

Algorithmic Shortlisting in Participatory Budgeting

Juan Ignacio Zambrano^{1,2}, Clément Contet^{1,3}, Jairo Gudiño-Rosero^{1,2,4},
Felipe Garrido-Lucero^{1,2}, Umberto Grandi^{1,2}, César A. Hidalgo^{4,5,6}

¹ IRIT, Toulouse, France

² Université Toulouse Capitole, Toulouse, France

³ Université de Toulouse, Toulouse, France

⁴ Center for Collective Learning, IAST, Toulouse School of Economics,
Université Toulouse Capitole, Toulouse, France

⁵ Center for Collective Learning, CIAS, Corvinus University of Budapest, Budapest, Hungary

⁶ AMBS, University of Manchester, Manchester, UK

Abstract

Participatory budgeting is a democratic innovation that allows citizens to propose and vote on public investment projects. To help organizers manage large volumes of submissions, we design and test AI methods for algorithmic shortlisting. Our algorithms predict which projects are likely to be funded using only project features and anonymous historical voting data. We demonstrate the limitations of a naive approach that uses a large language model to rank projects based on past success and propose a vote-based pipeline that enables state-of-the-art LLMs to perform on par with classical machine learning. Our findings indicate that user preferences in participatory budgeting are stable enough to allow algorithmic shortlisting to approximate an initial selection of projects effectively.

Code availability. The original code, prompts, and datasets used in this study can be accessed at <https://doi.org/10.5281/zenodo.21829180>.

1 Introduction

Cities around the world are increasingly giving citizens a more active role in public decision-making (Mitchell 1996; Peixoto and Fox 2016; Helbing et al. 2024) giving rise to a number of *democratic innovations* (Smith 2009), ranging from in-person processes such as deliberative mini-publics (Farrell et al. 2019) to digital participatory experiments such as petition platforms and participatory budgeting (Hagen 2018; Wampler, McNulty, and Touchton 2021; Li et al. 2025). A common factor of these participatory initiatives is the presence of large volumes of textual data, posing significant challenges in terms of analysis and scalability.

In this context, recent advances in large language models (LLMs) have enabled new approaches to democratic citizen participation. Notable examples include the popular deliberative platform Pol.is (Small et al. 2023), with LLMs producing a report of the 3,940 ideas commented and voted by 8,000 residents during a digital large-scale town-hall meeting in Bowling Green, U.S. (Jigsaw-Google 2025), or the use of LLMs to summarize contrasting opinions in deliberative experiments aimed at finding consensual statements on policy questions (Tessler et al. 2024; Konya et al. 2025).

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

In this article, we focus on *participatory budgeting* (PB), a widely adopted democratic process by which a portion of a municipal or regional budget is allocated to projects proposed and selected by its residents. PB processes are usually composed of three main phases (Wampler 2000; Cabannes 2004; Shah 2007): a *proposal phase*, where citizens submit project ideas and their costs are estimated by an expert entity; a *consolidation phase*, where proposals are shortlisted after removing infeasible or redundant projects; and a *voting phase*, where citizens select from among the finalists. At the end of the voting phase, votes are aggregated and combined with budget constraints to determine the winning projects using a predefined selection rule.

Researchers in computational social choice have focused on designing and testing rules for the voting phase (see, e.g., Talmon and Faliszewski (2019); Goel et al. (2019); Peters, Pierczyński, and Skowron (2021)). In contrast, little attention has been given to the consolidation phase, with the notable exception of Rey, Endriss, and de Haan (2021). Yet, in most PB campaigns, this phase remains rather obscure and poorly documented, typically involving in-person committees taking decisions without well-specified objectives.¹ We claim that a fully specified and documented algorithmic process for shortlisting could substantially improve transparency while reducing the costs associated with processing proposed projects, e.g., through a reliable and testable pre-screening mechanism.

A typical project proposal in PB is characterized by its name, description, category, geographical location, and associated cost (with both name and description being given by the proposer), constituting a rich corpus of textual data. The use of LLMs for treating project proposals therefore arises as a natural idea. Damle and Faltings (2025), for instance, tested the ability of LLMs to select winning projects from community metadata such as age, education, and location. Their work, however, pushes AI usage to a boundary of legitimacy that might be hard to accept by PB organizers or by the population as, e.g., their tools might be used to run

¹To get an example of scale, the City of Cambridge, MA, U.S., received over 1300 project proposals and shortlisted 20 of them in a consolidation process lasting months (Cambridge 2025).

PB elections without asking the citizens to actually vote.

Our objective is different: We aim at building a support-tool for PB organizers trying to minimize the AI intervention and leaving voters with the last say on winning projects. Motivated by the observation that a naive LLM shortlisting mechanism may fail to identify the most popular projects, we focus on estimating the success of project proposals in terms of number of votes, producing a coarse pre-ranking of projects to be used, e.g., when prioritizing those that might pass to feasibility study. Our methodology leverages historical voting data, as PB campaigns are often repeated in each city, and the rich textual and tabular data within project proposals. Importantly, we rely only on publicly available data, designing predictors that comply with principles of data minimization, essential in democratic applications to build trust and to encourage participation.

Contributions. Our contributions are the following.

- We show that a naive approach of prompt-based shortlisting is largely inefficient when tested on six target PB campaigns in three European cities with different linguistic and cultural contexts (Toulouse, France, Wrocław and Warszawa, Poland) which shared their data for multiple years on Pabulib.org (Faliszewski et al. 2023).
- We combine two classical machine learning approaches (ElasticNet and XGBoost) with textual preprocessing strategies, testing their capability to order projects by number of votes in a PB campaign when trained on the outcomes of the previous campaigns.
- We design three prompting strategies on off-the-shelf LLMs (GPT-4o mini and Mistral Large 3) provided with progressively richer context from previous PB campaigns, testing their capability to order projects by number of votes in subsequent campaigns.
- We construct and release datasets covering multi-year PB processes for the cities of Toulouse, France, and Wrocław, Poland, combining voting data from Pabulib.org with project descriptions publicly available on the organisers’ webpages.

Our results show that our pipeline for vote-based algorithmic shortlisting can meaningfully predict project success albeit with important limitations. The performance of LLMs is comparable to classical machine learning methods but requires a retrieval-augmented generation (RAG) setup to obtain consistent results, thus requiring computational skills that are typically not common in PB organising bodies. We also show that textual project information can improve predictive performance in some settings, highlighting the potential of augmenting classical voting datasets such as Preflib.org and Pabulib.org with textual descriptions of the alternatives (Grandi et al. 2026). This further highlights the potential of our approach, since our results match and often surpass those obtained using LLM personas based on participants’ demographic data (Yang et al. 2024a), demonstrating that historical data from same-city PB processes can effectively replace demographic inputs.

Related Work. Citizens’ involvement in proposing and selecting public projects began in Latin America in the 1980s and later spread globally. The AI research community has focused on the mechanism design and computa-

tional aspects of this process, with PB viewed as a form of multi-winner voting under budget constraints (Lackner and Skowron 2023). PB has been extensively studied in computational social choice (see the recent survey by Rey, Schmidt, and Maly, 2025). Key contributions include the analysis of proportionality axioms (e.g., Aziz and Lee, 2021) and the development of a fair and computationally tractable PB rule (Peters, Pierczyński, and Skowron 2021), which has been recently tested in European cities and in behavioral experiments (Yang et al. 2024b).² Rey, Endriss, and de Haan (2021) are the only ones to formally study a PB model including both a consolidation phase and a voting phase, proposing aggregation functions for shortlisting and analyzing incentive-compatibility.

Arana-Catania et al. (2021) and Grossi et al. (2024) argue that a range of computational methods must be mobilized for digital democracy to succeed. LLMs are among the most actively explored tools in this domain, from generating consensus statements in deliberative settings (Tessler et al. 2024; Konya et al. 2025), to enhancing participatory platforms (Davies et al. 2021; Small et al. 2023), to selecting proposals that meet representation guarantees (Fish et al. 2024). An important approach employs LLMs to create voting personas or digital twins—autonomous agents replicating individual voter behavior—within the broader framework of augmented democracy (Hidalgo 2018; Grandi 2019; García-Marzá and Calvo 2024). Key examples include Gudiño, Grandi, and Hidalgo (2024), who augmented voting data on Brazilian government proposals using fine-tuned LLMs, and Yang et al. (2024a), who simulated a voter population in a PB experiment using LLM personas generated via prompting. Constructing LLM personas typically requires demographic data, at odds with our goal of building a PB-support tool that relies solely on publicly available data.

2 Preliminaries

This section is dedicated to explain the context of a participatory budgeting process, the datasets we have used during the article, and to show how LLM-naive-shortlisting might struggle to find the most popular projects.

Context: Participatory Budgeting

A PB campaign starts with the announcement of an available budget by the organising body, followed by the collection of project proposals from the population. A shortlisting phase follows, and eligible citizens are then invited to vote on the consolidated list of projects during a pre-defined time window. Citizens are typically asked to select up to k projects ($k = 3$ in Toulouse, $k = 2$ in Wrocław, and $k = 10$ in Warszawa, the cities analyzed in this paper), either online or on-site. Once votes have been collected, a voting rule selects a set of projects to be funded given the budget constraint and the elicited preferences. The greedy rule³ is the most commonly used by municipalities due to its simplicity and ex-

²See also <https://equalshares.net/>

³The greedy rule selects projects in order of vote count, each time updating the remaining budget and skipping a project if there is not enough remaining budget to fund it (Aziz and Shah 2020).

ID	NAME	DESCRIPTION	CATEGORY	COST	DIST.	VOTES
201	PING-PONG TABLES AND VOLLEYBALL COURTS	MOVEMENT IS HEALTH. IN WROCLAW, THERE IS A HIGH DEMAND FOR PLACES WHERE ONE CAN ACTIVELY SPEND TIME OUTDOORS. ...	SPORT	1 000 000	1	3 974
694	"GREEN BUS STOPS" IN THE CITY CENTER	THE PROJECT PROPOSES THE CREATION OF GREEN BUS STOPS AT SELECTED LOCATIONS IN DOWNTOWN WROCLAW BY PLANTING CLIMBING PLANTS...	GREENERY	200 000	1	943

Table 1: Example of two projects translated to English from the 2017 Wroclaw PB campaign. Each project is characterized by an ID, a name, a textual description, a category, its cost, the city district, and the number of votes received.

plainability, but there are many other ways to select winning projects in PB, such as the method of equal shares (Peters, Pierczyński, and Skowron 2021), quadratic voting (Lalley and Weyl 2018), or knapsack methods (Goel et al. 2019).

Multi-Year Datasets

Voting data for PB campaigns is often made available by organisers, and a number of datasets have been collected on the public repository Pabulib.org (Faliszewski et al. 2023). Our objective of developing PB predictors requires multi-year datasets of PB instances in the same city using the same ballot format. Among those available on Pabulib, at least three cities satisfy these requirements: Toulouse 2022 and 2024 in France, Wroclaw 2016, 2017, and 2018 in Poland, and Warszawa 2021, 2022, 2023, and 2024 in Poland. We consolidated datasets for these cities using approval-election instances from Pabulib, which typically include a unique project ID, the project name, category, estimated cost, district or neighborhood, and vote count. For Toulouse 2022 and 2024 and Wroclaw 2016 and 2017, we further augmented the data with textual descriptions of each project, collected by web scraping on the webpages of PB organizers. These descriptions are exploited in the last experiment of the article. Examples of projects (with description) can be seen in Table 1. More examples and a detailed description of the datasets and their sources are available in the extended version of our paper (Zambrano et al. 2026).

For the city of Toulouse, our analysis focuses on the PB campaigns held in 2022 and 2024, during which 200 and 183 projects, respectively, were short-listed and submitted for public voting. For the city of Wroclaw, our analysis focuses on the PB elections held in 2016, 2017, and 2018, during which 52, 50, and 39 projects, respectively, were short-listed and submitted to public vote. It is important to note that the Wroclaw dataset in Pabulib includes some demographic information about voters which we removed to ensure consistency. Finally, for the city of Warszawa, our analysis focuses on the PB campaigns held from 2021 to 2024, during which 106, 129, 138, and 118 projects, respectively, were short-listed and submitted to public vote. Table 2 provides details of the mentioned campaigns.⁴

In Toulouse, voters supported on average 3 projects in each ballot, giving a total number of 11,606 votes in 2022

and 21,780 in 2024. In Wroclaw, voters could support up to 2 projects, giving 119,194 total votes in 2016, 111,961 in 2017, and 94,995 in 2018. In Warszawa, voters could support up to 10 projects, giving 958,990 total votes in 2021, 812,340 in 2022, 789,040 in 2023, and 752,550 in 2024. The budgets in Toulouse and Warszawa are 6 to 8 times the budget in Wroclaw. Regarding project’s average cost in euros, projects in Wroclaw have on average half the cost of those in Warszawa and have higher costs by a factor of 1.8 than those in Toulouse.

City	Year	Projects	Voters	Budget
Wroclaw, PL	2016	52	67 103	1 M €
Wroclaw, PL	2017	50	62 529	900k €
Wroclaw, PL	2018	39	53 801	900k €
Toulouse, FR	2022	200	4 532	8.0 M €
Toulouse, FR	2024	183	7 260	8.0 M €
Warszawa, PL	2021	106	95 899	5.8 M €
Warszawa, PL	2022	129	81 234	6.6 M €
Warszawa, PL	2023	138	78 904	7.1 M €
Warszawa, PL	2024	118	75 255	7.1 M €

Table 2: Summary of the PB campaigns studied in this article.

Naive Shortlisting

Datasets in PB campaigns present an important textual dimension. Additionally, the easy access to the use of LLMs by non-experts motivate us to assess the performance of naively trying to shortlist the projects by simply prompting the task to an LLM. Since the whole list of proposed projects (before the actual shortlisting) is not available, we asked two out-of-the-box LLMs, namely, GPT-4o mini and Mistral Large 3, to select the top $k\%$ of the projects with $k \in \{10, 20, 30, 50\}$, that is, to make a shortlisting of the already shortlisted list, in order to check whether they are capable to find the most popular projects. Moreover, we considered two versions for each LLM, namely, zero-shot, where the LLM receives no context, and full-context, where the LLM receives the list of winners of the previous campaign. Note that, for the sake of comparability, none of the predictions nor context included the projects’ descriptions.

Figure 1 shows the Jaccard index (cf. Equation (1)) obtained by naive shortlisting in six of our instances (the oldest PB campaign of each city is not considered as the “previous campaign” does not exist). We observe that all methods

⁴We converted budgets to euros using the 2026 exchange rate of 1 zł = 0.24 €.

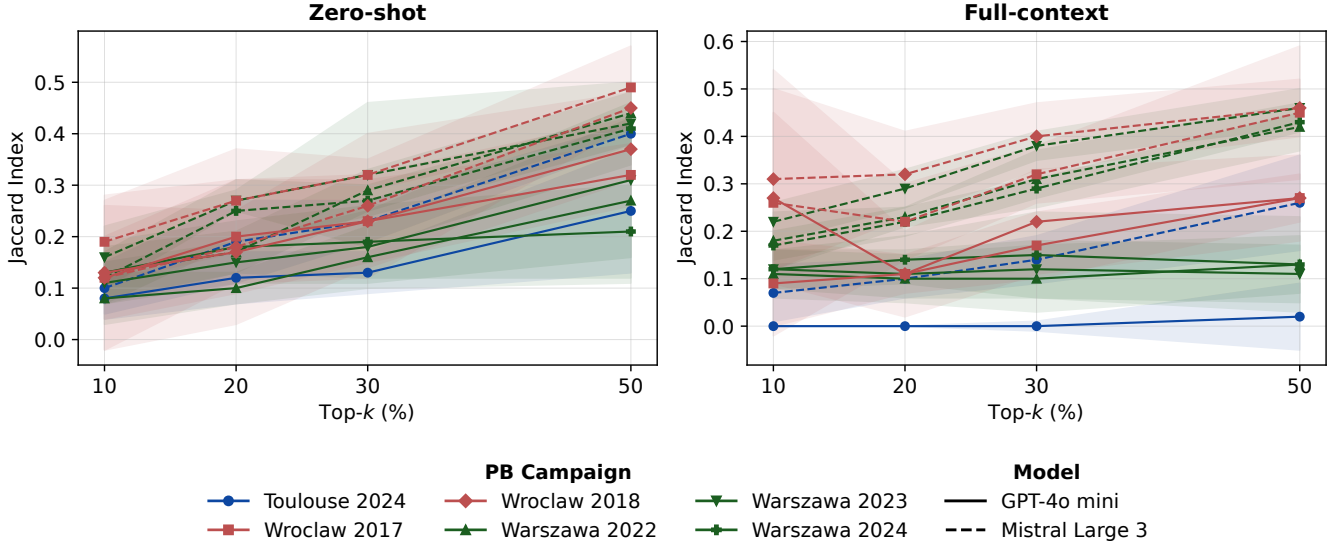


Figure 1: Jaccard index of naive shortlisting for two LLMs, with and without context, over the six PB campaigns. Results are averaged over 20 runs plotted with standard deviation. Performance has high variability and is close to random for all k .

struggle to identify popular projects. Surprisingly, the performance worsens when presented with context from previous years, in part due to a higher number of LLM hallucinations, showing that a more clever way of exploiting textual information and past voting data is required. See (Zambrano et al. 2026) for a detailed explanation of the prompts used as well as any computational specification.

Since a naive use of LLMs for shortlisting fails, in the following section we introduce a new shortlisting methodology based on predicting the vote count of each project.

3 Vote-Based Shortlisting

We propose a novel methodology based on predicting the level of public support, given by the vote count, of each project as an intermediate step for PB shortlisting. Given the projects submitted to public voting in a target PB campaign, we first predict the number of votes that each project will receive using either machine learning models or LLM-based predictors with different levels of contextual information. The predicted vote counts are then used to rank projects in descending order of expected public support. We compare this predicted ranking with the actual ranking observed in the corresponding PB campaign.

To test whether our predictions are accurate, we compute the **Kendall’s Tau coefficient** $\tau \in [-1, 1]$ between the real and predicted rankings, where $\tau = 1$ denotes perfect agreement and $\tau = -1$ denotes perfect disagreement. To get a more fine-grained analysis of our predictors, we also measure the agreement between the real and predicted rankings using the **Jaccard index**:

$$J(\hat{\text{top}}_k, \text{top}_k) := \frac{|\hat{\text{top}}_k \cap \text{top}_k|}{|\hat{\text{top}}_k \cup \text{top}_k|}, \quad (1)$$

where $\hat{\text{top}}_k$ denotes the predicted top- k projects and top_k the real ones. Unlike the Kendall’s tau coefficient, the Jac-

card index does not look at the actual order but only whether a given project is included in the top- k or not.

Machine Learning Predictors

We implemented classical supervised machine learning models using an incremental feature inclusion approach to predict the number of votes received by each project normalized by the total voter count. Models were trained using a PB elections in each city and evaluated on the following PB campaign: Toulouse 2022 was used to predict Toulouse 2024; Wroclaw 2016 and 2017 were used to predict Wroclaw 2017 and 2018, respectively; and Warszawa 2021, 2022, and 2023 were used to predict Warszawa 2022, 2023, and 2024, respectively.

We selected two representative architectures: a regression model (ElasticNet) and a gradient-boosted tree model (XGBoost), combined with two textual processing strategies:

- TF-IDF, from term frequency-inverse document frequency (Jones 1972), measures how important a word is to a document by balancing how often it appears in the document against how rare it is across an entire collection of documents.
- BGE-M3 (Chen et al. 2024), a multilingual encoders that transform textual data into embedding vectors, which was used to encode the full text directly.

Since our PB datasets contain a limited number of projects, we performed a principal component analysis (PCA) during training for each combination of ML model/semantic embedding in order to reduce the number of features representing the textual data. To obtain robust results, we made 20 PCA runs and perform predictions for each of them. Table 3 includes the median values for the PCA dimensions.

TARGET INSTANCE	ELASTICNET	XGBOOST
TOULOUSE 2024	2	4
WROCLAW 2017	8	5
WROCLAW 2018	2	20
WARSZAWA 2022	2	52
WARSZAWA 2023	20	79
WARSZAWA 2024	4	31

Table 3: Best-performing PCA dimensions for BGE-M3 embeddings computed from project titles only, on PB instances without project descriptions. Reported dimensions correspond to the median value over 20 runs.

As a benchmark, we also tested the NO-TEXT configuration for both architectures where each of them was trained only on tabular data (district, category, and project cost).

All experiments were conducted using Google Colab instances. Traditional machine learning models were trained on standard runtimes equipped with an Intel(R) Xeon(R) CPU @ 2.20GHz and 12.7 GB of RAM, with per-run training times ranging from 15 to 20 minutes for Wroclaw, between 30 minutes and 2 hours for Toulouse, and between 40 minutes and 1 hour and 30 minutes for Warszawa, depending on the model and feature variant. These times refer only to the training of the ML predictors.

Large Language Models

We design and compare three prompt variants that provide progressively richer contextual information to assess LLMs’ ability to predict project support. Our main hypothesis is that, when supplied with sufficient context, LLMs may capture complex multidimensional patterns that jointly reflect proposal attributes and cities’ social and territorial characteristics, consistent with recent evidence from cross-modal urban data integration (Feng et al. 2025a), spatio-temporal prediction (Li et al. 2024), and LLM-based urban research benchmarks (Feng et al. 2025b).

The shared prompt across all variants, illustrated in Figure 2, includes a brief description of the participatory budgeting process, specifies its geographical setting, and instructs the LLM to predict the total number of votes received by each project in the target PB campaigns: Toulouse 2024, Wroclaw 2017-2018, and Warszawa 2022-2024. For each variant—zero-shot, full context, and retrieval-augmented generation (RAG)—we fill the context box according to the following descriptions:

- **Zero-Shot.** The LLM receives no information from previous PB campaigns, predicting the votes for each project based solely on the shared prompt.
- **Full Context.** The LLM is presented with the list of projects from the previous election with their title, category, cost, district of submission, and vote count.
- **Retrieval-Augmented Generation (RAG).** The LLM receives information on the 20% of projects from the previous campaign that are most semantically similar to the target project (all their features: title, category, location, cost, and vote count). In addition, it also receives the

You are an expert model in analyzing participatory budgeting elections, specifically within the context of the city of Wroclaw, Poland. In Wroclaw, the municipality organizes participatory budgeting elections to fund citizen projects. Each resident can vote for 1 or 2 initiatives on the official website. Votes are anonymous.

Once the voting period ends, projects are ranked according to the number of votes obtained and assigned a rank. Then, the winning projects are selected using a greedy algorithm: starting with the project that received the most votes, subsequent most-voted projects are added as long as the available budget allows. When the remaining budget no longer allows for the funding of the next project in the list, the process moves to the next one that can be funded, and so on, until the total budget of 4,000,000 PLN is exhausted.

Context from previous election

Your task is to analyze the following project proposed in the context of a new participatory budget election:

Project: {project_name}, **Cost:** {cost} PLN, **District:** {district}

Based on:

- Your knowledge of Wroclaw, its urban context, as well as its social and environmental issues.
- The project’s characteristics (theme, cost, district, etc.).
- Typical voting dynamics in participatory elections.
- The given context.

Answer the following questions regarding this project:

What is a reasonable estimate of votes this project could get?

First, mentally analyze the scoring criteria (impact, cost, district) and compare them with historical results. Without writing your analysis at the beginning, provide the result directly in the form of a JSON dictionary:

```
""json [{"estimated_votes": XXXX}]""
```

Then develop your reasoning in 5 to 8 key points.

Figure 2: Prompt for the Wroclaw dataset. The *context from previous election* box represents the context given by the three prompt variants considered (zero-shot, full context, RAG).

three projects from the same district as the target project that are most semantically similar to it. Similar projects were computed by using the cosine similarity measure.

The implementation of the three prompt variants included the **Chain-of-Thought** technique (Wei et al. 2022), a reasoning strategy that forces the model to generate logical steps leading to a solution, as illustrated in the final part of the prompt in Figure 2. LLM-based predictors were accessed through external APIs and were not run locally. The LLM experiments were implemented using GPT-4o mini through the OpenAI API and Mistral Large 3 through the Mistral API. To reduce stochastic variation and facilitate reproducibility, we set the temperature to 0 for all prompts. In the zero-shot setting, for a given project and PB instance, the same prompt is generated. In the full-context setting, the order in which projects are presented in the prompt is randomized using a fixed seed. In the RAG setting, retrieved projects are ranked by cosine distance, which yields a deterministic ordering for a fixed embedding space and query. For all inferences, the maximum response length was constrained to 2,048 tokens. Since these settings do not guarantee strictly deterministic outputs across all providers, LLM results are averaged over multiple predictions when applicable. Despite

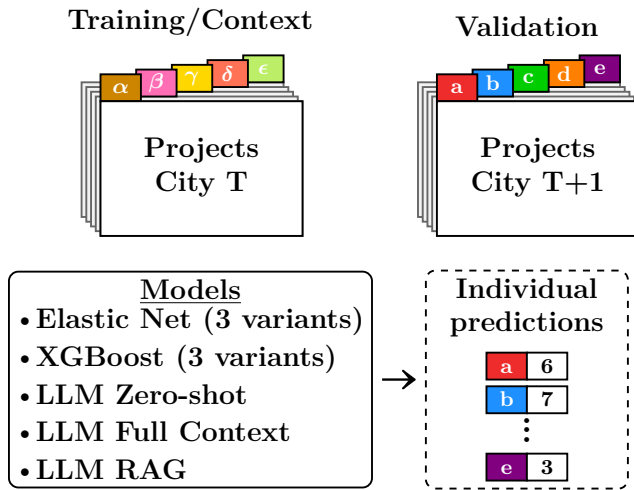


Figure 3: General workflow of our methodology to validate predictions of voting outcomes in PB elections. Projects and vote counts from a PB election are given as training data, with the objective of predicting the vote count of projects in the following election.

the differences in implementation, the time required for any model (ML or LLM-based) to produce an independent prediction remained under one minute, although API rate limits could increase the wall-clock time required to collect all LLM responses. For the SBERT and BGE-M3 embedding generation, we used GPU-accelerated Google Colab instances featuring an NVIDIA A100 Tensor Core GPU.

No Prior Data Exposure

To ensure that the language models used in this study had no prior exposure to the voting results or full project descriptions, we adapted the methodology proposed by Golchin and Surdeanu (2024). Their technique evaluates whether a model has been trained on a specific set of examples, by measuring the model’s ability to reconstruct exact or semantically similar text when presented with incomplete fragments, particularly in text completion or contextual recall. We selected a random sample from our dataset of projects in Toulouse 2022-2024 and Wroclaw 2016-2017, those for which project descriptions are available, focusing on their descriptions, districts, and costs, and implemented two types of test:

1. **Project Description Completion.** We prompted the LLM to generate the full description of a project under two conditions: (a) without providing any part of the original text (only the project title), and (b) by providing the first few lines of the original description. We analyzed both lexical and semantic similarity between the real and generated descriptions. In all cases, the LLM produced generic or irrelevant content and was unable to reproduce the original description, either partially or in full.
2. **Cost and District Retrieval:** In a second test, we provided the LLM with the project title and its full description and asked it to identify the district where the project

was proposed and its estimated cost. Again, the model failed to retrieve any of this information correctly.

Random Baseline

We implemented a random baseline by drawing uniformly at random a ranking of projects for each voter, a preference generation method known as the impartial culture in social choice (Mattei and Walsh 2013). For each voter, we kept the top-3 preferred projects in Toulouse, the top-2 in Wroclaw, and the top-10 in Warszawa, which in turn are aggregated to compute the predicted vote count for each project. Our results are averaged over 20 rounds of preference generation for the instances without project descriptions and 100 rounds for the instances with project descriptions.

Results

We evaluate our proposed pipeline for vote-based algorithmic shortlisting by comparing the predicted project support with the observed outcomes of the six PB campaigns we considered. We report results for both classical ML models and LLM-based predictors in Table 4.

First, we observe that all our models are better predictors, and many are significantly better, than the random baseline on most measures. Second, the LLM-based prompting approach achieves performances comparable to, and in some cases exceeding, that of the ML models: $\tau = 0.39$ in Toulouse for Mistral Large 3 RAG against $\tau = 0.30$ for XGBoost in its TF-IDF variant, the best ML model for this election; or $\tau = 0.56$ for Mistral Large 3 RAG against $\tau = 0.40$ for ElasticNet BGE-M3 in Warszawa. As expected, most of the time the best results for LLM prompting are obtained by the RAG variant, as the context is carefully tailored to the project under consideration, or by the Full Context variant. The latter, however, can give lower performances if the length of the context window is too large (such as in Toulouse), which can backfire and pollute the LLM reasoning (Liu et al. 2024). We still observe that the LLM predictions are surprisingly accurate, given that LLMs are made for predicting text and not scalars such as the vote count (see the discussion in Section 4 on different types of prompts), unlike the classical ML methods we implemented.

Digging further to analyze the LLM predictions, Figure 4 shows the predicted and real vote count of Mistral, respectively for zero-shot and RAG prompts. Without the context, the zero-shot prompt variant divides all projects into classes and predicts the same number of votes for each project within the same class, with a quite inaccurate scale for the number of votes. This shows that only prior knowledge about the city is not sufficient to retrieve the voters’ opinions. Moreover, it showcases that LLMs have an impressive capacity of aligning their predictions to the problem at hand when provided relevant context.

Adding Project Descriptions

For two datasets, Toulouse 2022–2024 and Wroclaw 2016–2017, we were able to complement our data project textual descriptions. Table 5 reports the results on these two datasets. Given the additional textual component, we only

		KENDALL τ					
MODEL	VARIANT	TOUL. 24	WRO.17	WRO.18	WAR.22	WAR.23	WAR.24
RANDOM	–	0.00 $\pm$ 0.03	0.00 $\pm$ 0.10	0.02 $\pm$ 0.08	-0.01 $\pm$ 0.07	-0.03 $\pm$ 0.06	-0.01 $\pm$ 0.05
ELASTICNET	NO-TEXT	0.22 $\pm$ 0.02	0.33 $\pm$ 0.04	0.14 $\pm$ 0.01	0.32 $\pm$ 0.01	0.37 $\pm$ 0.03	0.25 $\pm$ 0.01
	TF-IDF	0.25 $\pm$ 0.01	0.26 $\pm$ 0.05	0.39 $\pm$ 0.02	0.37 $\pm$ 0.04	0.33 $\pm$ 0.01	0.23 $\pm$ 0.02
	BGE-M3	0.21 $\pm$ 0.06	0.16 $\pm$ 0.04	0.28 $\pm$ 0.07	0.36 $\pm$ 0.01	0.40 $\pm$ 0.01	0.30 $\pm$ 0.02
XGBOOST	NO-TEXT	0.23 $\pm$ 0.00	0.29 $\pm$ 0.03	0.34 $\pm$ 0.00	0.37 $\pm$ 0.00	0.18 $\pm$ 0.00	0.24 $\pm$ 0.01
	TF-IDF	0.30 $\pm$ 0.01	0.44 $\pm$ 0.01	0.41 $\pm$ 0.00	0.36 $\pm$ 0.00	0.32 $\pm$ 0.00	0.24 $\pm$ 0.01
	BGE-M3	0.28 $\pm$ 0.01	0.26 $\pm$ 0.01	0.44 $\pm$ 0.01	0.40 $\pm$ 0.01	0.39 $\pm$ 0.00	0.37 $\pm$ 0.01
GPT-4O MINI	ZERO-SHOT	0.24 $\pm$ 0.02	0.36 $\pm$ 0.04	0.33 $\pm$ 0.06	0.26 $\pm$ 0.03	0.31 $\pm$ 0.01	0.23 $\pm$ 0.05
	FULL CONTEXT	0.22 $\pm$ 0.05	0.48 $\pm$ 0.04	0.43 $\pm$ 0.02	0.38 $\pm$ 0.02	0.44 $\pm$ 0.05	0.36 $\pm$ 0.02
	RAG	0.37 $\pm$ 0.01	0.34 $\pm$ 0.02	0.44 $\pm$ 0.00	0.44 $\pm$ 0.00	0.49 $\pm$ 0.00	0.53 $\pm$ 0.00
MISTRAL LARGE 3	ZERO-SHOT	0.09 $\pm$ 0.02	0.18 $\pm$ 0.01	0.27 $\pm$ 0.05	0.40 $\pm$ 0.03	0.50 $\pm$ 0.01	0.48 $\pm$ 0.02
	FULL CONTEXT	0.36 $\pm$ 0.01	0.54 $\pm$ 0.04	0.47 $\pm$ 0.04	0.47 $\pm$ 0.03	0.52 $\pm$ 0.01	0.46 $\pm$ 0.01
	RAG	0.39 $\pm$ 0.00	0.43 $\pm$ 0.01	0.44 $\pm$ 0.01	0.54 $\pm$ 0.01	0.56 $\pm$ 0.01	0.49 $\pm$ 0.01

Table 4: Kendall’s τ performance of vote-based algorithmic shortlisting on the six datasets. Values are reported as mean $\pm$ standard deviation, computed over 20 runs for ML models and 5 predictions for LLMs. Columns correspond to Toulouse 2024, Wrocław 2017–2018, and Warszawa 2022–2024. Results in boldface are the best model for each dataset. Results are statistically significant ($p < 0.05$), with p -values calculated on the rankings obtained from the average vote count of projects over the various runs of each table entry.

run the LLM-based pipelines. In addition to the Kendall-Tau distance, we also compute the lowest position of an actual winning project in the predicted ranking, to obtain an estimation of how coarse a shortlisting can be to include all the winning projects.

First, Comparing Tables 4 and 5, we observe that adding project descriptions improves or maintains performance across all settings, with the exception of Mistral in Wrocław. For Toulouse, the additional textual information improves all models, with Mistral Large 3 RAG increasing from $\tau = 0.39$ without descriptions to $\tau = 0.46$ with descriptions. For Wrocław, however, adding project descriptions can be detrimental: the Mistral RAG score decreases from 0.43 in Table 4 to 0.33 in the augmented setting. Second, the random baseline suggests that the Toulouse PB election results are easier to predict than the Wrocław one (lowest position of a greedy winner at 131 out of 183 in Toulouse against 17 out of 50 in Wrocław). This can be explained by the fact that in Wrocław the budget is smaller and so the number of accepted projects is lower (72 accepted projects out of 183 in Toulouse-2024, 5 accepted projects out of 50 in Wrocław-2017). Moreover, the high variability of the results in-city and between-city show that a shortlist that contains all real winning projects might actually be of size comparable to that of all projects, limiting the gains of algorithmic shortlisting.

4 Conclusion and Discussion

We introduced an innovative methodology for algorithmic shortlisting in PB processes, comparing traditional ML methods with three LLM prompts on multi-year PB datasets from three cities. Our results show that LLMs can be prompted effectively to obtain predictions of the ranking of most-voted projects, with context-aware prompting strate-

gies obtaining competitive and stable performances. Our findings show that LLMs can be a viable tool for PB organisers, as their natural language interface lowers the technical barriers to deploying algorithmic shortlisting tools. Their main limitation occurs when coupled with a voting rule such as the widely used greedy rule, where we highlighted a limited applicability to find a shortlist guaranteeing that all winning projects are present. Still, our tools can support innovations in digital democracy by helping PB organizers manage large numbers of proposals, as our pipelines can recover a substantial share of the top projects in the real rankings across the considered PB instances. Importantly, our proposed methodology has low data-collection barriers, and LLM-based implementations require fewer AI skills than traditional predictive pipelines, encouraging adoption by PB actors.

Demographics-based and voter-level predictions. Voting rules for participatory budgeting such as the methods of equal shares (Peters, Pierczyński, and Skowron 2021) require the full information on individual ballots to be computed (basically, which voters approved which project). Previous work by Yang et al. (2024a) and by Gudiño, Grandi, and Hidalgo (2024) attempted to obtain voter-level predictions by constructing LLM personas from the demographics of the participants—the first paper using data from a PB experiment, the second using pairwise comparisons over political proposals. Both papers indicate that accuracy for voter-level prediction is low (still, Gudiño, Grandi, and Hidalgo (2024) shows accuracies between 68% and 77% for individual choices, with slightly lower performance when comparing proposals from the same political candidate). Moreover, Yang et al. (2024a) showed important order-effects in PB elections, due to the ordering of presentation of projects on the ballots. However, at the aggregate level, LLM predic-

MODEL	VARIANTS	KENDALL τ		LAST GREEDY WINNER POS.	
		TOUL. 24	WRO. 17	TOUL.24	WRO.17
RANDOM	–	0.00 $\pm$ 0.00	0.00 $\pm$ 0.03	131.5 OUT OF 183	17.35 OUT OF 50
GPT-4o MINI	ZERO-SHOT	0.25 $\pm$ 0.02	0.40 $\pm$ 0.01	161	16
	FULL CONTEXT	0.28 $\pm$ 0.02	0.46 $\pm$ 0.06	164	16
	RAG	0.38 $\pm$ 0.01	0.22 $\pm$ 0.00	154	29
MISTRAL LARGE 3	ZERO-SHOT	0.13 $\pm$ 0.07	0.32 $\pm$ 0.13	140	19
	FULL CONTEXT	0.36 $\pm$ 0.01	0.54 $\pm$ 0.04	160	4
	RAG	0.46 $\pm$ 0.02	0.33 $\pm$ 0.01	149	22

Table 5: Performance of three prompting strategies for two LLMs on PB instances with project descriptions, for Toulouse and Wroclaw. Metrics include Kendall’s τ and the position of the last project selected under the greedy rule. Kendall’s τ is reported as mean $\pm$ standard deviation. For reference, the last project selected under the greedy rule is ranked 84th in Toulouse 2024 and 5th in Wroclaw 2017. Results are statistically significant ($p < 0.05$), with p -values calculated on the rankings obtained from the average vote count of projects over the various runs of each table entry.

tions recover the ground truth more reliably. A direct comparison between Yang et al. (2024a)’s results and our approach is not possible, as their data is a single-year instance. However, our results on the ordinal ranking are comparable to their aggregated findings (see Figure 4 in Yang et al. (2024a)), where GPT-4o mini achieved a Kendall- τ of 0.41 against a human baseline, while we obtain results between 0.29 and 0.44 with RAG prompts, and 0.35 and 0.38 with Full Context prompts. This suggests that demographic data may be less effective than project descriptions and historical election data in predicting project performance, while being publicly accessible (a finding also confirmed by Gudiño, Grandi, and Hidalgo (2024)’s ablation studies).

Cost bias. In most real-world instances the ratio between the vote count of a project and the total number of voters is close to the ratio between its cost and the total budget (Schmidt 2024). This means that running a greedy mechanism over projects ranked by their vote count produces biased results toward higher-cost projects. We have preliminary evidence that helping the ML methods to focus on cost can improve their performance, and future work could test whether this is also true of LLM prompting (e.g., with a cost-oriented RAG).

Type of prompt. Using the terminology of Licht et al. (2025), our prompts implement *pointwise scoring* (scalar prediction) and compare three strategies: zero-shot, few-shot, and many-shot in-context learning. Since LLMs are often more effective at *pairwise scoring*, future work could compare our approach to prompts that ask which of two projects is more likely to receive more votes, avoiding cardinal predictions at the cost of substantially more LLM queries to recover a full ranking. Alternatively, posterior pairwise scoring among the top 30% predicted most-voted projects with our methods, could be tested to recover the ranking of the highest scored projects.

Limitations. GPT-4o mini is a commercial LLM and may not be accessible to municipalities running PB campaigns. Open-source LLMs such as Mistral Large 3 are alternative options, and complementary experiments show they have similar and sometimes better performances. However, our

results show that LLMs cannot be used as out-of-the-box shortlisting tools by PB organizers but require some technical skills to set up, e.g., a RAG. Future work can experiment with more easily accessible prompting strategies building on the Full Context prompting results.

Our approach, like any shortlisting method based on historical data, assumes a certain degree of stability across consecutive PB processes. Social, economic, or political changes may alter citizens’ preferences, reducing the predictive capacity of any model, whether based on classical ML or LLMs. Such stability assumption is more plausible in online PB processes in the Global North, where the stakes are often relatively low, as in the three European cities we considered. Thus, validation across different institutional and geographical contexts is one of our primary directions for future work.

We did not assess the robustness of our methods to strategic user behavior, such as using specific language in project descriptions or voting tactically based on visible vote counts. Overreliance on algorithmic predictors at the proposal stage could even encourage similar examples of strategic behavior by proposers, but also reduce their predictive power over time or undermine the expected diversity of projects.

Finally, our approach to shortlisting optimizes exclusively with respect to predicted electoral support, reinforcing a status-quo bias. An evident example is that projects that have already been funded in a previous campaign would score high, but should clearly not be shortlisted. However, algorithmic shortlisting may and should incorporate additional normative criteria, such as representational fairness or diversity preservation. Our predictions can only be used as one of the factors in a multi-objective algorithmic shortlisting mechanism, generating an initial ranking for human review as part of an AI-assisted democratic process.

Ethical Statement

Data ownership. The data collected for this paper is publicly available on Pabulib.org. The datasets including project textual information were obtained by combining voting data from Pabulib.org with project descriptions scraped from the

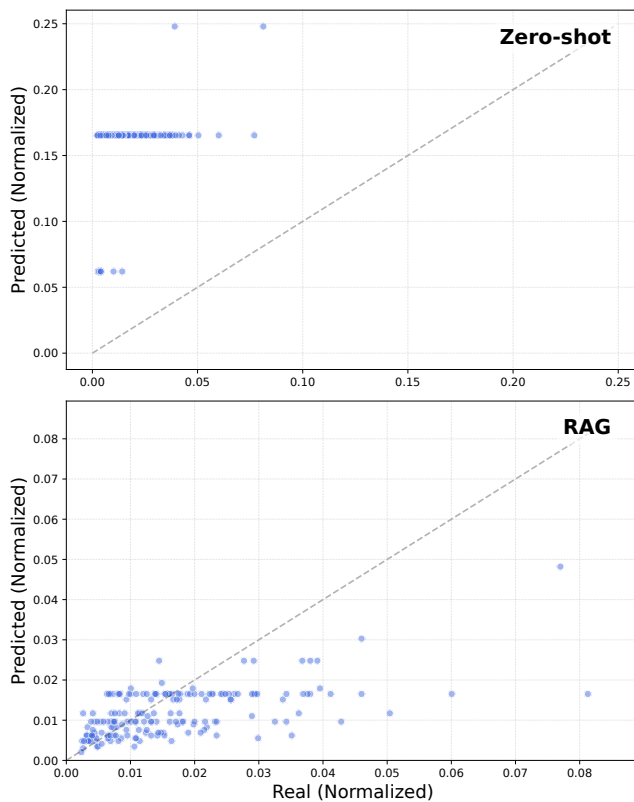


Figure 4: Predicted vs. real vote counts, normalized by number of voters for Mistral Large 3 on the Toulouse dataset. The figures compare zero-shot (top) and RAG (bottom) prompts. The dashed line is the identity line ($y = x$), indicating perfect prediction.

public webpages of the cities of Wrocław and Toulouse.

Potential risks. The use of LLMs in democratic applications raises ethical concerns, with examples such as pre-existing and documented biases of LLMs, the possibility of result manipulation, or the replacement of actual voters by AI agents at the voting phase. Our results are to be considered as an exploratory proof-of-concept which requires in-depth further scrutiny before being deployed or tested with human subjects. The use of black box ML algorithms, LLMs, or even interpretable methods that are still not accessible to the ordinary citizen is debatable. Still, if its use is accompanied by proper detailed documentation, our approach defines a shortlisting procedure that may be more transparent than some current practices, which often involve committees of volunteers selecting projects on loosely specified objectives.

Acknowledgments

This work was supported by the ANR LabEx CIMI (grant ANR-11-LABX-0040) within the French State Programme “Investissements d’Avenir.”

Funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and

do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. This work is supported by ERC grant 101166894 “Advancing Digital Democratic Innovation” (ADDI). This project was supported by the European Union LearnData, GA no. 101086712 a.k.a. 101086712-LearnDataHORIZON-WIDERA-2022-TALENTS-01 (<https://cordis.europa.eu/project/id/101086712>), IAST funding from the French National Research Agency (ANR) under grant ANR-17-EURE-0010 (Investissements d’Avenir program), and the European Lighthouse of AI for Sustainability [grant number 101120237-HORIZON-CL4-2022-HUMAN-02].



References

- Arana-Catania, M.; Lier, F.-A. V.; Procter, R.; Tkachenko, N.; He, Y.; Zubiaga, A.; and Liakata, M. 2021. Citizen participation and machine learning for a better democracy. *Digital Government: Research and Practice*, 2(3): 1–22.
- Aziz, H.; and Lee, B. E. 2021. Proportionally representative participatory budgeting with ordinal preferences. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*.
- Aziz, H.; and Shah, N. 2020. Participatory budgeting: Models and approaches. In *Pathways Between Social Science and Computational Social Science: Theories, Methods, and Interpretations*, 215–236. Springer.
- Cabannes, Y. 2004. Participatory budgeting: a significant contribution to participatory democracy. *Environment and urbanization*, 16(1): 27–46.
- Cambridge. 2025. Participatory Budgeting. <https://www.cambridgema.gov/participatorybudgeting/pbcycles/pbcycle11>. Accessed: 2026-01-09.
- Chen, J.; Xiao, S.; Peitian, Z.; Luo, K.; Pan, Z. L.; and Ke, X. 2024. BGE M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. In *Findings of the Association for Computational Linguistics (ACL)*.
- Damle, S.; and Faltings, B. 2025. LLMs for resource allocation: A participatory budgeting approach to inferring preferences. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI)*.
- Davies, J.; Arana-Catania, M.; Procter, R.; van Lier, F.-A.; and He, Y. 2021. Evaluating the application of NLP tools in mainstream participatory budgeting processes in Scotland. In *Proceedings of the 14th International Conference on Theory and Practice of Electronic Governance (ICEGOV)*.

- Faliszewski, P.; Flis, J.; Peters, D.; Pierczynski, G.; Skowron, P.; Stolicki, D.; Szufa, S.; and Talmon, N. 2023. Participatory Budgeting: Data, Tools and Analysis. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*.
- Farrell, D. M.; Curato, N.; Dryzek, J. S.; Geißel, B.; Grönlund, K.; Marien, S.; Niemeyer, S.; Pilet, J.-B.; Renwick, A.; Rose, J.; and Setälä, M. 2019. Deliberative mini publics: Core design features. *The Centre for Deliberative Democracy & Global Governance Working Paper Series*, 2019(5).
- Feng, J.; Wang, S.; Liu, T.; Xi, Y.; and Li, Y. 2025a. UrbanLLaVA: A Multi-modal Large Language Model for Urban Intelligence with Spatial Reasoning and Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- Feng, J.; Zhang, J.; Liu, T.; Zhang, X.; Ouyang, T.; Yan, J.; Du, Y.; Guo, S.; and Li, Y. 2025b. CityBench: Evaluating the capabilities of Large Language Models for urban tasks. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.
- Fish, S.; Gözl, P.; Parkes, D. C.; Procaccia, A. D.; Rusak, G.; Shapira, I.; and Wüthrich, M. 2024. Generative Social Choice. In *Proceedings of the 25th ACM Conference on Economics and Computation (EC)*.
- García-Marzá, D.; and Calvo, P. 2024. Digital twins: on algorithm-based political participation. In *Algorithmic democracy: a critical perspective based on deliberative democracy*, 61–79. Springer.
- Goel, A.; Krishnaswamy, A. K.; Sakshuwong, S.; and Aitamurto, T. 2019. Knapsack voting for participatory budgeting. *ACM Transactions on Economics and Computation (TEAC)*, 7(2): 1–27.
- Golchin, S.; and Surdeanu, M. 2024. Time Travel in LLMs: Tracing Data Contamination in Large Language Models. In *The 12th International Conference on Learning Representations (ICLR)*.
- Grandi, U. 2019. Agent-mediated social choice. In *The Future of Economic Design: The Continuing Development of a Field as Envisioned by Its Researchers*, 89–96. Springer.
- Grandi, U.; Bredereck, R.; Delemazure, T.; Endriss, U.; Maly, J.; Mattei, N.; Maudet, N.; Nardi, O.; and Szufa, S. 2026. Social Choice with Text: Collective Decision Making in the LLM Era. Preprint available at <https://hal.science/hal-05548117>.
- Grossi, D.; Hahn, U.; Mäs, M.; Nitsche, A.; Behrens, J.; Boehmer, N.; Brill, M.; Endriss, U.; Grandi, U.; Haret, A.; Heitzig, J.; Janssens, N.; Jonker, C. M.; Keijzer, M. A.; Kistner, A.; Lackner, M.; Lieben, A.; Mikhaylovskaya, A.; Murukannaiah, P. K.; Proietti, C.; Revel, M.; Rouméas, É.; Shapiro, E.; Sreedurga, G.; Swierczek, B.; Talmon, N.; Turini, P.; Terzopoulou, Z.; and Putte, F. V. D. 2024. Enabling the Digital Democratic Revival: A Research Program for Digital Democracy. *CoRR*, abs/2401.16863.
- Gudiño, J. F.; Grandi, U.; and Hidalgo, C. 2024. Large language models (LLMs) as agents for augmented democracy. *Philosophical Transactions A*, 382(2285): 20240100.
- Hagen, L. 2018. Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Information processing & management*, 54(6): 1292–1307.
- Helbing, D.; Mahajan, S.; Carpentras, D.; Menendez, M.; Pournaras, E.; Thurner, S.; Verma, T.; Arcaute, E.; Batty, M.; and Bettencourt, L. M. 2024. Co-creating the future: participatory cities and digital governance. *Philosophical Transactions A*, 382(2285): 20240113.
- Hidalgo, C. 2018. A bold idea to replace politicians. *Vancouver: TED*.
- Jigsaw-Google. 2025. Reimagining the Town Hall Meeting for the AI Age. <https://jigsaw.google/our-work/reimagining-the-town-hall-meeting/>. Accessed: 2026-01-09.
- Jones, K. S. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1): 11–21.
- Konya, A.; Thorburn, L.; Almasri, W.; Leshem, O. A.; Procaccia, A.; Schirch, L.; and Bakker, M. 2025. Using collective dialogues and AI to find common ground between Israeli and Palestinian peacebuilders. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.
- Lackner, M.; and Skowron, P. 2023. *Multi-winner voting with approval preferences*. Springer Nature.
- Lalley, S. P.; and Weyl, E. G. 2018. Quadratic voting: How mechanism design can radicalize democracy. In *AEA Papers and Proceedings*, volume 108, 33–37.
- Li, H.; De, S.; Revel, M.; Haupt, A.; Miller, B.; Coleman, K.; Baxter, J.; Saveski, M.; and Bakker, M. A. 2025. Scaling Human Judgment in Community Notes with LLMs. *Journal of Online Trust and Safety*, 3(1).
- Li, Z.; Xia, L.; Tang, J.; Xu, Y.; Shi, L.; Xia, L.; Yin, D.; and Huang, C. 2024. UrbanGPT: Spatio-temporal Large Language Models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.
- Licht, H.; Sarkar, R.; Wu, P. Y.; Goel, P.; Stoehr, N.; Ash, E.; and Hoyle, A. M. 2025. Measuring scalar constructs in social science with LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Liu, N. F.; Lin, K.; Hewitt, J.; Paranjape, A.; Bevilacqua, M.; Petroni, F.; and Liang, P. 2024. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12: 157–173.
- Mattei, N.; and Walsh, T. 2013. Preflib: A library for preferences <http://www.preflib.org>. In *International conference on algorithmic decision theory*, 259–270. Springer.
- Mitchell, W. J. 1996. *City of bits: space, place, and the infobahn*. MIT press.
- Peixoto, T.; and Fox, J. 2016. *When does ICT-enabled citizen voice lead to government responsiveness?* World Development Indicators Washington, D.C. : World Bank Group.
- Peters, D.; Pierczyński, G.; and Skowron, P. 2021. Proportional participatory budgeting with additive utilities.

Advances in Neural Information Processing Systems, 34: 12726–12737.

Rey, S.; Endriss, U.; and de Haan, R. 2021. Shortlisting Rules and Incentives in an End-to-End Model for Participatory Budgeting. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*.

Rey, S.; Schmidt, F.; and Maly, J. 2025. The (Computational) Social Choice Take on Indivisible Participatory Budgeting. *arXiv:2303.00621*.

Schmidt, F. 2024. *Cost Bias in Participatory Budgeting*. Ph.D. thesis, Technische Universität Wien.

Shah, A. 2007. *Participatory budgeting*. World Bank Publications.

Small, C. T.; Vendrov, I.; Durmus, E.; Homaei, H.; Barry, E.; Cornebise, J.; Suzman, T.; Ganguli, D.; and Megill, C. 2023. Opportunities and risks of LLMs for scalable deliberation with Polis. *arXiv preprint arXiv:2306.11932*.

Smith, G. 2009. *Democratic Innovations: Designing Institutions for Citizen Participation*. Cambridge University Press.

Talmon, N.; and Faliszewski, P. 2019. A Framework for Approval-Based Budgeting Methods. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*.

Tessler, M. H.; Bakker, M. A.; Jarrett, D.; Sheahan, H.; Chadwick, M. J.; Koster, R.; Evans, G.; Campbell-Gillingham, L.; Collins, T.; Parkes, D. C.; et al. 2024. AI can help humans find common ground in democratic deliberation. *Science*, 386(6719).

Wampler, B. 2000. *A guide to participatory budgeting*. International Budget Partnership Washington, DC, USA.

Wampler, B.; McNulty, S.; and Touchton, M. 2021. *Participatory Budgeting in Global Perspective*. Oxford University Press.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Yang, J. C.; Dalisan, D.; Korecki, M.; Hausladen, C. I.; and Helbing, D. 2024a. LLM Voting: Human Choices and AI Collective Decision-Making. In *Proceedings of the 7th AAAI/ACM Conference on AI, Ethics, and Society (AIIES)*.

Yang, J. C.; Hausladen, C. I.; Peters, D.; Pournaras, E.; Hnggli Fricker, R.; and Helbing, D. 2024b. Designing digital voting systems for citizens: Achieving fairness and legitimacy in participatory budgeting. *Digital Government: Research and Practice*, 5(3): 1–30.

Zambrano, J.; Contet, C.; Gudiño-Rosero, J.; Garrido-Lucero, F.; Grandi, U.; and Hidalgo, C. 2026. Algorithmic Shortlisting in Participatory Budgeting. *arXiv preprint arXiv:2508.06577*.