputation achieves the highest Top-1 accuracy and lowest MAE, while achieving the second highest Top-3 Accuracy. Hence, using entropy to select statements results in a 1-unit smaller deviation from the original party match score compared to the short VAA baseline. Nevertheless, 35.4% of users still do not receive fully accurate Top-3 party matching, while only 27.1% do not receive accurate Top-1 party advice. For HA, Top-1 change is the best-performing item selection strategy, leading to only 27% of users receiving a wrong first-ranked party recommendation. Entropy, on the other hand, performs less well, while uncertainty and Top-3 change are also good performance selection strategies. The reasons for these differences between the two datasets could stem from differences in party numbers and statements, and this highlights that there is no one-size-fits-all strategy. Overall, however, the results demonstrate that our active learning strategy for an adaptive short VAA with additional imputation, based on Ridge Regression, outperforms a regular short VAA and helps counter advice degradation in short VAAs, providing users with more accurate advice. Hence, **RQ2** can be answered by saying that the regular short VAA performs worse (even with imputation) compared to at least the best-performing fully adaptive strategy.

### Cross-Party Fairness

While the initial test shows that our ML approach can improve short VAAs, we argue that cross-party fairness also needs to be evaluated and tested, since good performance metrics do not necessarily imply that opinions that appear less often are accurately learned and represented. While we already aim to account for cross-party fairness by using a

| Dataset | Condition          | Top-1 Accuracy           | Top-3 Accuracy           | MAE (avg.)                  |
|---------|--------------------|--------------------------|--------------------------|-----------------------------|
| SB      | Observed only      | 64.4 [63.8, 65.1]        | 49.3 [48.7, 50.0]        | 4.464 [4.438, 4.492]        |
|         | Observed + imputed | <b>67.5</b> [66.9, 68.0] | <b>61.4</b> [60.8, 62.0] | <b>4.042</b> [4.008, 4.074] |
| HA      | Observed only      | 60.2 [59.8, 60.6]        | 36.0 [35.6, 36.4]        | 4.792 [4.772, 4.814]        |
|         | Observed + imputed | <b>65.2</b> [64.8, 65.6] | <b>47.3</b> [46.9, 47.7] | <b>4.060</b> [4.038, 4.082] |

Table 1: Performance for observed only versus observed plus imputed across all settings with 50% stopping (regular short VAA setting). Bootstrap confidence intervals reported in brackets. Best scores are bold. For Top-1 and Top-3 Accuracy, higher is better. For MAE, lower is better.

| Dataset | Selection method   | Top-1 Accuracy           | Top-3 Accuracy           | MAE (avg.)                  |
|---------|--------------------|--------------------------|--------------------------|-----------------------------|
| SB      | Entropy            | <b>72.9</b> [71.0, 74.7] | 64.6 [62.4, 66.7]        | <b>2.661</b> [2.596, 2.728] |
|         | Top-1 change       | <u>71.7</u> [69.8, 73.7] | 62.9 [60.8, 64.9]        | 2.898 [2.834, 2.963]        |
|         | Uncertainty        | 69.0 [67.0, 71.1]        | 62.2 [60.0, 64.2]        | <u>2.717</u> [2.663, 2.770] |
|         | Top-3 change       | 68.9 [66.9, 70.9]        | <b>66.3</b> [64.3, 68.3] | 2.908 [2.846, 2.981]        |
|         | Variance reduction | 66.1 [64.0, 68.2]        | 58.5 [56.3, 60.6]        | 3.125 [3.045, 3.218]        |
|         | Short VAA + imp    | 67.9 [65.8, 70.0]        | 63.9 [61.9, 66.2]        | 2.789 [2.734, 2.847]        |
|         | Short VAA          | 64.2 [62.1, 66.2]        | 54.4 [52.2, 56.6]        | 3.690 [3.625, 3.754]        |
| HA      | Entropy            | 65.2 [63.8, 66.6]        | 48.7 [47.3, 50.0]        | 2.916 [2.863, 2.966]        |
|         | Top-1 change       | <b>73.0</b> [71.8, 74.1] | <u>50.2</u> [48.7, 51.6] | 2.783 [2.745, 2.824]        |
|         | Uncertainty        | 65.6 [64.3, 66.9]        | 47.1 [45.7, 48.5]        | <b>2.737</b> [2.700, 2.776] |
|         | Top-3 change       | <u>69.6</u> [68.3, 70.9] | <b>53.2</b> [51.7, 54.7] | 2.829 [2.792, 2.870]        |
|         | Variance reduction | 63.2 [61.9, 64.6]        | 44.7 [43.2, 46.1]        | 3.456 [3.386, 3.525]        |
|         | Short VAA + imp    | 64.4 [63.1, 65.7]        | 45.7 [44.2, 47.1]        | <u>2.771</u> [2.734, 2.809] |
|         | Short VAA          | 57.4 [56.0, 58.7]        | 35.6 [34.2, 36.9]        | 3.897 [3.855, 3.939]        |

Table 2: Results for the different selection strategies across all settings with 50% stopping (regular short VAA setting) and observed + imputed using Ridge Regression. Bootstrap confidence intervals reported in brackets. Best scores are bold, and second-best scores are underlined. For Top-1 and Top-3 Accuracy, higher is better. For MAE, lower is better.

stratified split that correctly distributes top-ranked parties during model training, we still need to test how fair the baseline model (i.e., the regular short VAA) and the best model with an adaptive statement order and imputation are. For this, we plot, in Figure 2, the Top-1 accuracy for both the baseline with imputation and our best item selection strategy with imputation, grouped by each HA party in ascending order of how often this party was ranked on top.<sup>7</sup>

At first sight, Figure 2 shows that the two ideological extremes AfD (extreme right) and Linke (extreme left) were recommended the most often in the HA VAA. This is also roughly in line with the actual voting behavior in the HA election, with the AfD receiving the majority of votes with 21.2% (Stadt Halle (Saale) 2024). In general, the Top-1 curve shows that for both the baseline and the best model, Top-1 accuracy decreases as a party appears less frequently at the top of the voting advice. Nevertheless, this effect is steeper for the baseline than for the best adaptive strategy. Hence, to answer **RQ3**, the less often a party appears on the first rank, i.e., the fewer users have opinions that match such parties, the less accurate the advice in short VAAs becomes. This means that minority opinions are not as well represented in short VAAs, leading not only to advice degra-

dation but also to unfair party treatment, biasing short VAA advice further. Our best model, while exhibiting similar behavior, nevertheless predicts more evenly across the parties, indicating that the stratified split, as well as our adaptive design, also benefits minority opinions. Therefore, we find that methods that aim to address advice degradation in short VAAs must account for cross-party fairness to avoid biasing models trained on imbalanced party preferences, leading specific users to receive less accurate advice. Another reason both AfD and Linke are more often correctly ranked might be their positional uniqueness: no other party shares their political stances, making it easier to differentiate them from others. This would also explain why parties that are more ideologically close, like Volt and the Grüne (Greens), are rated roughly in the same quality, but not as good as other parties, because it is harder to distinguish between them. As a result, we can demonstrate how important it is for VAA developers to choose their statements so that parties are portrayed more uniquely and can be better distinguished from each other, thereby increasing VAA statement salience.

### Cold-Start Problem

As our proposed methodology relies on fully answered VAA user observations for training, we are—similar to candidate-based systems (Bachmann et al. 2025)—impacted by the cold-start problem, i.e., cases when too little data is available for effective training. Because of this, we investigate the ef-

<sup>7</sup>As the relationships found in this investigation are similar between SB and HA, and HA offers more parties, we only report the cross-party fairness of HA at this point and refer to the Appendix for our investigation of SB.

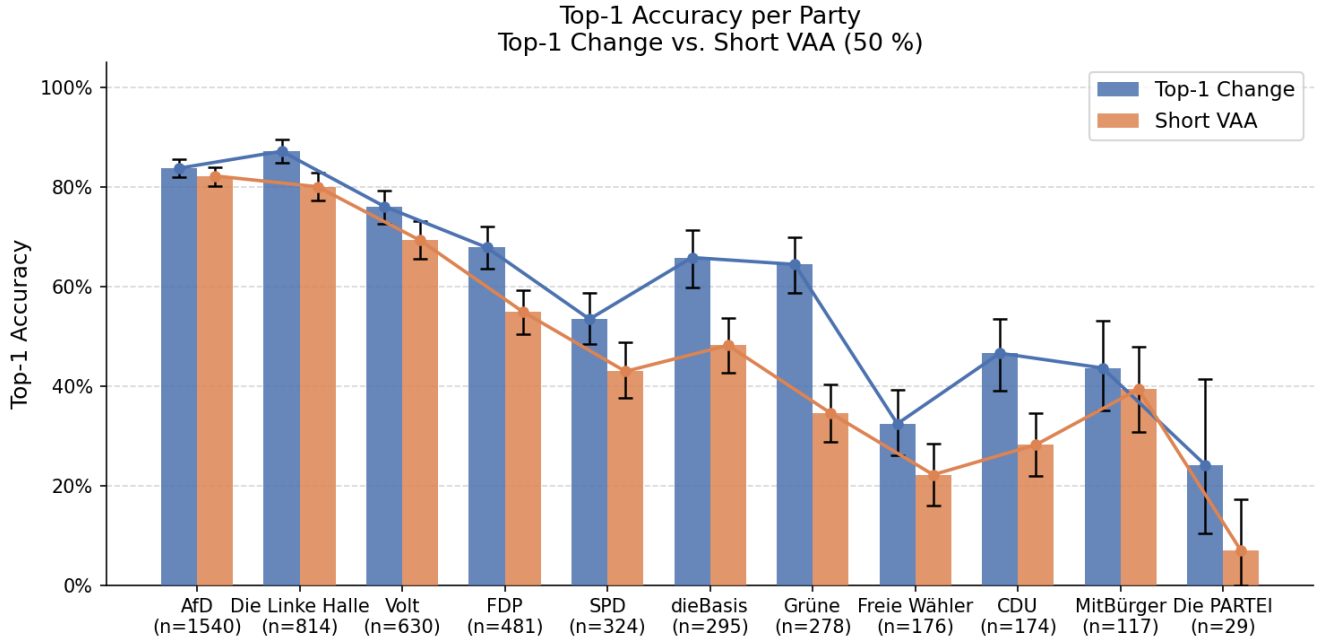


Figure 2: The HA Top-1-accuracy from the best-performing selection strategy (top-1 change) vs the short VAA (random strategy) for each party, sorted by the number of people having this party as the ground-truth top result. Both methods with imputation of missing items after 50% stopping.

fects of reducing the training size on the model performance in this section. This test will help communicate how many user inputs our adaptive models require to be used in short VAAs. To limit computational effort, we conduct this test on the smaller SB dataset—since the performance in SB and HA is similar. On the SB data, we then reduce the train set size incrementally while keeping the test set as is, narrowing the steps to smaller sizes. Figure 3 shows the performance when the train set is reduced for all three metrics, grouped by each party, to also test whether reducing size also impacts general cross-party fairness. Similar to HA, the models trained using SB also show performance differences between parties depending on how popular their opinions were among the users. As a result, the left party Linke can be predicted the least well, while the FDP and bunt.saar can be predicted the best, with Top-1 accuracies above 75%. Looking at the model accuracies as the train set size decreases shows that performance is relatively stable until the train set size drops below 500 users. This means that our models require only a small amount of user data to reach peak performance, exhibiting behavior similar to that reported in a candidate VAA setting (Bachmann et al. 2025) and for other machine learning models trained on VAA statement data, in different contexts (Katakis et al. 2014), giving an answer to **RQ4**. Differentiating between the parties additionally demonstrates that accuracies remain stable for parties that appear often in the top rank, while those that appear less often in the top rank exhibit greater variation with decreasing training sizes, which is sensible, as the parties’ already sparse representation is reduced even further. Therefore, smaller training

sizes influence these minority opinion parties more drastically than those represented more. Hence, developers need to explore new ways to build cross-party fairness in short VAA applications, especially when training data is scarce, to enable politically neutral election tools. Another conceptualization of the cold start problem would not be limited to the size of the initial training data, but rather the level of input we have from a specific user  $r_i$  we want to recommend the next statement to. Hence, as described in our experiment pipeline, we, in addition to the regular short VAA stopping criterion of 50%, also tested how many statements in total we would need to arrive at the correct voting advice. Detailed results of this can be found in the appendix and show that our methods require far fewer statements to accurately recommend the Top-1 party than 50%, while this effect is heavily influenced by cross-party fairness, with parties appearing less often in top ranks needing more statements.

## 6 Limitations and Future Work

Our study clearly demonstrates the value of imputation and fully personalized strategies for short VAAs to address advice degradation. At the same time, we shed light on the problem of cross-party fairness and show that, while adaptive strategies are slightly better at representing parties that are less often ranked at the top than the baseline, they still clearly have this problem. As such, a clear recommendation for future work and for developers of short VAAs is to: a) consider and evaluate this problem, and b) develop methods to counter it. One option that may help improve results, particularly for the less prominent parties, is to define a fixed,

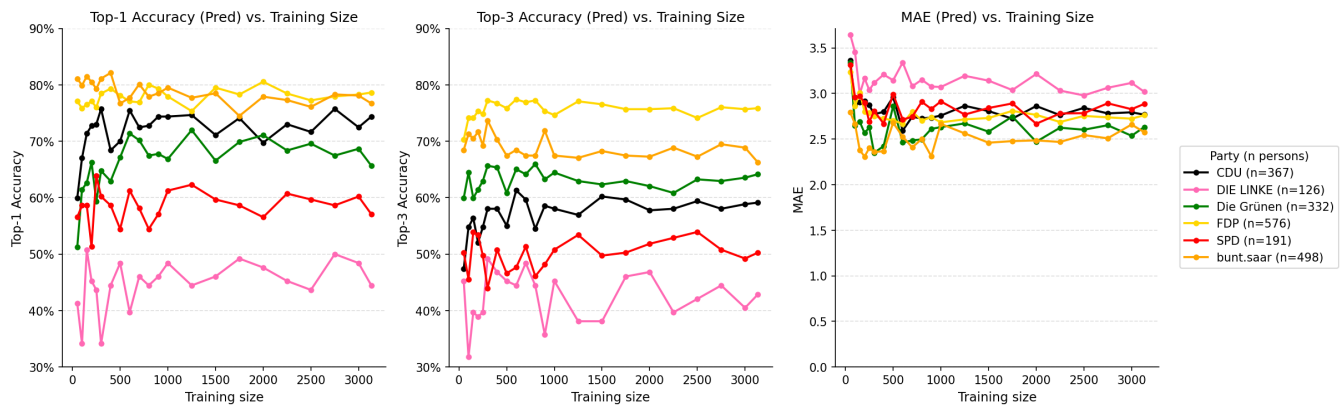


Figure 3: Results for varying training data sizes on the same test set, with 50% stopping and Ridge Regression with entropy as the selection method, for the SB dataset, separated by party.

small set of initial items selected primarily for their ability to differentiate between smaller parties. Having a small, fixed set of items may also reduce the computational effort required by the methods, as fewer models need to be trained initially. Aside from this, further research should test whether sampling methods can improve ML models’ handling of imbalanced data. Generally, our results here should be seen as a proof of concept that adaptive VAAs with imputation have advantages over the regular short VAAs, but we do not claim to have found a one-size-fits-all strategy for adaptive questionnaires.

Beyond the pure technical aspects, it should be noted that existing (long) VAAs may not be perfect and may also not represent niche parties or positions. The adaptive methods, as showcased here, may eventually allow developers to construct much longer VAAs to better differentiate parties and include niche political positions. In fact, we expect the positive effect of adaptive methods to be much more pronounced when the full VAAs are longer. Of course, adaptive questionnaires require training data, and with longer VAAs, it may become increasingly hard to motivate people to complete them. However, as we demonstrated, not much training data is actually needed to get stable results, and one could also add synthetic users’ answers via large language models for initial training until enough real user data is available (Bachmann et al. 2025). Furthermore, since our tests were conducted solely with German municipal elections, further research should retest the proposed methods on other elections, such as national elections, across different social and cultural contexts to further test generalizability. Lastly, while we show that imputation and fully personalized strategies can counter advice degradation, a different argument could be that this actually makes VAAs less comparable and neutral since not every user sees every statement. This difference in what users see may eventually steer users toward or away from certain parties (simply due to the presence of topics and parties in the statements). This point of view has not seen much attention in the discussion surrounding short VAAs so far, and while going beyond the scope of this work is something that should also be critically discussed

and evaluated. Regarding the development of VAAs, we argue that practitioners should keep this in mind and inform users when adaptive or imputation techniques are used to increase transparency.

## 7 Conclusion

Driven by the newly emerging problem of advice degradation caused by short VAA statements, which allow users to skip many VAA statements, this paper investigated the feasibility of an active learner strategy to adaptively order VAA statements and impute missing ones. Our results show that by imputing statements and adapting VAAs to each user, we can significantly reduce advice degradation. Furthermore, we highlighted that regular short VAAs treat minority and majority opinions unequally. While our models weigh more equally, we stress the need for further research to address cross-party fairness so that VAAs remain politically neutral and useful to all. Lastly, we show that our models require only a small amount of initial training data. Overall, this study contributes to the ongoing improvement of democratic recommender systems by introducing new methods and highlighting potential caveats that researchers should be aware of when building ML models for (short) VAAs.

## References

- Albertsen, A. 2022. How do the characteristics of voting advice application users change over time? Evidence from the German election studies. *German Politics*, 31(3): 399–419.
- Anckar, C. 2000. Size and party system fragmentation. *Party Politics*, 6(3): 305–328.
- Bachmann, F.; Sarasua, C.; and Bernstein, A. 2024. Fast and adaptive questionnaires for voting advice applications. In *Joint European Conference on machine learning and knowledge discovery in databases*, 365–380. Springer.
- Bachmann, F.; van der Weijden, D.; Heitz, L.; Sarasua, C.; and Bernstein, A. 2025. Adaptive political surveys and GPT-4: Tackling the cold start problem with simulated user interactions. *PLoS One*, 20(5): e0322690.

- Bachmann, F.; van der Weijden, D.; Sarasua, C.; and Bernstein, A. 2026. Estimating the Recommendation Certainty in Candidate-Based Voting Advice Applications. *Politics and Governance*, 14.
- Berdoz, F.; Brunner, D.; Vonlanthen, Y.; and Wattenhofer, R. 2025. Recommender systems for democracy: toward adversarial robustness in voting advice applications. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI '25*. ISBN 978-1-956792-06-5.
- Blais, A.; Daoust, J.-F.; Dassonneville, R.; and Péloquin-Skulski, G. 2019. What is the cost of voting? *Electoral Studies*, 59: 145–157.
- Bundeszentrale für politische Bildung. 2025. Bundestagswahl 2025.
- Dalton, R. J. 2014. Partisan dealignment and voting choice. *Developments in German politics*, 4: 57–77.
- Dassonneville, R. 2015. *Stability and Change in Voting Behaviour: Macro and Micro Determinants of Electoral Volatility*. Ph.D. thesis.
- Dassonneville, R.; Blais, A.; and Dejaeghere, Y. 2015. Staying with the party, switching or exiting? A comparative analysis of determinants of party switching and abstaining. *Journal of Elections, Public Opinion and Parties*, 25(3): 387–405.
- De Graaf, J. 2010. “The irresistible rise of the StemWijzer,”. In Cedroni, L.; and Garzia, D., eds., *Voting Advice Applications in Europe: The State of the Art*, 35–46.
- Dieing, T.; and Stecker, C. 2025. Denkkzettel für höhere Ebenen? Kommunalwahlen als Protest- und Nebenwahl. In *Gemeinderatswahlen im deutschen Mehrebenensystem: Wählerverhalten und politische Einstellungen auf kommunaler Ebene*, 177–197. Springer.
- Dieing, T. I. 2024. AI-Driven Dialogue: Leveraging Generative AI in Conversational Agent Voting Advice Applications (CAVAAs). In *International Symposium on Chatbots and Human-Centered AI*, 161–180. Springer.
- Dieing, T. I. 2025a. Just can’t get enough – profiling users of multiple Voting Advice Applications. *Journal of Information Technology & Politics*, 0(0): 1–14.
- Dieing, T. I. 2025b. Traditional and Deep Learning Methods for Trustworthy Intent Classification in Conversational Agent Voting Advice Applications (CAVAAs). In *Proceedings of the Mensch Und Computer 2025, MuC '25*, 749–754. New York, NY, USA: Association for Computing Machinery. ISBN 9798400715822.
- Dieing, T. I.; and Siefken, L. 2025. Read Between the Lines-How Linguistics Can Unearth Parties’ True Voting Advice Application Positions. *German Politics*, 1–22.
- Enyedi, Z. 2015. The Influence of Voting Advice Applications on Preferences, Loyalties and Turnout: An Experimental Study. *Political Studies*, 64(4): 1000–1015.
- Ferland, B.; and Dassonneville, R. 2021. Shifting parties, rational switchers: Are voters responding to ideological shifts by political parties? *Party Politics*, 27(1): 114–124.
- Fivaz, J.; and Nadig, G. 2010. Impact of voting advice applications (VAAs) on voter turnout and their potential use for civic education. *Policy & Internet*, 2(4): 167–200.
- Garzia, D.; and da Silva, F. F. 2026. How many people use voting advice applications? Survey evidence from 15 democracies worldwide. *Politics and Governance*, 14.
- Garzia, D.; and Marschall, S. 2019. Voting Advice Applications. In Thompson, W. R., ed., *Oxford research encyclopedia of politics*, Oxford research encyclopedias. New York, NY: Oxford University Press. ISBN 0-19-022863-6.
- Gemenis, K. 2024. Artificial intelligence and voting advice applications. *Frontiers in Political Science*, 6.
- Gemenis, K.; and Rosema, M. 2014. Voting advice applications and electoral turnout. *Electoral studies*, 36: 281–289.
- Germann, M.; Mendez, F.; and Gemenis, K. 2023. Do Voting Advice Applications Affect Party Preferences? Evidence from Field Experiments in Five European Countries. *Political Communication*, 1–19.
- Heinsohn, T.; Fatke, M.; Israel, J.; Marschall, S.; and Schultze, M. 2019. Effects of voting advice applications during election campaigns. Evidence from a panel study at the 2014 European elections. *Journal of Information Technology & Politics*, 16(3): 250–264.
- heise online. 2009. Weiter Streit um Wahl-O-Mat.
- Janega, J.; and Zymová, K. 2025. When too many is too much: the effects of party system fragmentation on voter turnout in Europe. *European Political Science*, 24(4): 635–653.
- Kamoen, N.; Holleman, B.; Krouwel, A.; van de Pol, J.; and de Vreese, C. 2015. The Effect of Voting Advice Applications on Political Knowledge and Vote Choice. *Irish Political Studies*, 30(4): 595–618.
- Kamoen, N.; and Liebrecht, C. 2022. I Need a CAVAA: How Conversational Agent Voting Advice Applications (CAVAAs) Affect Users’ Political Knowledge and Tool Experience. *Frontiers in Artificial Intelligence*, 5.
- Katakis, I.; Tsapatoulis, N.; Mendez, F.; Triga, V.; and Djouvas, C. 2014. Social voting advice applications-definitions, challenges, datasets and evaluation. *IEEE transactions on cybernetics*, 44(7): 1039–1052.
- Liebrecht, C.; Kamoen, N.; and Aerts, C. 2023. Voice Your Opinion! Young Voters’ Usage and Perceptions of a Text-Based, Voice-Based and Text-Voice Combined Conversational Agent Voting Advice Application (CAVAA). In Følstad, A.; Araujo, T.; Papadopoulos, S.; Law, E. L.-C.; Luger, E.; Goodwin, M.; and Brandtzaeg, P. B., eds., *Chatbot Research and Design*, 34–49. Cham: Springer International Publishing. ISBN 978-3-031-25581-6.
- Manavopoulos, V.; Triga, V.; Marschall, S.; and Wurthmann, L. C. 2018. The impact of VAAs on (non-voting) aspects of political participation: Insights from panel data collected during the 2017 German federal elections campaign. *Statistics, Politics and Policy*, 9(2): 105–134.
- Marschall, S.; Hagemann, D. C.; and Bongartz, J. 2026. *Der Wahl-O-Mat*. Springer.

- Mendez, F. 2017. Modeling proximity and directional decisional logic: What can we learn from applying statistical learning techniques to VAA-generated data? *Journal of Elections, Public Opinion and Parties*, 27(1): 31–55.
- Metag, J.; and Gurr, G. 2023. Too much information? A longitudinal analysis of information overload and avoidance of referendum information prior to voting day. *Journalism & Mass Communication Quarterly*, 100(3): 646–667.
- Moreno, G. R.; Padilla, J.; and Chueca, E. 2022. Learning VAA: A new method for matching users to parties in voting advice applications. *Journal of Elections, Public Opinion and Parties*, 32(2): 339–357.
- Munzert, S.; and Ramirez-Ruiz, S. 2021. Meta-Analysis of the Effects of Voting Advice Applications. *Political Communication*, 38(6): 691–706.
- Pianzola, J.; Trechsel, A. H.; Vassil, K.; Schwerdt, G.; and Alvarez, R. M. 2019. The impact of personalized information on vote intention: Evidence from a randomized field experiment. *The Journal of Politics*, 81(3): 833–847.
- Riker, W. H. 1986. Duverger's law revisited. *Electoral laws and their political consequences*, 19–42.
- Schultze, M. 2013. Effekte des Wahl-O-Mat auf politisches Wissen über Parteipositionen. *ZPol Zeitschrift für Politikwissenschaft*, 22(3): 367–391.
- Sigfrid, K. 2024. IRT for voting advice applications: a multi-dimensional test that is adaptive and interpretable. *Quality & Quantity*, 58(5).
- Stadt Halle (Saale). 2024. Ergebnisse der Stadtratswahl 2024.
- Stecker, C.; Dieing, T. I.; Oblong, J.; Strobel, S.; Braun, D.; Deiss-Helbig, E.; Gessler, T.; Hein, M.; Höhne, B.; Kuepers, A.; Linhart, E.; Müller, J.; Pravemann, S.; Reiser, M.; Steiner, N. D.; Thomeczek, J. P.; Völkl, K.; Weisskircher, M.; Wenzelburger, G.; and Zissel, P. 2025. Voting Advice Applications for Local Elections.
- Terán, L.; and Mancera, J. 2019. Dynamic profiles using sentiment analysis and twitter data for voting advice applications. *Government Information Quarterly*, 36(3): 520–535.
- Tromborg, M. W.; and Albertsen, A. 2023. Candidates, voters, and voting advice applications. *European Political Science Review*, 15(4): 582–599.
- Tsapatoulis, N.; and Agathokleous, M. 2017. Forecasting elections from VAA data: What the undecided would vote? In *2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, 46–52.
- van Camp, K.; Lefevere, J.; and Walgrave, S. 2014. The content and formulation of statements in voting advice applications: a comparative analysis of 26 VAAs. In Diego, G.; and Stefan, M., eds., *Matching Voters with Parties and Candidates: Voting Advice Applications in Comparative Perspective*, 11–31. ECPR Press.
- van de Pol, J.; Holleman, B.; Kamoen, N.; Krouwel, A.; and de Vreese, C. 2014. Beyond young, highly educated males: a typology of VAA users. *Journal of Information Technology & Politics*, 11(4): 397–411.
- van de Pol, J.; Kamoen, N.; Krouwel, A.; de Vreese, C.; and Holleman, B. 2019. Same but different: A typology of Voting Advice Application users in first- and second-order elections. *Acta Politica*, 54(2): 225–244.
- Westle, B.; Begemann, C.; and Rütter, A. 2014. The “Wahl-O-Mat” in the course of the German Federal Election 2013 – Effects of a German VAA on users' election-relevant political knowledge. *Zeitschrift für Politikwissenschaft*, 24(4): 389–426.