

Use of LLMs for Generating Synthetic ARDs to Make Estimates in Electoral Processes

Lucia Soria Gonzalez

IMDEA Software Institute

Milene Gutierrez Santibanez

University of Chile

Sergio Diaz

IMDEA Networks

Juan Marcos Ramirez

IMDEA Networks

Rosa Lillo

Universidad Carlos III de Madrid

Antonio Fernandez Anta

IMDEA Software Institute

Jose Aguilar

`jose.aguilar@imdea.org`

IMDEA Software Institute

Research Article

Keywords: Indirect surveys, Network Scale-Up Methods, Large language models, Synthetic data

Posted Date: August 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10534310/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Use of LLMs for Generating Synthetic ARDs to Make Estimates in Electoral Processes

Lucia Soria González ^{1,4}, Milene Gutiérrez Santibáñez ²,
Sergio Díaz ³, Juan Marcos Ramírez³,
Rosa Lillo ⁴, Antonio Fernández Anta¹, Jose Aguilar^{1,5},

¹IMDEA Software Institute, Madrid, Spain,

²Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Chile

³IMDEA Networks Institute, Madrid, Spain,

⁴Universidad Carlos III, Madrid, Spain,

⁵CEMISID, Universidad de Los Andes, Mérida, Venezuela,

100497666@alumnos.uc3m.es, milenegutierrez@gmail.com, sergio.diaz@imdea.org

juan.ramirez@imdea.org, lillo@est-econ.uc3m.es,

antonio.fernandez@imdea.org, jose.aguilar@imdea.org

August 1, 2026

Abstract

Context: Accurately estimating electoral results remains a challenge in political science due to response biases in traditional direct surveys. Indirect survey techniques, which infer voting intentions from respondents’ social networks, have emerged as a promising alternative by reducing response bias and improving scalability. In parallel, large language models (LLMs) have shown a remarkable ability to capture and reproduce social patterns, suggesting their potential use as synthetic respondents in survey-based research. While indirect survey techniques—asking about the voting intentions of acquaintances—mitigate the biases in traditional direct surveys, the emergence of LLMs offers a novel opportunity to simulate these processes using synthetic agents.

Objective: This study aims to develop a robust methodological framework to employ LLMs as artificial respondents in indirect electoral surveys, assessing their ability to reflect collective social behavior.

Method: To this end, LLMs are prompted to answer indirect survey questions regarding perceived voting intentions within social circles, and the resulting synthetic data are processed using Network Scale-Up Methods (NSUM).

Specifically, this research proposes a methodology using specific prompt templates to elicit synthetic responses from an LLM. These responses, focused on perceived voting intentions within social circles, are processed using NSUM.

Results: This study provides methodological guidelines and prompt templates that support the integration of LLMs into electoral analysis frameworks. The methodology is applied to two empirical cases: the latest regional elections in Spain and recent national elections in Chile. The study systematically compares different LLM versions, deployment environments, and demographic configurations (age, sex), and contrasts LLM-based estimates with results from real-world indirect surveys. The results show that NSUM-based estimates derived from LLM-generated surveys reproduce key aggregate patterns observed in empirical indirect survey data, although variability across models and settings is observed.

Findings: The main finding of this work is that LLMs can infer a collective pattern of electoral behavior, offering a scalable and reproducible complement to traditional indirect surveys, positioning LLMs as a powerful complementary tool for modern political analysis and social pattern estimation. The results demonstrate that the synthetic data generated by LLM can effectively replicate voting behaviors comparable to those obtained through indirect human surveys.

Keywords:

Indirect surveys, Network Scale-Up Methods, Large language models, Synthetic data.

1 Introduction

Accurately estimating the results of electoral processes remains a challenge in political science. In particular, traditional direct surveys often struggle to capture truthful responses, leading to systematic errors. Indirect survey techniques, based on questions about how respondents' acquaintances will vote, have emerged as promising alternatives by reducing response bias, maintaining anonymity, and being more scalable by leveraging social networks.

At the same time, large language models (LLMs) have demonstrated an exceptional capacity to learn and reproduce complex patterns from a vast corpus of text, including content on political analysis, collective narratives, and social behaviors. Beyond their conventional use as text generators or analytical tools, LLMs can be seen as synthetic agents capable of simulating human-like responses under controlled conditions. This opens up new methodological possibilities for their use as artificial respondents.

Using LLMs as survey participants offers several advantages for electoral analysis. First, survey-based LLMs can be surveyed on a large scale, at low cost, and without the logistical and ethical limitations associated with human subjects. Second, they provide a reproducible and controllable environment,

allowing for the exploration of different indirect questioning strategies under varying assumptions and contexts, such as specific demographic analyses. This is particularly valuable for developing a methodological framework for using LLMs as respondents, modeling the voter rather than the individual voter’s behavior.

In this work, we develop a methodology for employing LLMs as respondents in indirect election surveys focused on perceived voting intentions within respondents’ social circles. The collected synthetic responses are then processed using Network Scaling Methods (NSUM) to estimate aggregate election results. This approach proposes prompts for designing established indirect survey processes, analyzes different LLM environments (app and online), different LLM versions, and demographic analyses (age, sex, etc.) to directly compare synthetic data generated by LLMs with real-world indirect survey data.

Specifically, this work studies the cases of the most recent regional elections in Spain, which were estimated using human indirect surveys, and the last elections in Chile. To this end, we compare NSUM-based estimates derived from LLM-generated surveys with those obtained from actual indirect surveys. In doing so, we assess the extent to which LLMs can serve as significant indicators of collective social knowledge in electoral contexts and explore their potential as complementary tools for electoral analysis, developing a robust methodological framework for conducting electoral surveys. The main contributions are the following:

- The development of a methodology for using LLMs as respondents in indirect surveys for electoral processes.
- The definition of prompt templates for generating synthetic data from indirect respondents.
- A comprehensive analysis of the behavior of synthetic data generated from LLM responses compared to human responses and real data.

This work is organized as follows. Section two describes works related to our own, first, those that have used indirect surveys to estimate the results of electoral processes, and second, those that have used LLMs to generate synthetic surveys. Next, we present the basic theoretical framework, introducing the NSUM and LLMs used. Following this, we describe the methodological framework proposed in this work, which consists of three steps: Prompt design, generation of synthetic data using the prompts in the LLMs, and estimation of voting intention using NSUM techniques. Next, we present the results for the last elections in Spain and Chile, comparing the results with both the actual results and indirect human surveys. In addition, we carry out a general discussion of the results. Finally, we present conclusions and directions for future work.

2 Related work

2.1 Indirect surveys and NSUM in election contexts

Indirect surveys have emerged as an effective strategy for rapidly and anonymously collecting information, which can then be used to estimate the size of hidden subpopulations. To infer the size of these unknown groups, it is essential to gather data about the contact networks of the respondents. The foundational idea behind the NSUM was introduced by Bernard et al. [JKR89], who proposed using aggregated relational data (ARD) to estimate each participant’s network size (degree) and then leveraging this information to infer the size of the hidden subpopulation. Subsequent studies have expanded on this approach [KJM⁺98, HDK15]. Particularly important is the estimator developed by Killworth et al. [KMB⁺98], which has become the most widely used NSUM estimation method. NSUM methods have been applied in epidemiology [FUM⁺16, TPC⁺19], as well as in research on health and social issues such as cancer [CSMG17], abortion [SGAD20], sex work [JLC⁺18], and drug use [SFM⁺10].

On the other hand, a number of studies have focused on improving the NSUM estimation process itself. New Bayesian models have been introduced to correct for various sources of bias [MZ12, TPC⁺19] and to account for correlations between groups [LBN23]. Other works incorporate multiple estimates from different sources of information to refine existing ones [PNB22]. Additionally, ARD has been employed to reconstruct networks [BCMP20], estimate network properties [BCL⁺23], and derive latent network representations [MZ15]. Two review articles on NSUM and ARD [LBN21, McC20] summarize most of the approaches currently available in the literature.

In the context of electoral surveys, several studies have examined the use of indirect surveys—particularly surveys that ask about the broader population or respondents’ immediate social circles—across a range of elections in different countries. For instance, Murr [Mur11] investigated the “wisdom of crowds” in election forecasting. His approach relies on citizen predictions of constituency-level results. Using data from the 2010 British Election Study, the analysis showed that the size of the reference group plays a crucial role in prediction accuracy. Although this method can directly forecast seat shares, it can sometimes lead to errors as high as 15%. Moreover, the approach does not address the question of how large a respondent sample is needed for reliable forecasts, nor does it account for the influence of electoral campaigns on respondents’ expectations. In another line of work, Galesic et al. [GBdBD⁺18] demonstrated that social-circle questions—those asking about the voting intentions of respondents’ acquaintances—yield improved forecasts. These questions successfully predicted the outcomes of the 2016 U.S. and 2017 French presidential elections at national, state, and individual levels. They also helped capture last-minute shifts in voting intentions and provided insight into the structure of echo chambers among supporters of different candidates. Although the article details the survey design (including multiple waves and different types of questions regard-

ing own voting intention, social-circle intention, and expected election winner), it does not provide information on how the survey responses were analyzed or how predictions were computed. Graefe [Gra14] compared the accuracy of vote-expectation surveys for U.S. presidential elections from 1932 to 2012 against four established forecasting methods: expert judgment, traditional polls, prediction markets, and quantitative models. In a separate study, Graefe [Gra16] evaluated similar surveys in the context of the 2013 German federal election, concluding that representative expectation surveys offer a valuable and low-cost forecasting tool—particularly for small-scale elections where conventional polling or modeling approaches may not be practical or cost-effective. However, a key concern with expectation-based forecasting is that individuals’ predictions may be influenced by their own preferences; voters tend to expect their favored candidate to win. While this approach draws on Condorcet’s Jury Theorem—according to which the probability of a group reaching a correct decision increases with the number of independent estimates—such preference-driven bias is not necessarily eliminated, unless the sample is well balanced across candidate preferences.

Recent work by Bruine de Bruin et al. [BdBGB⁺22] found that social-circle questions outperform respondents’ own voting-intention reports when predicting elections involving many political parties. Their findings, based on the 2017 general election in the Netherlands and the 2018 general election in Sweden, suggest that social-circle questions remain effective even in complex multiparty systems. Nonetheless, the study does not provide details on sample characteristics, contextual factors, or the specific analytical procedures used to obtain these results. Finally, the work [ARDA⁺25] introduces a methodology for electoral forecasting that relies on indirect questioning and network scale-up methods. The proposed approach gathers information not about respondents’ own voting intentions, but about the voting intentions of people within their social networks. After designing an indirect survey to capture these network-based perceptions, the method applies data-cleaning filters and sample-stratification strategies, and ultimately uses NSUM techniques to estimate the vote share for each political group. The methodology is tested using data from the 2023 Spanish general elections, focusing on the regions of Madrid, Andalusia, and Valencia. Results show that the proposed approach produces estimates that are closer to the actual election outcomes than those published by major public and private polling organizations.

2.2 LLMs in surveys

Recent advances in LLM have increased interest in its use for generating synthetic data. One such possibility is generating synthetic survey data as an alternative to human responses. Below, we present some recent works where LLMs have been applied to this task in different domains.

The work by Kim and Lee [KL24] proposed a method for the use of LLMs as synthetic respondents in direct surveys, fine-tuning LLMs with the responses to real questions from surveys previously administered to human respondents. Their study analyzes several tasks inherent to the survey context, including

missing data imputation, retrodiction, and zero-shot prediction, within a unified methodological framework that conditions model responses on survey questions. The results show that LLMs can reproduce human response patterns with high accuracy in scenarios where comparable human data are available, particularly for missing data imputation and the reconstruction of past opinions. In contrast, performance decreases substantially when models are required to answer entirely new questions for which no prior human observations exist. Of particular relevance for the use of LLMs as survey respondents, the authors also document systematic variation in accuracy across population groups, with lower performance for individuals with lower socioeconomic status, non-partisan political affiliations, and racial minority backgrounds. These results suggest that LLMs reproduce not only aggregate opinion patterns but also representational gaps present in the underlying survey data. As a result, they argue that LLMs cannot replace human respondents, but can instead complement existing data through imputations and historical reconstructions.

Motoki et al. [MMMR24] explored the potential of LLMs to generate synthetic survey data for organizational research. They introduced a structured prompting framework that incorporates persona creation and impulse variables to simulate realistic human responses and increase variability. Using validated constructs from organizational deviance literature, the study evaluated whether synthetic responses align with theoretical expectations. Results indicate that LLMs can generate structured data resembling real-world patterns, supporting efficient pre-testing and survey refinement. However, challenges remain, including response homogeneity, inflated reliability estimates, and model-specific biases. Huang et al [HWW25] examined how LLMs can simulate human survey responses while incorporating uncertainty quantification to produce reliable population estimates. The proposed method transformed imperfect synthetic responses into confidence sets that account for distribution shifts between simulated and real populations. A central contribution is adaptively determining the optimal number of simulated responses, balancing overly narrow intervals with poor coverage against excessively wide estimates.

Other studies have suggested the possibility of using LLMs to predict public opinion by mimicking human responses. For example, Aher et al. [AAK23] carried out a Turing Experiment about how language models simulate human behavior using representative samples rather than individuals. The method replicates classic studies in economics, psycholinguistics, and social psychology. Results shown successful replication but also revealed distortions like “hyper-accuracy”, which may impact real-world applications, including education, decision making, and creative fields. Argyle et al. [ABF⁺23] argued that LLMs can act as proxies for human subpopulations in social research. Using GPT-3, it shown biases are demographically structured and conditionable, enabling accurate simulation of diverse groups. This “algorithmic fidelity” produced nuanced, realistic responses, suggesting language models can be powerful tools for studying human attitudes and behavior. Dillion et al. [DTGG23] examined whether LLMs can replace human participants in psychological studies, reviewed emerging evidence, proposed a theoretical framework, and highlighted impor-

tant limitations, risks, and conditions under which LLMs may or may not serve as reliable experimental subjects. Also, Mei et al. [MXYJ24] evaluated LLMs behavior using a Turing-style test, games, and personality surveys. ChatGPT-4 shown human-like traits across trust, fairness, and cooperation, adapting behavior as if learning. Its actions fell within human distributions but tend toward greater altruism and cooperation, suggesting LLMs can approximate human behavior in diverse social and psychological contexts.

Suhyun et al. [CKK24] examined whether LLMs can infer an individual’s opinions from their speech and subsequently respond in ways that reflect those learned views. They also explored how such models could be applied to survey research. The study introduced the concept of a “Surveyed LLM,” a model specifically trained for survey tasks to replicate previously learned human responses. Villarreal-Zegarra et al. [VZBB26] evaluated the consistency and psychometric plausibility of synthetic data generated by an LLM to simulate responses in Peru’s ENSUSALUD 2016 national health survey. Using GPT-OSS-20B, researchers produced Spanish-language responses conditioned on sociodemographic and clinical profiles. Synthetic data showed strong agreement with key characteristics such as sex, age, and chronic disease status. Depression and anxiety scales (PHQ-9, PHQ-8, GAD-7) demonstrated good factor fit and high internal consistency, with meaningful correlations to real measures. However, deviations emerged, including overestimation of obesity and irregular distributions, highlighting the need for rigorous validation before applying synthetic data to research or policy decisions. Also, Kaiser et al. [KKM⁺25] developed methods for synthetic survey sampling using ASPIRE, a system that creates digital twins based on demographic profiles and generates responses through LLMs. Comparing synthetic data to a representative U.S. sample ($n = 461$) on soft drink preferences, results showed better-than-chance accuracy and similar aggregate rankings. However, synthetic responses overestimated positive ratings and showed lower variance. Jansen et al. [JgJS23] discussed the use of LLMs to generate responses to survey items. They concluded that LLMs can address some of the challenges associated with survey research regarding response bias, lack of clarity and understanding, but cannot yet correct for sampling or non-response bias challenges.

For elections, Bisbee et al. [BCD⁺24] examined how accurately ChatGPT can replicate public opinion by prompting it to adopt various personas and assign feeling thermometer scores to 11 sociopolitical groups. The average scores produced by ChatGPT closely match those from the 2016–2020 American National Election Study (ANES). However, its sampling performance was unreliable for statistical inference. The responses showed less variation than real survey data, and regression coefficients often differed significantly from ANES estimates. These results raise serious concerns about the quality, reliability, and reproducibility of synthetic survey data generated by LLMs.

As can be seen from the reviewed work, there are very few studies that have used LLMs in an electoral context to simulate voter responses. Furthermore, there are no previous studies that simulate responses in indirect surveys, making our work the first to our knowledge.

3 Background

3.1 NSUM

Collecting social data is often infeasible due to legal constraints and the population’s reluctance to disclose sensitive information. This is especially relevant in survey-based studies, where participants may not be willing to participate or respond honestly, introducing biases in the data and subsequently affecting the results of the studies.

NSUM is a relatively new technique [JKR89, LBN21] designed to estimate unknown groups of a social network without direct information from the participants. NSUM relies on the aggregated relational data (ARD), which consists of the number of contacts that participants know belonging to certain groups. This is collected by questions of the type ”How many people do you know with a university degree?”. To infer the sizes of the unknown groups, this information is scaled in different ways with the participants’ personal network size, usually denoted as the *degree* of the participants, following conventions of graph theory. This degree is collected by directly asking respondents about their personal network size or by estimating it from the ARD of other groups [DAAR⁺25].

NSUM does not require information about the sampled nodes of the social network. This is especially relevant in survey scenarios, since this approach increases the participation and honesty of the responses, thereby improving overall data quality and conclusions. Additionally, every sample individual reports information about many acquaintances, obtaining much more information about the network. In particular, the use of ARD has been shown to reduce the cost of more than 70% in socio-economic networks while yielding similar conclusions [BCMP20]. For these advantages, NSUM has been a tool applied to estimate sensitive populations such as sex workers [JLC⁺18], reproductive health [SGAD20], COVID-19 [GAOHR⁺21], and trafficking in child sexual exploitation material [NAMS⁺26]. It is also employed by reputable organizations such as the WHO to estimate HIV infections [UNA10]. In addition, NSUM has been applied to areas other than surveys, such as preferences on online social networks [DAAR⁺25] and IPFS characterization [RDABF26].

An emerging application of the NSUM is the estimation of voting intention. Arevalillo et al. [ARDA⁺25] applied several NSUM estimators to predict the outcomes of multiple elections in Spain using surveys conducted one week before the elections. Their study showed that the NSUM estimates were close to the actual results and, despite relying on much smaller samples, showed better performance than the main private polling organizations and the official pollsters with substantially greater resources.

3.2 LLMs

In this subsection, we provide a concise overview of the large language models (LLMs) employed in this study. For each model, we summarize its release timeline and highlight its main architectural and functional characteristics that are

relevant to our experimental design. This background is intended to clarify the technological differences between the models and to support the interpretation of the results presented in the subsequent sections.

- **GPT-4 (OpenAI):** GPT-4 was released by OpenAI on March 14, 2023 [Ope23] as the successor to GPT-3.5, representing a significant advance in large language model capabilities. GPT-4 is a multimodal model that can process both text and images, allowing it to perform tasks such as image description, diagram interpretation, and reasoning over mixed input formats [BD23]. It also introduced much larger context windows, enabling the model to handle longer inputs and maintain coherence over extended interactions. GPT-4 was designed to be more reliable, creative, and better at following nuanced instructions, involving iterative safety improvements and enhanced control mechanisms [Ope23].
- **GPT-4o (OpenAI):** GPT-4o (where “o” stands for “omni”) was announced and released by OpenAI on May 2024 [Ope24] as the next flagship in the GPT-4 family. GPT-4o extends GPT-4’s capabilities by natively integrating text, image, and audio modalities within a single model, allowing real-time reasoning across these input types [Ope24]. It supports voice-to-voice interaction without relying on external sub-models and features a context window of up to 128K tokens. GPT-4o also delivers faster response times and enhanced efficiency, emphasizing natural conversational interaction and broader accessibility.
- **GPT-5 (OpenAI):** The official release of GPT-5 on August 2025, introduced a unified ecosystem encompassing Standard, Mini, and Nano variants, representing a groundbreaking advancement in adaptive model architecture [Leo25]. This generation is defined by its adaptive reasoning engine, a system capable of determining within milliseconds whether to provide an instantaneous response or engage a sophisticated “Thinking” mode for complex problem-solving. This trajectory was further refined with the deployment of GPT-5.1 in November 2025, which optimized autonomous task routing to maximize both computational efficiency and precision [SMS⁺26]. Technically, GPT-5 features a substantial context window of up to 400,000 tokens and a redesigned tokenization process that significantly outperforms its predecessors in information density [Leo25]. The model was trained in high-performance NVIDIA Blackwell infrastructure with a knowledge cutoff of September 2024; a rigorous emphasis on data curation resulted in a 45% reduction in hallucinations compared to previous generations, reinforcing its reliability in critical sectors such as healthcare [Leo25]. Additionally, the refinements introduced alongside the 5.1 update included eight distinct personality presets, ensuring a more nuanced and human-centric interaction model than earlier system message frameworks ¹.

¹For further details regarding the training and specific characteristics of each model, see: <https://openai.com/es-ES/news/>

4 Methodology

Our method consists of using LLMs as generators of synthetic data in indirect surveys, which are based on questions that take the form: “From your close social circle, how many people do you know who voted for this candidate or party?”. In this context, LLMs act exclusively as simulated respondents, replacing the data collection stage that is traditionally carried out through surveys administered to real individuals.

In particular, the models are not used to directly estimate voting intention or to perform inference of the population. Their sole role is to produce individual responses that emulate the behavior of a human respondent who, based on information from their close social circle, reports the number of acquaintances who voted for a given candidate or party. These responses constitute the ARD used as input for the NSUM estimators.

The methodology is structured in the following steps:

1. Design of the prompt for LLMs.
2. Querying the LLM and generation of synthetic data (synthetic ARD).
3. Estimation of voting intention using NSUM techniques.

In what follows, the steps of the methodology are described in detail.

4.1 Step 1: Design of the prompt for LLMs.

The first step of the methodology consists of designing the prompt, understood as the input text that defines the context, the instructions, the questions, and the expected output format that will later be used to query the LLMs.

The prompt is designed so that the model adopts the role of a single respondent and limits its behavior to answering an indirect survey question. In particular, the model is instructed to respond as an individual who reports information based exclusively on their close social circle, without relying on global knowledge, real electoral outcomes, or population-level inference.

To increase the realism of the generated responses, each prompt instance includes a demographic profile associated with the synthetic respondent. This profile includes the country of residence, age, and, in some cases, income level. These attributes are assigned proportionally according to the available census information for each country considered in the study. The text of the prompt is identical across all queries, with only the demographic attributes varying between synthetic respondents.

Two variants of the prompt are considered. The first is a standard prompt that mirrors the wording used in indirect surveys administered to human respondents. The second, referred to as the Perfect Prompt, is a refined version of the standard prompt obtained using an automatic instruction optimization

tool for language models.² In both cases, the core question asks for the number of individuals in the respondent’s close social circle who voted for each candidate or party.

Finally, the prompt explicitly enforces that the model output consists only of numerical values, with no explanations or additional text. This restriction is imposed directly on the input to ensure immediate compatibility with the NSUM estimators and to reduce the occurrence of unstructured or inconsistent responses. No explicit size is imposed on the close social circle, leaving this aspect to be implicitly determined by the model in each response. A complete example of both prompts used in this study is provided in the Appendix (see A.2.3 and A.2.4).

4.2 Step 2: Querying the LLM and generation of synthetic data (synthetic ARD)

The second step of the methodology consists of querying the language models using the prompts defined in Step 1, with the objective of generating the synthetic responses that constitute the ARD used in the subsequent analysis. Each query corresponds to a single synthetic respondent and is treated as an independent observation.

Two separate channels were used to query the language models: programmatic access via the API and interaction through the standard conversational web interface. These two channels are treated as independent experimental settings and are analyzed separately.

4.2.1 API

The queries to the language models were conducted through programmatic API calls using Python. In each call, the model receives as input the corresponding prompt together with the demographic information associated with the synthetic respondent. Each API call is treated as an independent respondent answering the full set of survey questions.

Different models from the GPT family were employed, including GPT-4, GPT-4o-mini, and GPT-5. For each respondent, the complete set of survey questions constitutes the input, and the output consists exclusively of numerical values associated with each candidate or party; this is a parameter specified in the query format. The generated responses are stored grouped by survey round (200 respondents), which allows for subsequent processing and their direct use as the input dataset for the NSUM estimators.

²We used PromptPerfect (<https://promptperfect.jina.ai/>), an AI-based tool designed to improve the clarity, precision, and consistency of prompts used to query large language models.

4.2.2 Online

In this channel, a parallel set of queries was administered through the standard conversational web interface. To ensure methodological rigor and prevent data contamination, each simulation was conducted within isolated "Projects" or workspaces. This structural isolation ensures that the model does not engage in cross-session learning, thereby maintaining the independence of each synthetic respondent. Within these projects, every query was initiated as a new, single-turn conversation consisting of a single prompt and numerical response in order to replicate the discrete nature of a human survey participant.

In line with the selection criteria used for the API, several iterations of the GPT family were employed, including GPT-4o-mini, GPT-5, and GPT-5.1. Due to the rolling update cycle of web-based LLM platforms, data collection was aligned with the latest available releases, transitioning the methodology to each current frontier model as it became available to ensure the use of the highest available reasoning capabilities.

4.3 Step 3: Estimation of voting intention using NSUM techniques

NSUM estimates the proportion of a group in a social network, using the responses of the synthetic participants about the number of contacts in several groups that share a common characteristic. There are different methods to estimate the size or proportion of a group, see [DAAR⁺25] for a simulated evaluation. In this work, we will restrict ourselves to the NSUM estimators that have been employed to predict the voting intention presented in the study of Arevalillo et al. [ARDA⁺25].

Let $\{1, \dots, Q\}$ be a set of indexes representing different mutually exclusive groups in a social network, and p_u be the proportion of the group $u \in \{1, \dots, Q\}$ within the network. The NSUM estimators estimate the proportion of these groups, which are denoted as *unknown groups*, using the ARD. In our case, we use the ARD of the synthetic participants. Concretely, the synthetic participant i reports the number of contacts it knows in the group u , which we denote as y_{iu} . NSUM uses the degree (personal network size) of synthetic users, d_i , to estimate the proportion of unknown groups. An option is to utilize the fact that the unknown groups are mutually exclusive, adding up all the collected ARD of the unknown groups for each participant i .

$$\hat{d}_i^{(1)} = \sum_{u=1}^Q y_{iu}. \quad (1)$$

Traditional NSUM estimators estimate the values d_i with the ARD of other groups whose proportion p_k is known for $k = 1, \dots, L$, which are denoted as the *known groups*. The value of the ARD of the participant i for the group k is

denoted as y_{ik} . One option is to take the average of the ratios y_{ik}/p_k

$$\hat{d}_i^{(2)} = \frac{1}{L} \sum_{k=1}^L \frac{y_{ik}}{p_k}. \quad (2)$$

Another method assumes that $y_{ik} \sim \text{Bin}(d_i, p_k)$, and estimate d_i with a Maximum Likelihood estimation

$$\hat{d}_i^{(3)} = \frac{\sum_{k=1}^L y_{ik}}{\sum_{k=1}^L p_k}. \quad (3)$$

For the estimation of the unknown groups, the first strategy uses only the information obtained for the ARD of the unknown groups, estimating the degree by summing the ARD of the groups, to obtain the expression

$$\hat{p}_u^{Naive} = \frac{1}{n} \sum_{i=1}^n \frac{y_{iu}}{\hat{d}_i^{(1)}} \quad (4)$$

Other alternatives with estimated degrees with the ARD of the known groups are the Mean of Sums (MoS) [HDK15]

$$\hat{p}_u^{MoS} = \frac{1}{n} \sum_{i=1}^n \frac{y_{iu}}{\hat{d}_i^{(2)}} = \frac{1}{n} \sum_{i=1}^n \frac{y_{iu}}{\frac{1}{L} \sum_{k=1}^L \frac{y_{ik}}{p_k}}, \quad (5)$$

the Plug-in MLE (PiMLE) [KJM⁺98]

$$\hat{p}_u^{PiMLE} = \frac{1}{n} \sum_{i=1}^n \frac{y_{iu}}{\hat{d}_i^{(3)}} = \frac{1}{n} \sum_{i=1}^n \frac{y_{iu}}{\frac{\sum_{k=1}^L y_{ik}}{\sum_{k=1}^L p_k}}, \quad (6)$$

and the estimator based on the assumption that y_{iu} are $\text{Bin}(d_i, p_u)$ independent random variables, Maximum Likelihood Estimator (MLE) [KMB⁺98], which uses a maximum likelihood estimation to obtain the following expression

$$\hat{p}_u^{MLE} = \frac{\sum_{i=1}^n y_{iu}}{\sum_{i=1}^n \hat{d}_i^{(3)}} = \left(\sum_{k=1}^L p_k \right) \frac{\sum_{i=1}^n y_{iu}}{\sum_{i=1}^n \sum_{k=1}^L y_{ik}}. \quad (7)$$

5 Experiments

5.1 Protocol

This section presents all the aspects required for experimentation. Specifically, the LLM versions used, the performance metrics calculated to determine the quality of our approach, and finally, the details of the test scenarios used (Spanish and Chilean general elections).

5.1.1 Used Language Models

In this study, we employed LLMs as synthetic respondents to generate voting-intention data through indirect survey methods and to evaluate whether model-generated answers can reproduce the patterns observed in indirect surveys applied to humans. All interactions with LLMs followed a standardized protocol designed to ensure consistency, reproducibility, and control over variation across models. The models used were:

- **GPT-4 (OpenAI):** used exclusively via API.
- **GPT-4o (OpenAI):** accessed both through the web interface and via API.
- **GPT-5 (OpenAI):** accessed both through the web interface and via API.

The main difference between using these models through the API and interacting with them through the chat interface rests primarily on how we communicate with the LLM rather than the difference in the model itself. Due to the characteristics of our experiment, we suppressed the conversational framework that differentiates the online version by only interacting once with the LLM and extracting its response. However, this means that the model’s behavior in the online version is therefore shaped not only by its internal features but also by conversation-structuring heuristics that influence how prompts are interpreted and how responses are generated.

5.1.2 Performance Metrics

In order to analyze the accuracy of the estimations, the following metrics were used: Mean Squared Error (MSE) and the Root Mean Squared Error (RMSE).

The Mean Squared Error (MSE) is a statistical measure that evaluates the average of the squares of the errors. More specifically, the average squared difference between the actual observed values (y_i) and the predicted or estimated values ($\hat{y}_i$). Mathematically, it is expressed as:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

This metric was utilized to quantify the variance and bias within the model’s predictions. The primary reason for employing MSE is its property of squaring the error terms, which effectively penalizes larger errors more severely than smaller ones. By giving higher weight to significant deviations, MSE ensures that the model is sensitive to outliers, helping to identify whether the estimations are consistently close to the true values or if they suffer from occasional large inaccuracies.

Building upon this, the Root Mean Squared Error (RMSE) was employed to provide a more interpretable measure of accuracy. It is calculated by taking the square root of the MSE, denoted as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

The RMSE represents how spread out the prediction errors are. While MSE is mathematically advantageous for optimization, its units are squared, which can make direct comparison with the original data difficult. By applying the square root, RMSE maintains the original units of the variables, allowing a more intuitive assessment of the model’s performance.

These metrics were applied to assess the accuracy of NSUM estimations across all data samples, whether generated through simulation or obtained via standard empirical methods. For each round of surveys, we applied the four NSUM estimators. For every iteration of these estimators, we calculated both types of errors, thereby obtaining a standardized measure for all surveys regardless of the model utilized, the prompt administered, or the acquisition channel.

5.1.3 Scenarios

- **Spanish elections**

We evaluate the performance of LLM-based indirect surveys using the database of indirect surveys of Arevalillo et al. [ARDA⁺25], collected for the Spanish general elections held on July 23, 2023. This dataset contains the responses of three of the most populated regions in Spain. Notice that Spain has a parliamentary electoral system in which seats in the national legislature are allocated according to election results within each of the 17 regions.

- **Survey rounds.**

- * **Human surveys (2023):** We use the indirect surveys collected prior to the Spanish general elections held in July 2023, originally reported in [ARDA⁺25]. These surveys were conducted during the pre-electoral period, capturing voting intentions in a real electoral context.
- * **LLM surveys (2025):** The surveys generated using large language models (ChatGPT in chat modality and GPT API) were conducted later, in 2025. However, to mitigate temporal inconsistency in data collection time with the original electoral scenario, the models were explicitly instructed to respond under the context of July 2023, that is, the period preceding the Spanish general elections. In particular, the prompts included explicit instructions requiring the models to situate their responses within that temporal framework.

- **Data sources.**

1. **Human respondents.** For each of the three regions, the dataset consists of responses from 200 participants regarding the number of close contacts, including family members, friends, and coworkers, who intend to vote for a particular political party. In addition, each dataset includes demographic information, such as age groups and unemployment status, about the reported contacts, which can be used as control variables required by various NSUM estimators. More specifically, each dataset contains information on 12 variables: six related to voting intention, three corresponding to predefined age groups (18-34, 35-54, and older than 54), one indicating unemployment status, one identifying self-employed workers, and one indicating the number of contacts that are medical doctors [ARDA⁺25]. The English version of the questionnaire is provided in the Supplementary Material, Section A.1.5.

Note that the datasets were collected using a stratified sampling design based on age. Specifically, predefined age groups were constructed based on population proportions reported by the Spanish National Statistics Institute (INE, Instituto Nacional de Estadística), thereby ensuring demographic representativeness across age strata. Finally, to mitigate the influence of extreme or potentially erroneous observations, we employ a robust outlier-filtering procedure based on the Median Absolute Deviation (MAD). This filtering criterion is applied independently to each column (i.e., each variable) of the dataset. For a given variable $\mathbf{x} = \{x_1, \dots, x_n\}$, we first compute its median $\tilde{x}$. The MAD is then defined as

$$\text{MAD} = b \cdot \text{median}(|x_i - \tilde{x}|_{i=1}^n) \quad (8)$$

where $b = 1.4826$ is a constant associated with the assumption that the variable is normally distributed [LLK⁺13]. An observation x_i is classified as an outlier for the corresponding variable if it falls outside the interval

$$[\tilde{x} - 3 \cdot \text{MAD}, \tilde{x} + 3 \cdot \text{MAD}].$$

This procedure is applied to all variables in the dataset. A row (i.e., an individual response) is removed if at least one of its entries is identified as an outlier according to the above criterion in its corresponding column. Compared to the inter-quartile range (IQR) rule, MAD-based filtering is substantially more robust to skewed distributions and heavy-tailed noise. This property makes it particularly well-suited to survey data and count variables, such as those commonly encountered in NSUM applications.

2. **ChatGPT (chat modality):** responses generated under two prompts with different questionnaires:

- (a) an *original prompt*, identical to the questionnaire administered to human respondents;
- (b) an *enhanced prompt* (hereafter, *perfect prompt*), designed to improve instruction clarity and response consistency while preserving the original survey content.

In this case, two models were used across the three regions:

- * GPT-5: For each type of prompt and demographic setting (age, age+income), we generated $n = 200$ surveys in each region, obtaining the following factorial design:

$2 \text{ demographic settings} \times 2 \text{ prompts} \times 3 \text{ regions} = 12 \text{ configurations.}$

- * GPT-4o: Since this model was launched prior to our experiment, we had limited access via the online interface. This restricted our generation to the following scenarios:
 - (a) Madrid: Original Prompt (Age and Age + income), Perfect Prompt (Age + income)
 - (b) Andalusia: Original Prompt (Age + income)
 - (c) Valencia: Original Prompt (Age + income)

For each of the five scenarios, we generated $n = 200$ surveys.

3. **GPT API:** responses were generated using structured prompts under two prompting conditions (original prompt and perfect prompt), across three different models:
 - (a) GPT-4,
 - (b) GPT-4o mini,
 - (c) GPT-5.

For each model, we considered two demographic configurations provided as contextual information to the model: (i) age, and (ii) age and income. Simulations were conducted separately for each of the three regions under study: Madrid, Valencia, and Andalusia.

This yields the following factorial design:

$3 \text{ models} \times 2 \text{ demographic settings} \times 3 \text{ regions} \times 2 \text{ prompts} = 36 \text{ configurations.}$

For each configuration, we generated $n = 200$ simulated respondents. In total, the API experiments comprise 36 survey batches of 200 simulated respondents each (7,200 observations overall).

- **Indirect nature of the survey.** In all cases, both human respondents and language models report:
 - * the number of contacts who would vote for each party,
 - * the number of contacts who would blank or null vote, or who do not know their voting intention.

- * the number of contacts who are medical doctors, autonomous workers, and unemployed.

- **Chilean elections.**

For the case of Chile, two different electoral scenarios were considered: the first presidential round and the second round or runoff. In both scenarios, an indirect survey was applied with the aim of estimating voting intentions based on the contact networks of the respondents. In this study, we conducted surveys with human participants and generated responses through both the ChatGPT online interface and the GPT API in an automated manner.

- **Survey rounds.**

- * **First round:** a full round of surveys was conducted during the official campaign period, covering all 8 presidential candidates.
- * **Second round:** a second round of surveys was conducted during the two weeks preceding the runoff election, considering only the two candidates who advanced from the first round.

- **Data sources.**

In each survey round, data were collected from the following sources:

1. **Human respondents:** We followed the same survey design, sampling strategy, and preprocessing methodology as in the Spanish case study. In particular, we conducted four survey waves: one during the first electoral round and three during the second electoral round. In each survey wave, we collected responses from 200 participants regarding the voting intentions of their close contacts, together with the corresponding demographic control variables required by the NSUM estimators. Participants were sampled using an age-stratified design, and the same MAD-based outlier filtering procedure was applied. The main differences with respect to the Spanish dataset concern the electoral context, the source used to construct the age strata, and the participant recruitment process. Specifically, respondents were sampled proportionally to the age distribution reported by the Chilean electoral census, and recruitment was conducted through the Pollfish platform (<https://www.pollfish.com/>). The questionnaire was adapted to the Chilean electoral process while preserving the same survey design and variable structure.
2. **ChatGPT (chat modality):** responses generated with GPT-5 under the same type of prompts as in the Spanish scenario:
 - (a) an *original prompt*, replicating the human respondent survey;
 - (b) an *enhanced prompt* (*perfect prompt*), an optimized version of the previous one.

For each prompting condition, we generated $n = 200$ surveys per electoral round.

3. **GPT API:** responses were generated using structured prompts under two prompting conditions (original prompt and perfect prompt), across three different models:
 - (a) GPT-4,
 - (b) GPT-4o mini,
 - (c) GPT-5.

For each electoral round, the factorial design consists of:

$$3 \text{ models} \times 2 \text{ prompts} = 6 \text{ configurations.}$$

For each configuration, we generated $n = 200$ simulated respondents. In total, the API experiments comprise 6 survey batches of 200 simulated respondents each (1,200 observations per electoral round).

In the first electoral round, a single batch was collected, whereas in the second round, three independent batches were collected.

- **Indirect nature of the survey.** In all cases, both human respondents and language models report:
 - * the number of contacts who would vote for each candidate,
 - * the number of contacts who would cast a blank or null vote, or who do not know their voting intention.
 - * the number of contacts who are medical doctors, autonomous workers, and unemployed.

5.2 Results

This section presents the results from two perspectives. First, the accuracy of the LLM surveys is compared with real results. Second, we compare accuracy against results obtained from human responses. Previously, a sensitivity analysis was performed on the LLM-based surveys to determine the optimal values for the following dimensions: LLM version, the optimal combination of demographic variables, the LLM interface used (API or online), the NSUM method to use, and the ideal prompting strategy (manual or optimized). Using this optimal combination of values for each dimension, the results of the surveys based on our approach were compared against the others (real results, Results from indirect surveys in humans, etc.) for the general elections in three Spanish provinces and the first and second rounds of the Chilean presidential elections.

5.2.1 LLM Survey Accuracy Based on Real Results

Spanish Elections

• Dimensions Analysis

The appendix shows the detailed results of all four NSUM estimators using the synthetic data generated with the LLMs. Following the findings in [ARDA⁺25], which identify PIMLE as the superior performer among the four NSUM techniques, we adopt this approach to evaluate the Spanish electoral process. The following analysis uses PIMLE to discuss the best-performing LLM, the optimal combination of demographic variables, and the most effective prompting strategy.

Model	MADRID		ANDALUSIA		VALENCIA	
	Age	Age + Income	Age	Age + Income	Age	Age + Income
<i>Panel A: API</i>						
GPT 4	10.84	11.67	10.96	11.24	11.58	10.8
GPT 4o - mini	6.86	7.66	6.65	7.37	7.12	6.91
GPT 5	3.77	3.72	3.95	3.92	4.28	4.25
<i>Panel B: Online</i>						
GPT 5	4.10	3.46	3.86	3.28	4.21	3.42

Table 1: Comparison by region and input demographic variables using RMSE and PIMLE, considering blank vote (Bold values indicate the best performing scenario in each case).

As previously mentioned, because of technical difficulties in working with GPT 4o through the Chat interface (Online), certain iterations had to remain unexplored. Tables SM1 and SM2 in the appendix show the complete study, including the available results for said model.

The results in Table 1 specifically evaluate how the inclusion of additional demographic variables, such as income, affects predictive accuracy. The comparison shows that for more advanced model generations like GPT 5, the Age + Income configuration consistently achieved lower RMSE values than the Age-only approach across all regions and interfaces. This suggests that the model effectively takes into account the socioeconomic data to refine its estimations.

However, this improvement is not uniform across older or smaller models. In the API channel, GPT-4 and GPT-4o-mini often performed better when restricted to Age alone. This discrepancy implies that while GPT-5 successfully processes the multidimensional relationship between income and voting intent, less sophisticated models may experience a noise effect when additional variables are introduced, leading to a slight degradation in precision.

The results also highlight Chat as a premium interface, as the Chat results for GPT-4o and GPT-5 show significantly lower error rates than their API

counterparts. Notably, GPT-5 in the Chat interface reached its highest levels of accuracy when utilizing the combined demographic variables.

Model	MADRID		ANDALUSIA		VALENCIA	
	Original P.	Perfect P.	Original P.	Perfect P.	Original P.	Perfect P.
<i>Panel A: API</i>						
GPT 4	11.54	10.97	11.25	10.95	11.04	12.09
GPT 4o - mini	6.91	7.61	6.52	7.51	6.57	7.46
GPT 5	3.98	3.51	4.30	3.57	4.28	4.22
<i>Panel B: Online</i>						
GPT 5	4.14	3.42	4.17	2.97	3.82	3.80

Table 2: Comparison between Original and Perfect prompt across regions using RMSE and PIMLE, considering blank vote (Bold values indicate the best performing scenario for each model. Missing gaps are due to computational difficulties in working with specific models through the Chat interface).

Moving onto the next dimension of the analysis, the results in Table 2 compare the Original Prompt against the optimized architecture previously defined, the Perfect Prompt, across the three regions. The data reveals that the Perfect Prompt serves as a superior architecture for high-capacity models, particularly using Chat interfaces. Since the Perfect Prompt was specifically designed to leverage the conversational and reasoning depth of Chat environments, GPT-5 achieved its most accurate results when utilizing this optimized structure in that specific medium.

Nevertheless, the effectiveness of the prompt architecture appears to correlate with model sophistication. While GPT-4 and GPT-5 generally benefited from the optimization, the smaller GPT-4o-mini model performed better with the Original Prompt in certain regions. This suggests that simpler models might be more sensitive to the increased complexity of the Perfect Prompt, whereas larger models possess the heuristic depth required to process these structured instructions effectively.

Ultimately, these findings reveal that the Perfect Prompt architecture, when paired with the Age + Income demographic approach, delivers the most robust results in advanced models. Furthermore, our dimensional analysis demonstrates that the Chat interface consistently outperforms the API. This suggests that LLMs may draw upon real-time information available online to provide more precise responses.

- **Madrid**

Table 3 shows a clear evolution in how LLMs predict election outcomes compared to traditional human surveys. While older models like GPT 4 struggled by significantly overestimating the performance of the opposition, the newer generations tell a different story. When provided with age

Survey	PP	PSOE	VOX	SUMAR
Real Results	16	10	5	6
Human Surveys	16	11	5	5
API · P.original · Age - GPT 5	14	10	7	6
API · P.original · Age and income - GPT 5	14	10	6	7
API · P.original · Age - GPT 4	12	16	5	4
API · P.original · Age and income - GPT 4	12	17	5	3
API · P.original · Age - GPT 4o-mini	14	11	6	6
API · P.original · Age and income - GPT 4o-mini	13	12	5	7
API · P.perfect · Age - GPT 5	15	9	7	6
API · P.perfect · Age and income - GPT 5	15	9	7	6
API · P.perfect · Age - GPT 4	14	16	5	2
API · P.perfect · Age and income - GPT 4	13	16	6	2
API · P.perfect · Age - GPT 4o-mini	14	11	5	7
API · P.perfect · Age and income - GPT 4o-mini	12	13	5	7
ONLINE · P.original · Age - GPT 5	14	10	7	6
ONLINE · P.original · Age and income - GPT 5	13	12	5	7
ONLINE · P.original · Age - GPT 4o	15	10	6	6
ONLINE · P.original · Age and income - GPT 4o	16	10	5	6
ONLINE · P.perfect · Age and income - GPT 4o	13	11	5	8
ONLINE · P.perfect · Age - GPT 5	15	9	6	7
ONLINE · P.perfect · Age and income - GPT 5	12	10	6	6

Table 3: Seat distribution in Madrid using PIMLE (Highlighted in bold are the distribution that gives back the lowest RMSE value in comparison with real results. This value is 1.22 for the iterations in the API and exactly 0 for GPT 4o using the Original Prompt, considering Age and Income in the Chat interface).

and income data, GPT 4o matched the actual results perfectly. Interestingly, both GPT 4o and GPT 5 produced very similar seat distributions throughout different combinations of prompt and demographic variables. This suggests that as these models get better at processing demographic data, they move beyond simple guessing and start to accurately simulate human behavior.

- **Andalusia**

The results for Andalusia in Table 4 follow a similar pattern to the Madrid data, with newer models showing a much stronger grasp of voting dynamics. While GPT 4 struggled significantly by overestimating PSOE’s and undercounting the PP by a wide margin, the more recent models were much more accurate. Specifically, the Online version of GPT 4o matched PP’s actual 25 seats exactly when demographic factors were included. Even though GPT 4o-mini and GPT 5 also showed solid performance, these findings reinforce the idea that as LLMs incorporate specific data

like income, they become increasingly reliable at reflecting the pattern of regional voting behavior.

Survey	PP	PSOE	VOX	SUMAR
Real Results	25	21	9	6
Human Surveys	25	20	8	8
API · P. original · Age - GPT-5	22	18	12	9
API · P. original · Age and income - GPT-5	22	18	12	9
API · P. original · Age - GPT-4	17	31	8	5
API · P. original · Age and income - GPT-4	17	31	8	5
API · P. original · Age - GPT-4o-mini	24	19	10	8
API · P. original · Age and income - GPT-4o-mini	21	20	9	11
API · P.perfect · Age - GPT-5	23	19	11	8
API · P.perfect · Age and income - GPT-5	23	19	11	8
API · P.perfect · Age - GPT-4	18	31	9	3
API · P.perfect · Age and income - GPT-4	17	32	9	3
API · P.perfect · Age - GPT-4o-mini	26	17	11	7
API · P.perfect · Age and income - GPT-4o-mini	16	25	8	12
ONLINE · P. original · Age - GPT-5	22	17	12	10
ONLINE · P. original · Age and income - GPT-5	22	19	11	9
ONLINE · P. original · Age and income - GPT-4o	25	18	10	8
ONLINE · P. Perfect · Age - GPT-5	23	18	11	9
ONLINE · P. Perfect · Age and income - GPT-5	23	19	10	9

Table 4: Seat distribution in Andalusia using PIMLE (Highlighted are the best performing iterations in both the API and Online channels. More specifically, the lowest RMSE value for the API was 1.58 and through the Chat interface 1.87, both achieved by the respective iteration shown in bold).

• Valencia

The results for Valencia in Table 5 confirm the trends observed in previous regions, with newer models consistently outperforming their predecessors. GPT 4 continued to struggle with identifying the majority party, likely because its training data reflected a different political landscape. However, more advanced iterations like GPT 4o and GPT 5 significantly closed the gap, showing much higher accuracy. In this region, the Online version of GPT 4o and GPT 5 performed equally well when demographic data was included, proving that as LLMs mature, they can provide a level of precision that rivals traditional human surveys.

Survey	PP	PSOE	VOX	SUMAR
Real Results	13	11	5	4
Human Surveys	12	10	6	5
API · P.original · Age – GPT 5	12	9	6	6
API · P.original · Age and income – GPT 5	12	9	6	6
API · P.original · Age – GPT 4	10	15	4	4
API · P.original · Age and income – GPT 4	10	15	4	4
API · P.original · Age – GPT 4o-mini	14	9	5	5
API · P.original · Age and income – GPT 4o-mini	11	11	4	7
API · P.perfect · Age – GPT 5	12	9	6	6
API · P.perfect · Age and income – GPT 5	12	9	6	6
API · P.perfect · Age – GPT 4	11	15	5	2
API · P.perfect · Age and income – GPT 4	11	16	4	2
API · P.perfect · Age – GPT 4o-mini	9	13	5	6
API · P.perfect · Age and income – GPT 4o-mini	10	12	5	6
ONLINE · P.original · Age – GPT 5	12	9	6	6
ONLINE · P.original · Age and income – GPT 5	11	10	6	6
ONLINE · P.original · Age and income – GPT 4o	14	9	5	5
ONLINE · P.perfect · Age – GPT 5	13	9	5	6
ONLINE · P.perfect · Age and income – GPT 5	12	10	5	6

Table 5: Seat distribution using in Valencia PIMLE (Marked in bold are shown the most accurate seat distributions for each channel, API and Online. In the case of the Chat interface, there is a tie between GPT 4o using the Original Prompt and GPT 5 using Perfect Prompt, both considering Age and income. The best performance of the iterations in this region all coincide in an RMSE value of 1.22).

Chilean Elections

• Dimensions Analysis

Table 6 provides a comprehensive summary of the mean Root Mean Square Error (RMSE) of the estimators calculated across the multiple iterations. The temporal distribution of these queries is of particular methodological significance. For the online interface, five distinct observations were recorded. The first two observations (8/11 and 21/11) were timed slightly before and after the first round of the Chilean presidential elections, capturing the model’s predictive and reflective states immediately prior to and following the electoral event. The other three were generated before the second round, gaining insight into the model’s accuracy in predicting the unknown. The API-based queries were partitioned to provide an overall view of performance, with one round dedicated to the initial electoral phase (8/11) and two subsequent rounds (24/11 & 12/12) focused on the second phase.

Estimator	8/11	21/11	24/11	4/12	12/12
<i>Panel A: API</i>					
NAIVE	9.5521	–	14.8210	–	15.2632
MOS	9.5535	–	14.7309	–	15.1249
MLE	9.5615	–	14.7545	–	15.1568
PIMLE	9.5515	–	14.7148	–	15.0856
<i>Panel B: Online</i>					
NAIVE	8.137	8.391	10.867	10.809	12.101
MOS	8.126	8.109	10.569	10.497	11.804
MLE	8.301	8.150	10.633	10.550	11.954
PIMLE	8.107	8.141	10.519	10.501	11.785

Table 6: Performance comparison of the average RMSE for Chilean Presidential Election Forecasts using different NSUM estimators and access channels (API and Online), including blank vote.

A comparative analysis of the estimators reveals that the PIMLE consistently demonstrates superior predictive accuracy. As illustrated in Panel B, the PIMLE estimator uniformly outperforms all competing models within the API environment, yielding the lowest error rates across all three rounds. While the results from the online interface (Panel A) show a more competitive landscape, PIMLE maintains its status as the most robust estimator; it is only marginally surpassed by the MOS estimator in two specific instances (the 21/11 and 4/12 queries). This pattern of high-fidelity performance is not an isolated occurrence. The PIMLE estimator exhibited a similar degree of precision in prior studies concerning Spanish electoral data [ARDA⁺25], suggesting a high level of cross-contextual reliability. Consequently, given its demonstrated robustness and consistent minimization of error across both online and API platforms, we have designated PIMLE as the primary estimator for all subsequent analytical phases of this study.

The results presented in Table 7 delineate the performance hierarchy of the models across the iterations. By fixing the estimator to PIMLE, we can isolate the impact of model architecture and the interface type (Online vs API) on the resulting predictive accuracy. A significant performance gap is observed within the API channel (Panel A). While the “mini” variant (GPT-4o - mini) initially showed competitive results, its error margin widened drastically during the runoff phase, ultimately gaining the highest error rates in the study. In contrast, GPT-5 demonstrated superior robustness, consistently maintaining the lowest error rates for the API channel in all late-stage queries. This suggests that the greater number of parameters and reasoning depth of the GPT-5 architecture are essential for handling the complexity of the Chilean electoral data, particularly when

Model	08/11	21/11	24/11	04/12	12/12
<i>Panel A: API</i>					
GPT 4	8.68	–	14.35	–	15.23
GPT 4o - mini	10.48	–	18.61	–	18.65
GPT 5	9.49	–	11.18	–	11.37
<i>Panel B: Online</i>					
GPT 5	8.11	8.14	10.52	10.50	11.79

Table 7: Longitudinal analysis of Chilean presidential election forecasts (Values represent Average RMSE for blank-vote adjusted estimations, disaggregated by model, access channel, and temporal iteration).

managing the nuances of blank vote estimation. The data also reveals a consistent “performance premium” for the Online interface compared to the API.

Building on the observation that all models experienced a precision decay following the first round, the magnitude of this decline varied significantly across model generations. While older architectures like GPT-4 saw a sharp decline in accuracy, GPT-5 Online exhibited a more stable transition, reinforcing its resilience to the shifting dynamics of a polarized runoff. This robustness is further supported by the data in Table 8, which justifies the Perfect Prompt as a superior architecture for high-stakes predictive tasks. By analyzing the evolution of the RMSE across the Chilean electoral cycle, it becomes clear that this optimized framework provides the necessary reliability to mitigate accuracy loss during the transition between electoral phases.

In the initial stage of the election (see Table 8 Panel A), the RMSE values remain relatively stable across both prompting strategies. This suggests that during the first round, the higher number of candidates and observations allows for a more distributed error margin, which can be effectively managed even by simpler query structures. In this multi-candidate environment, the models demonstrate a baseline level of predictive competence that does not yet fully demand the specific constraints of the optimized architecture.

However, the superiority of the Perfect Prompt as an optimization tool becomes decisive during the second-round runoff. As the electoral landscape narrowed to only two candidates, the margin for error decreased significantly, making the models more susceptible to large fluctuations. The results from Panels C through E illustrate that the Perfect Prompt acts as a critical calibration layer in this polarized scenario. For instance, the transition observed in GPT-5 shows that the optimized architecture consistently yields lower RMSE values compared to the Original Prompt, which struggled to maintain precision when the variables were reduced to a

Model and Access	Original Prompt	Perfect Prompt
<i>Panel A: 08/11</i>		
<i>API</i>		
GPT 4	8.31	9.05
GPT 4o - mini	10.48	10.48
GPT 5	9.41	9.57
<i>Online</i>		
GPT 5	7.86	8.36
<i>Panel B: 21/11</i>		
<i>Online</i>		
GPT 5	9.89	6.40
<i>Panel C: 24/11</i>		
<i>API</i>		
GPT 4	14.50	14.20
GPT 4o - mini	20.20	17.02
GPT 5	11.35	11.01
<i>Online</i>		
GPT 5	11.70	9.33
<i>Panel D: 04/12</i>		
<i>Online</i>		
GPT 5	11.04	9.96
<i>Panel E: 12/12</i>		
<i>API</i>		
GPT 4	14.29	16.17
GPT 4o - mini	20.20	17.10
GPT 5	11.17	11.58
<i>Online</i>		
GPT 5	14.47	9.10

Table 8: Performance comparison of the Average RMSE for Chilean presidential election forecasts by iteration, model and access channel between Original and Perfect Prompts, including blank vote estimation (Note: In Panel A, GPT 4o - mini accessed via API obtained the following scores: Original Prompt: **10.4816** - Perfect Prompt: **10.4810**, therefore the Perfect Prompt shows a slightly better performance than the Original Prompt).

binary choice. Ultimately, the Perfect Prompt proves to be an essential infrastructure for maintaining model stability. By providing a more rigorous framework for the model to process information, this query optimization effectively mitigates the escalation of the RMSE that typically occurs as an election approaches its final stage. These findings confirm that a structured query architecture is not merely a preference but a methodological requirement for ensuring high-fidelity results in longitudinal electoral forecasting.

• Results First Round

Survey	FP	JJ	MEO	JK	JAK	EA	EM	HM	Blank
<i>Panel A: 08/11</i>									
Human Surveys	6.9	22.5	2.4	13.6	28.0	2.1	13.3	2.8	8.4
API - GPT 4 - Original Prompt	15.1	12.7	14.1	2.8	21.7	1.8	16.2	6.7	8.8
API - GPT 5 - Perfect Prompt	10	12.7	7.8	3.7	20.4	3.7	29.2	2.7	9.9
Online - GPT 5 - Original Prompt	20.2	11.8	7.5	0.9	26.8	0.1	19.8	6.0	6.9
Online - GPT 5 - Perfect Prompt	14	14.4	6	0.7	23.7	2.5	24.9	3.8	9.9
<i>Panel B: 21/11</i>									
Online - GPT 5 - Perfect Prompt	16.3	18.6	8.3	1.9	24.3	2.2	17.2	1.8	9.3
Real Results	19.71	26.85	1.2	13.94	23.92	0.66	12.46	1.26	1.06

Table 9: Comparison between voting estimates for each of the presidential candidates in the Chilean elections: traditional surveys, GPT-based models (via API and Chat), and Real Results (FP: Parisi, JJ: Jara, MEO: Ominami, JK: Kaiser, JAK: Kast, EA: Artés, EM: Matthei, HM: Mayne-Nicholls).

The results in Table 9 evaluate how various configurations estimated electoral support during the first round of the Chilean presidential race. A primary finding is that JAK held the plurality in every LLM iteration, aligning with the actual outcome. This trend is particularly evident in the Online GPT 5 Perfect Prompt on November 8, which reached a precision of 23.7% compared to the real 23.92%, confirming the models’ ability to identify the leading candidate.

However, despite identifying the frontrunner, the LLMs consistently projected a more distributed vote than the Real Results. While the actual election was highly polarized, leaving candidates like Ominami (MEO) and Artés (EA) with negligible support, the models assigned significantly higher shares to these fringe options. This suggests that the models reflect a broader digital discourse rather than the tactical consolidation seen in real-world voting behavior.

Finally, although the Original Prompt for GPT 4 API and GPT 5 Online proved highly competitive in the first iteration, GPT 5 Perfect Prompt is the preferred choice for its longitudinal robustness. While specific earlier configurations showed isolated strengths, this optimized query architecture provides greater stability across the entire study and eventually narrows the gap for candidates like JJ in the post election reflection.

• Results Second Round

The results for the Chilean presidential runoff in Table 10 demonstrate a high degree of homogeneity across all estimation methods. Both human surveys and LLM configurations consistently place José Antonio Kast (JAK) as the leader over Jeannette Jara (JJ), mirroring the actual electoral direction. While these estimations remain stable throughout the

three panels, they present a much tighter race than the final results, where the gap between candidates eventually widened to over 16 points.

Within this consistent landscape, the second iteration in Panel B provides a notable methodological exception. In this specific timeframe, the Original Prompt for the API GPT 5 performed slightly better than the Perfect Prompt, recording an RMSE of 8.77 compared to 8.87. This suggests that in a binary runoff scenario with fewer variables, the standard query structure can occasionally match or exceed the precision of the optimized architecture.

A significant point of divergence remains the estimation of the blank vote, which is radically higher in all models than the final result of 1.33%. Human surveys and LLM iterations consistently projected blank vote shares between 12% and 17%, failing to capture the massive consolidation of undecided voters that occurred on election day. This persistent discrepancy indicates that while the models are effective at identifying the winner, they struggle to predict the degree to which voters ultimately commit to a valid preference in the final stage of the race.

Survey	JAK	JJ	Blank
<i>Panel A: 24/11</i>			
Human Surveys	48.00	38.00	12.00
API - GPT 5 - Perfect Prompt	45.86	38.00	16.15
Online - GPT 5 - Perfect Prompt	46.77	39.81	13.43
<i>Panel B: 04/12</i>			
Human Surveys	48.45	39.00	12.56
Online - GPT 5 - Perfect Prompt	46.52	38.97	14.51
<i>Panel C: 12/12</i>			
Human Surveys	46.26	40.94	12.81
API - GPT 5 - Original Prompt	45.31	38.51	16.18
API - GPT 5 - Perfect Prompt	45.39	37.64	16.97
Online - GPT 5 - Perfect Prompt	47.10	39.73	13.17
Real Results	57.39	41.29	1.33

Table 10: Comparison of voting estimates for each presidential candidate in the Chilean presidential runoff between human surveys, GPT-5 model iterations (API vs. online), and official results, including the percentage of blank votes.

5.2.2 LLM Surveys Emulating Human Responses

Throughout the research process, a recurring pattern emerged: LLMs’ responses aligned more closely with human survey data than with actual electoral outcomes. Based on this observation, this section explores the AI’s emulation of human sociopolitical behavior. By adopting the human survey as our primary

benchmark, we shift the definition of the "real" value away from the final election results to better analyze the model's behavioral approximation.

Spanish Elections

The data visualized in Figure 1 illustrates five box-plots about the estimated vote percentages for Spain's primary political parties and blank ballots across Madrid, Valencia, and Andalusia. This figure represents the various iterations of human-emulated results obtained via API and Online channels. While the results show high consistency in median estimations across the different methodologies, the nuances in variance and outliers provide critical insights into the intersection of statistical modeling and the characteristics of the Spanish political landscape.

A prominent feature of this dataset is the high prevalence and significant dispersion of blank votes, a phenomenon driven by a combination of structural elements inherent to the surveying environment. Primarily, this dispersion is attributed to the "social circle" effect, where respondents or emulated agents lack clear knowledge regarding the voting intentions of their immediate environment. Furthermore, this category acts as a statistical catch-all; the absence of a stated vote is commonly recorded as a blank, as are votes intended for smaller political forces not explicitly contemplated in the survey parameters. This accumulation of "noise" translates into higher statistical uncertainty, as the models must account for a wide spectrum of non-aligned or excluded preferences.

In turn, PSOE's results exhibit significant outliers, reflecting an overestimation of the party's popularity by specific older models, such as GPT-4. Because these previous models were trained in a very different political moment characterized by peak institutional visibility for PSOE, they tend to carry forward an inflated popularity that does not align with more recent, grounded scenarios. In contrast, a factor to highlight is the relative stability of both Vox and PP, whose distributions remain tightly clustered. This stands in marked opposition to the outlying behavior and higher variance observed in the left-wing bloc (Sumar and PSOE). Ultimately, these findings suggest that the emulation of results is as varied as the opinion of the population itself, mirroring the fragmentation and volatility of the current Spanish electorate.

Based on the statistical distribution of the results, it is necessary to examine how these emulations perform against different benchmarks. Table 11 illustrates the shift in precision that occurs when synthetic data is measured directly against actual electoral results. The high RMSE values across Madrid, Andalusia, and Valencia reveal a significant reality gap. From this perspective, the models appear to be inconsistent predictors of objective political outcomes.

However, when the same synthetic data is measured against human survey results in Table 12, the error margins drop substantially. For example, the error for GPT 5 in Andalusia falls over two whole points when shifting the comparison. This increased precision across both API and Chat interfaces suggests that the internal logic of these models is inherently synchronized with human response

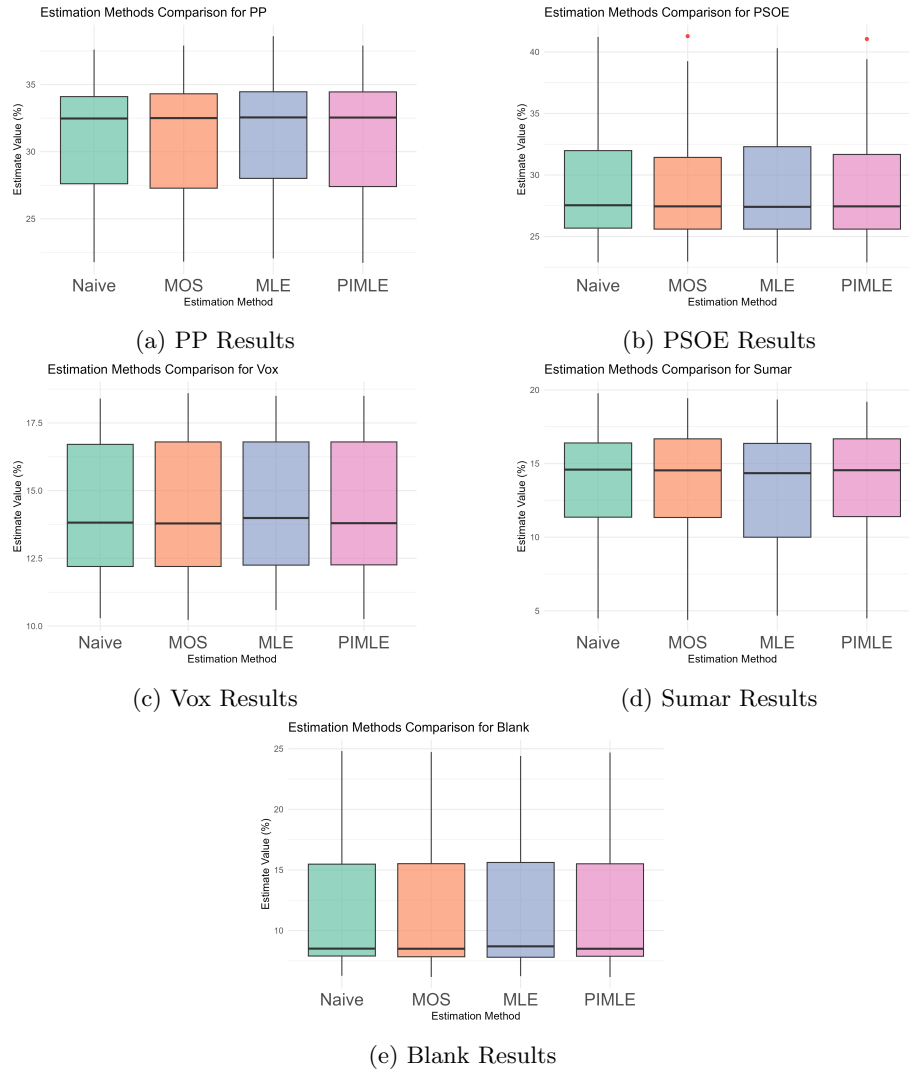


Figure 1: General comparison across the different political parties

patterns. These results indicate that the LLMs may be functioning more as high-fidelity mirrors of survey sentiments than as tools for objective electoral forecasting, a trend that invites further analysis in other political contexts.

Model	MADRID	ANDALUSIA	VALENCIA
<i>Panel A: API</i>			
GPT 4	11.26	11.10	11.31
GPT 4o - mini	7.26	7.01	6.55
GPT 5	3.75	3.94	4.26
<i>Panel B: Online</i>			
GPT 4o	3.11	3.48	3.99
GPT 5	3.78	3.57	3.81

Table 11: Comparison of RMSE values using PIMLE across different regions and interfaces, considering blank vote (Bold values indicate the lowest error for each interface category).

Model	MADRID	ANDALUSIA	VALENCIA
<i>Panel A: API</i>			
GPT 4	10.68	10.12	9.07
GPT 4o - mini	6.72	4.47	3.67
GPT 5	3.19	1.39	1.40
<i>Panel B: Online</i>			
GPT 4o	3.17	1.21	2.48
GPT 5	3.29	1.61	1.58

Table 12: Comparison of RMSE values with human results using PIMLE, considering blank vote (Bold values indicate the highest precision (lowest RMSE) achieved per interface. This table highlights the capacity of LLMs to simulate human-like response distributions).

Chilean Elections

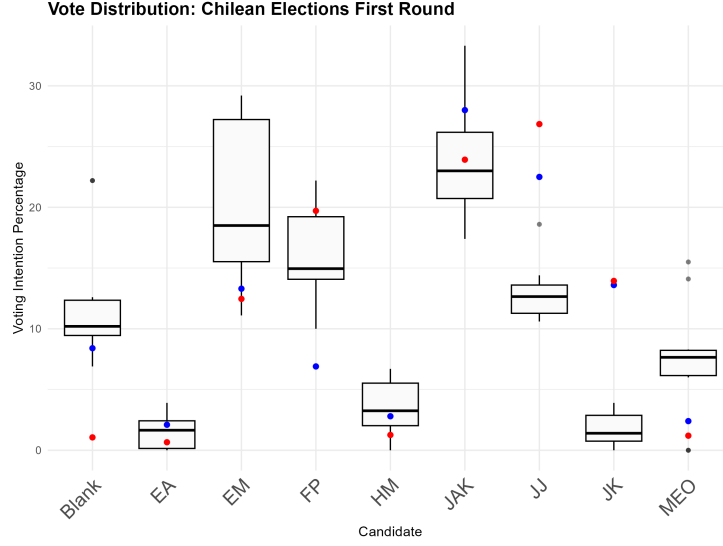
The results in the Chilean scenario indicate that the model works even more effectively as an emulator of human biases in surveys than as a direct predictor of final election results than in the case of the Spanish elections. As illustrated in Figure 2a, the first round results show that the model’s interquartile ranges (IQRs) often encompass or sit adjacent to the human polling data (blue dots), even when those polls significantly diverge from the real results (red dots). This is particularly evident for candidate EM, where the model and the single available human survey are closely aligned in their overestimation of support relative to the final vote count. The distribution of the different iterations appears to capture the polling logic, reflecting the same underlying uncertainties

and systemic errors present in traditional surveying methods during the multi-candidate stage.

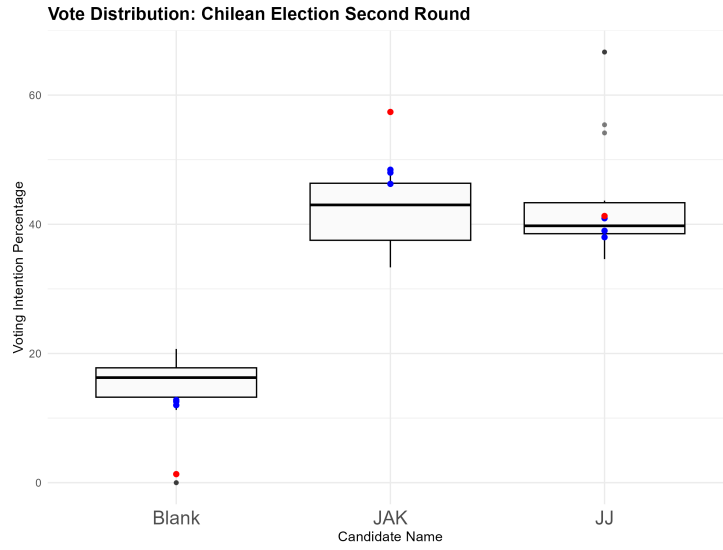
The second round, depicted in Figure 2b, provides a more robust validation of this emulative behavior due to the inclusion of three distinct human surveys. These surveys show a high degree of consensus among human pollsters, and the model’s boxplots demonstrate a remarkable ability to approximate this specific cluster of human results. While both the polls and the model’s median slightly under-predicted the final surge for candidate JAK, the emulation of the human results remains remarkably consistent across all three survey benchmarks. Notably, while the general model distribution emulates the consensus of polling sentiment, the precision in this round reaches a peak in the third and final human survey; for candidate JJ, this specific survey iteration achieves an almost perfect estimation of the final result. A persistent discrepancy, once again, is the estimation of the Blank vote. While human surveys and LLM predictions are closely aligned, the final result is consistently closer to zero; a reality that only a single model iteration was able to capture.

The longitudinal analysis of the Chilean results further supports the pattern observed in Spain, showing a distinct shift in accuracy depending on the point of comparison. In Table 7 (see 5.2.1, Chilean elections, Dimensional analysis), where the synthetic data is measured against the current electoral results, the results show a consistent discrepancy. Across the various temporal iterations, the error margins remain high, suggesting that the models struggle to bridge the gap between their generated outputs and the objective reality of the final vote.

In contrast, the results in Table 13 show a significant increase in precision when the same data is compared to human survey results. This is particularly evident in the final rounds of the election, where the error for advanced models like GPT 5 drops by several points compared to the electoral benchmark. The fact that this alignment holds steady over time and repeats across both the Spanish and Chilean studies suggests that the models’ internal logic is structurally tuned to human sociopolitical sentiment. Ultimately, these findings indicate that the capacity for behavioral emulation is not a localized fluke but a fundamental characteristic of LLMs, positioning them as sophisticated social simulators rather than reliable tools for objective political prediction.



(a) First Round Election Results.



(b) Second Round Election Results.

Figure 2: Distribution of voting intention in the Chilean elections using PIMLE (Red dots represent real results, blue dots represent human polls, and boxplots show model iterations. FP: Parisi, JJ: Jara, MEO: Ominami, JK: Kaiser, JAK: Kast, EA: Artés, EM: Matthei, HM: Mayne-Nicholls).

Model	08/11	21/11	24/11	04/12	12/12
<i>Panel A: API</i>					
GPT 4	7.61	–	6.60	–	7.36
GPT 4o	9.37	–	16.22	–	16.25
GPT 5	7.69	–	2.88	–	2.99
<i>Panel B: Online</i>					
GPT 5	7.41	7.24	2.69	2.31	3.06

Table 13: Longitudinal analysis of LLM emulation accuracy of the RMSE for estimations across five temporal iterations using PIMLE, comparing API and Online access channels (Bold values indicate the highest precision achieved in the Online interface)

5.2.3 Correlational Analysis between LLM and human surveys

To empirically evaluate the models’ capacity for human emulation, we leverage correlation matrices to quantify these behavioral patterns numerically. Due to the large volume of data generated in this study, the depth-based estimations will be limited to the Spanish scenario. Within this scope, we will work with the combination of factors that has proven to be the most efficient: the GPT-5 model, the perfect prompt strategy, and the inclusion of both demographic variables: age and income.

To quantify this structural similarity, we construct correlation matrices for the political parties under study, contrasting the patterns found in real-world surveys against those from the synthetic data. Along these lines, we will use the Euclidean distance and the matrix correlation coefficient to measure the disparity between survey data and synthetic results. The Euclidean distance measures the absolute difference in magnitude between the cells of both matrices. Given two correlation matrices R_1 and R_2 , the Euclidean distance is defined as the square root of the sum of the squared differences of all their elements:

$$d(R_1, R_2) = \sqrt{\sum_{i=1}^n \sum_{j=1}^m (R_{1,ij} - R_{2,ij})^2}.$$

To isolate structural patterns independently of matrix scale, we utilize a matrix correlation coefficient (ρ_{mat}) calculated across the upper triangle elements, a framework mathematically formalized by Mantel [Man67]. Since correlation matrices are symmetric, the Pearson correlation is calculated over their off-diagonal elements. Let $\text{vech}^*(R)$ be the operator that extracts and vectorizes the elements R_{ij} such that $i < j$. The metric is defined as:

$$\rho_{\text{mat}} = \frac{\sum (r_{1,k} - \bar{r}_1)(r_{2,k} - \bar{r}_2)}{\sqrt{\sum (r_{1,k} - \bar{r}_1)^2 \sum (r_{2,k} - \bar{r}_2)^2}},$$

where $r_{1,k}$ and $r_{2,k}$ are the elements of the vectors $\mathbf{r}_1 = \text{vech}^*(R_1)$ and $\mathbf{r}_2 = \text{vech}^*(R_2)$, respectively. This measure is fundamentally a structural proximity

coefficient. While the Euclidean distance is sensitive to scale changes, ρ_{mat} is invariant under linear transformations.

The comparative analysis of the correlation matrices (Tables 14, 15 and SM12 to SM15 from the appendix) reveals that, although language models successfully identify the block structure of Spanish politics, they tend to project a hyper-polarized view. While human matrices show a smoother fragmentation, with negative correlations close to zero, the Chat interface simulates radical segregation. This suggests that LLMs assume mutual exclusion within social circles that physical reality disproves, evidencing actual coexistence among voters of both blocks. This tendency toward polarization does not mean that the models fail to effectively replicate internal affinities, but it does show the polarizing effects that these artificial intelligence-based tools have, according to the literature. Both access channels identify the cohesion of the right-wing block (PP-VOX, with values above 0.83); however, human data present a more diverse regional consistency. For instance, in Andalusia, the left-wing block (PSOE-Sumar) reaches a solidity of 0.79, surpassing the affinity of the right-wing block, which shows a more moderate relationship in social reality than in the synthetic simulation.

Table 14: Correlation matrix: Madrid, Human Surveys

	PP	PSOE	Vox	Sumar
PP	1.0000	-0.0012	0.4428	-0.0366
PSOE	-0.0012	1.0000	-0.0329	0.3700
Vox	0.4428	-0.0329	1.0000	-0.0987
Sumar	-0.0366	0.3700	-0.0987	1.0000

Table 15: Correlation matrices: Madrid (GPT 5 - P. Perfect - Age and income)

(a) Online Interface					(b) API Interface				
	PP	PSOE	VOX	SUMAR		PP	PSOE	VOX	SUMAR
PP	1.0000	-0.0264	0.8899	-0.3566	PP	1.0000	0.4623	0.8662	0.1626
PSOE	-0.0264	1.0000	0.0344	0.3694	PSOE	0.4623	1.0000	0.4246	0.5117
VOX	0.8899	0.0344	1.0000	-0.2609	VOX	0.8662	0.4246	1.0000	0.1589
SUMAR	-0.3566	0.3694	-0.2609	1.0000	SUMAR	0.1626	0.5117	0.1589	1.0000

The formalization of these observations using the metrics in Table 16 confirms that structural fidelity is not a uniform phenomenon but is conditioned by region and access channel. High correlations between real and simulated data are observed in Madrid and Valencia; however, the Euclidean distance values warn that this similarity is fundamentally morphological, confirming that the intensity of the conflict is greater in synthetic agents than in citizens. This disconnection phenomenon reaches its peak in Andalusia, where LLMs impose

a generic political framework that fails to capture the idiosyncrasies of a more transversal electorate, resulting in maximum structural disparity (ρ_{mat} values below 0.2). The contrast between interfaces is equally revealing. Systematically, the user interface shows a greater capacity to replicate the architecture of human relationships than the developer-oriented version (API). This divergence suggests that the conversational nature of Chat identifies the political stereotypes of the electorate more efficiently, while the API tends to smooth out these relationships, possibly due to stricter technical alignment processes that dilute the underlying correlation structure.

Table 16: Comparison of Distance and Correlation Metrics for Correlation Matrices by Region and Access Channel

Region	Access Channel	Euclidean Distance	Correlation (ρ_{mat})
Madrid	Chat	0.8171	0.9240
	API	1.2079	0.8428
Andalusia	Chat	1.4928	0.1686
	API	1.8942	−0.0189
Valencia	Chat	0.8236	0.8201
	API	1.1545	0.5697

These results confirm that synthetic data possess a coherent internal structure and that LLMs act as sophisticated social simulators capable of understanding dynamics of exclusion and affinity.

5.3 General Discussion of the Results

This section analyzes the main results presented in the tables and figures of the previous section, with the aim of identifying general patterns and discussing their methodological implications. In particular, four central aspects are addressed: the effect of prompt optimization, the comparison between different LLMs, the relationship between model performance and release date, and the similarity between estimates obtained using LLMs and those derived from human surveys.

A first relevant observation is that, for both the Spanish and Chilean elections, the use of the prompt optimizer (*Perfect Prompt*) tends to improve performance in the majority of cases, although not uniformly. While consistent improvements are observed across several models and configurations—particularly in the case of GPT-5 in the online setting for Chile (see Table 8), where the reduction in error reaches up to 5.37 points. There are also scenarios in which the difference is marginal or even slightly unfavorable to the optimized prompt. In general, the most substantial improvements are concentrated in specific models.

Regarding model comparison, the results indicate that more recently released models tend to exhibit better overall performance than earlier versions. In

particular, a consistent improvement is observed in the performance of the GPT-5 model, both in the API and Chat settings. This trend is consistent with the temporal evolution of the models and suggests that architectural and training improvements translate into more accurate aggregate-level estimates.

In turn, the similarity observed between the results obtained using LLMs and those derived from indirect human surveys is particularly noteworthy. For both electoral processes analyzed, the estimates produced by language models more closely resemble survey-based results than the actual electoral outcomes. Figures 3 and 4 show a systematic reduction in error when comparing LLM outputs with data from human surveys. This behavior is consistent with biases in indirect survey methodologies, suggesting that LLMs may be implicitly reproducing response patterns similar to those observed among human respondents.

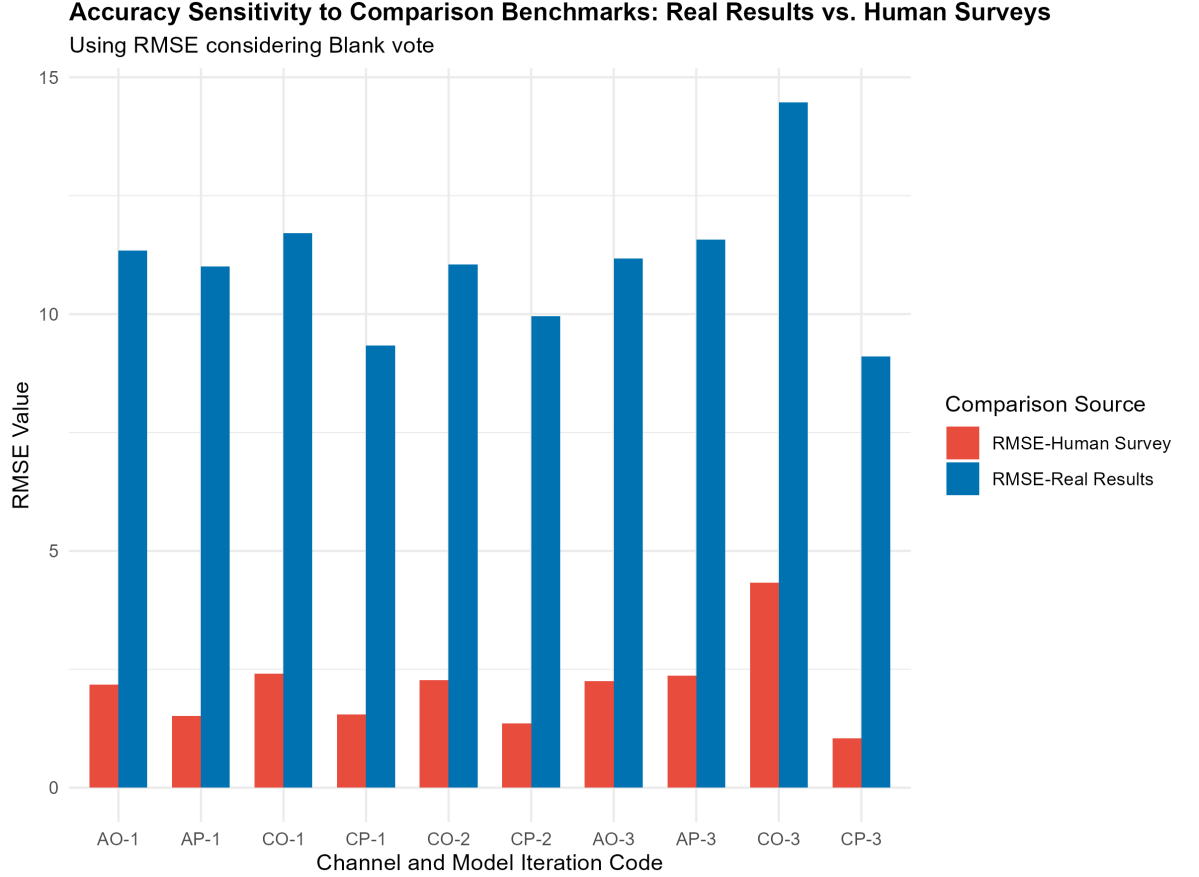


Figure 3: Accuracy Sensitivity to Comparison Benchmarks: Real Results vs. Human Surveys, in Chile (The figure compares the RMSE across different GPT-5 iterations for the second round of the Chilean elections. A consistent reduction in error is observed when the benchmark shifts from real results (blue) to human polling data (red), highlighting the models’ structural alignment with human sociopolitical sentiment. Iteration codes denote: model interface (A: API, C: Chat), prompting strategy (O: Original, P: Perfect), and data collection date (1: Nov 24, 2: Dec 4, 3: Dec 12)).

In Table 17, we observe that removing the blank vote leads to a significant reduction in the estimation error, as measured by the RMSE, across all the surveys considered. This improvement suggests that the blank vote introduces an additional source of noise in the context of indirect surveys, particularly when its magnitude is large. In this type of question, it is reasonable to assume that respondents resort to the blank vote when they are unaware of the electoral preferences of their contacts, which tends to artificially inflate its proportion. Interestingly, language models reproduce this same behavioral pattern, similarly

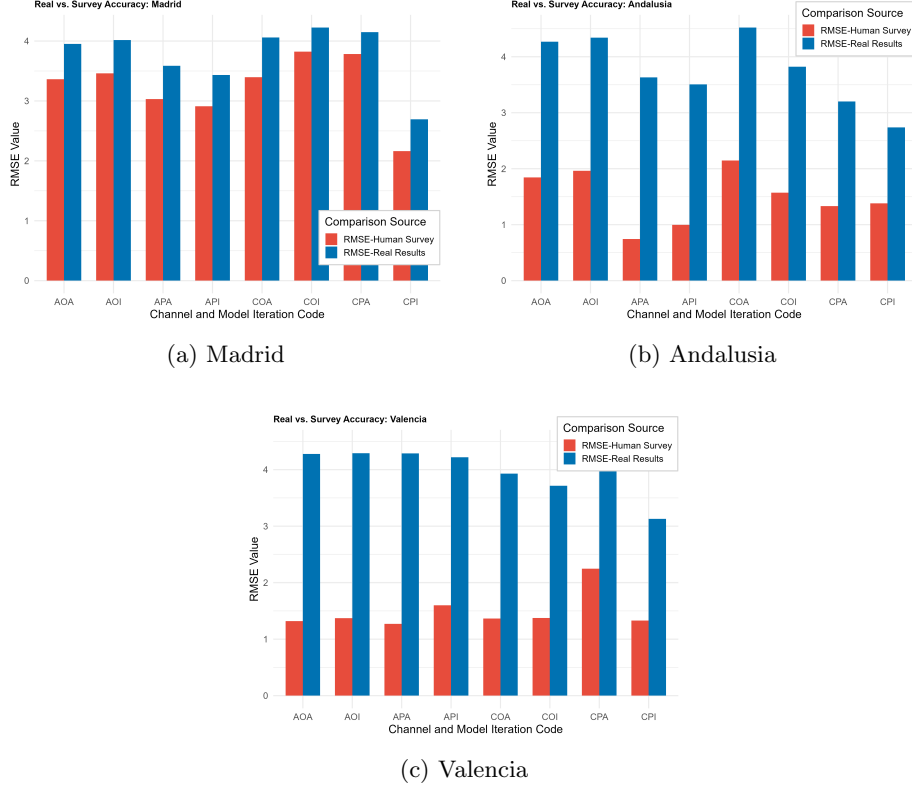


Figure 4: Accuracy Sensitivity to Comparison Benchmarks: Real Results vs. Human Surveys, in Spain (This figure tracks how GPT-5 performance varies across Andalusia, Valencia, and Madrid when measured against different ground truths. We observe the same pattern as in the Chilean scenario: the model aligns more closely with human survey data (red) than with the real results (blue), suggesting that AI outputs may naturally mirror collective social perceptions. The labels represent combinations of: channel (A: API, C: Chat), prompting technique (O: Original, P: Perfect), and demographic variables combination (A: Age, I: Age + Income)).

to human respondents. By excluding this category, the estimates align more closely with the real results, suggesting that the treatment of the blank vote is a key aspect for improving the performance of methods based on indirect questions, both in human surveys and in those generated by language models.

Survey	JAK	JJ	Blank	RMSE
<i>Panel A: With blank vote</i>				
Humans	48.00	38.00	12.00	7.03
API – Perfect Prompt – GPT-5	45.86	38.00	16.15	8.48
Chat – Perfect Prompt – GPT-5	46.77	39.81	13.43	7.58
<i>Panel B: Without blank vote</i>				
Humans	55.81	44.19	–	2.33
API – Perfect Prompt – GPT-5	54.69	45.31	–	3.43
Chat – Perfect Prompt – GPT-5	54.02	45.98	–	4.08
<i>Panel C: Real results</i>				
Real Results (with blank)	57.39	41.29	1.33	–

Table 17: Comparison of vote intention percentages and RMSE across Chilean elections using PIMLE, considering scenarios with and without blank votes, where results without blank votes are obtained by reweighting.

Taken together, these results indicate that LLMs are capable of capturing aggregate-level regularities that are closer to estimates obtained from human surveys than to observed ground-truth values. From this perspective, LLMs emerge as a potentially complementary tool for social and political analysis, particularly within indirect approaches based on methodologies such as NSUM.

6 Conclusion

This study proposes a methodological framework for using LLMs as synthetic respondents in indirect electoral surveys. Operationalizing LLMs as synthetic agents capable of answering questions based on contact networks and integrating their responses with NSUM allows for modeling perceived voting intentions at an aggregate level. The proposed framework integrates carefully designed prompts to simulate indirect questioning, the systematic generation of synthetic survey data, and subsequent estimation using NSUM. The results demonstrated that LLMs can be operationalized not merely as text generators but as controllable agents that emulate a key aspect of social perception: reporting on the voting intentions of one’s acquaintances. This establishes a foundational methodological pipeline for leveraging synthetic data in electoral research.

Specifically, the results indicate that the NSUM-based estimates derived from LLM-generated survey data capture meaningful patterns of collective social knowledge, showing a notable alignment with estimates obtained from real-world indirect surveys in both Spain and Chile. While variations were observed

across different LLM versions, deployment environments, and demographic conditions, overall trends suggest that LLMs are able to internalize and reproduce socially shared perceptions of electoral behavior. These findings support the interpretation of LLMs not as predictors of individual voting choices, but as models of socially aggregated narratives and expectations embedded in public discourse.

In the analyzed cases of Spain and Chile, the synthetic data produced aggregate vote intention distributions that captured the relative standing of the major political forces with notable similarity to benchmark surveys. This suggests that LLMs, trained on vast corpora reflecting public discourse, internalize a form of collective social knowledge about electoral tendencies that can be extracted through indirect survey protocols. However, discrepancies were observed in the precise magnitude of certain party shares and in the stability of estimates across different LLM versions and interfaces (app vs. online), highlighting that the synthetic responses are sensitive to model-specific architectures, updates, and prompting environments.

Despite these promising results, several limitations must be acknowledged. LLMs are trained on large but opaque corpora that may reflect historical biases, uneven media coverage, or temporal mismatches with the electoral events studied. Consequently, synthetic responses may overrepresent dominant narratives while underrepresenting marginalized or rapidly evolving opinions. In addition, LLMs' "knowledge" is inherently retrospective and shaped by their training data, which may not fully capture last-minute electoral shifts or the nuanced, localized context of a specific election cycle. In addition, the indirect nature of NSUM amplifies uncertainties related to network size assumptions and homophily, which may interact with the internal representations of LLMs in non-trivial ways. In turn, the evaluation is limited to a small number of electoral contexts, constraining the generalization of the findings. Finally, the ethical and epistemological considerations of substituting or complementing human respondents with AI systems require a deeper, ongoing scrutiny beyond the technical scope of this paper.

Future work should focus on several directions. Firstly, expanding the methodology to incorporate more complex network structures and homophily parameters within the simulation prompts could enhance the realism of synthetic social circles. Secondly, a systematic investigation into prompt engineering is needed to improve the robustness and reduce the variance of responses across different LLM providers and versions. Third, it is also necessary to systematically study prompt sensitivity, calibration techniques, and hybrid approaches that combine synthetic and human survey data. Lastly, a fruitful line of inquiry would be to develop hybrid models that intelligently combine synthetic LLM data with traditional survey data to improve cost-efficiency, scale, and possibly the accuracy of electoral forecasts.

Data Availability Statement

The synthetic data generated by the LLMs, as well as the prompts and codes designed in this work, can be found in https://github.com/GCGImdea/NSUM_LLMs.git. (*este link es provisional*)

Author Contribution Statement

All authors contributed to different parts of the work: writing, experimentation, discussion of results, etc.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported by grant TSI-064100-2023-34 (INES), funded by the Spanish Ministry of Digital Transformation and Civil Service and the European Union-NextGenerationEU/PRTR. The research is part of the projects PID2022-140560OB-I00 (DRONAC) funded by MICIU/AEI /10.13039/501100011033 and ERDF, and MADQuantum-CM, financed by the European Union NextGeneration-EU, and Madrid Government.

References

- [AAK23] Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org, 2023.
- [ABF⁺23] L. Argyle, E. Busby, N. Fulda, J. Gubler, C. Rytting, and D. Wingate. Hello gpt-4o: Our new flagship model can reason across audio, vision, and text in real time. *Political Analysis*, 31:337—351, 2023.
- [ARDA⁺25] J. Arevalillo, J. Ramirez, S. Díaz-Aranda, J. Aguilar, A. Fernandez Anta, and R. Lillo. Network Scale-up Methods on Aggregated Relational Data to Estimate the Outcome of Elections. *Submitted to publication*, 2025.

- [BCD⁺24] James Bisbee, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416, 2024.
- [BCL⁺23] Emily Breza, Arun G Chandrasekhar, Shane Lubold, Tyler H McCormick, and Mengjie Pan. Consistently estimating network statistics using aggregated relational data. *Proceedings of the National Academy of Sciences*, 120(21):e2207185120, 2023.
- [BCMP20] Emily Breza, Arun G Chandrasekhar, Tyler H McCormick, and Mengjie Pan. Using aggregated relational data to feasibly identify network structure without network data. *American Economic Review*, 110(8):2454–84, 2020.
- [BD23] Jawid Ahmad Baktash and Mursal Dawodi. Gpt-4: A review on advancements and opportunities in natural language processing. *arXiv preprint arXiv:2305.03195*, 2023.
- [BdBGB⁺22] Wändi Bruine de Bruin, Mirta Galesic, Rasmus Bååth, Jochem de Bresser, Lars Hall, Petter Johansson, Thomas Strandberg, and Arthur van Soest. Asking about Social Circles Improves Election Predictions Even with Many Political Parties. *International Journal of Public Opinion Research*, 34(1):edac006, 04 2022.
- [CKK24] Suhyun Cho, Jaeyun Kim, and Jang Hyun Kim. Llm-based dopelgänger models: Leveraging synthetic data for human-like responses in survey simulations. *IEEE Access*, 12:178917–178927, 2024.
- [CSMG17] Giulia Carletti, Nicola Soriani, Martina Mattiazzi, and Dario Gregori. Suppl-1, m6: A social network approach to the estimation of perceived quality of health care. *The Open Nursing Journal*, 11:219, 2017.
- [DAAR⁺25] Sergio Díaz-Aranda, Jose Aguilar, Juan Marcos Ramírez, David Rabanedo, Antonio Fernández Anta, and Rosa E Lillo. Performance analysis of NSUM estimators in social-network topologies. *The American Statistician*, 79(2):247–264, 2025.
- [DTGG23] Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600, 2023.
- [FUM⁺16] Dennis M Feehan, Aline Umubyeyi, Mary Mahy, Wolfgang Hladik, and Matthew J Salganik. Quantity versus quality: A survey experiment to improve the network scale-up method. *American journal of epidemiology*, 183(8):747–757, 2016.

- [GAOHR⁺21] Augusto Garcia-Agundez, Oluwasegun Ojo, Harold A Hernández-Roig, Carlos Baquero, Davide Frey, Chryssis Georgiou, Mathieu Goessens, Rosa E Lillo, Raquel Menezes, Nicolas Nicolaou, et al. Estimating the COVID-19 prevalence in Spain with indirect reporting via open surveys. *Frontiers in Public Health*, 9:658544, 2021.
- [GBdBD⁺18] Mirta Galesic, Wändi Bruine de Bruin, Marion Dumas, A Kapteyn, JE Darling, and Erik Meijer. Asking about social circles improves election predictions. *Nature Human Behaviour*, 2(3):187–193, 2018.
- [Gra14] Andreas Graefe. Accuracy of Vote Expectation Surveys in Forecasting Elections. *Public Opinion Quarterly*, 78(S1):204–232, 06 2014.
- [Gra16] Andreas Graefe. Forecasting proportional representation elections from non-representative expectation surveys. *Electoral Studies*, 42:222–228, 2016.
- [HDK15] Patrick Habecker, Kirk Dombrowski, and Bilal Khan. Improving the network scale-up estimator: Incorporating means of sums, recursive back estimation, and sampling weights. *PloS one*, 10(12):e0143406, 2015.
- [HWW25] Chengpiao Huang, Yuhang Wu, and Kaizheng Wang. Uncertainty quantification for LLM-based survey simulations. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 25947–25971. PMLR, 13–19 Jul 2025.
- [JgJS23] Bernard J. Jansen, Soon gyo Jung, and Joni Salminen. Employing large language models in survey research. *Natural Language Processing Journal*, 4:100020, 2023.
- [JKR89] E. C. Johnsen, P. D. Killworth, and S. Robinson. Estimating the size of an average personal network and of an event subpopulation. *The Small World*, page 159–175, 1989.
- [JLC⁺18] L Jing, Q Lu, Y Cui, H Yu, and T Wang. Combining the randomized response technique and the network scale-up method to estimate the female sex worker population size: an exploratory study. *Public Health*, 160:81–86, 2018.
- [KJM⁺98] Peter D Killworth, Eugene C Johnsen, Christopher McCarty, Gene Ann Shelley, and H Russell Bernard. A social network approach to estimating seroprevalence in the united states. *Social networks*, 20(1):23–50, 1998.

- [KKM⁺25] Carolin Kaiser, Jakob Kaiser, Vladimir Manewitsch, Lea Rau, and Rene Schallner. Simulating human opinions with large language models: Opportunities and challenges for personalized survey data modeling. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP Adjunct '25, page 82–86, New York, NY, USA, 2025. Association for Computing Machinery.
- [KL24] Junsol Kim and Byungkyu Lee. Ai-augmented surveys: Leveraging large language models and surveys for opinion prediction, 2024.
- [KMB⁺98] Peter D Killworth, Christopher McCarty, H Russell Bernard, Gene Ann Shelley, and Eugene C Johnsen. Estimation of seroprevalence, rape, and homelessness in the united states using a social network approach. *Evaluation review*, 22(2):289–308, 1998.
- [LBN21] Ian Laga, Le Bao, and Xiaoyue Niu. Thirty years of the network scale-up method. *Journal of the American Statistical Association*, 116(535):1548–1559, 2021.
- [LBN23] Ian Laga, Le Bao, and Xiaoyue Niu. A correlated network scale-up model: Finding the connection between subpopulations. *Journal of the American Statistical Association*, pages 1–10, 2023.
- [Leo25] Maikel Leon. Gpt-5 and open-weight large language models: Advances in reasoning, transparency, and control. *Information Systems*, page 102620, 2025.
- [LLK⁺13] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, 2013.
- [Man67] Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2.Part.1):209–220, 1967.
- [McC20] Tyler H McCormick. The network scale-up method. *The Oxford Handbook of Social Networks*, page 153, 2020.
- [MMMR24] Fabio Motoki, Januário Monteiro, Ricardo Malagueño, and Victor Rangel. Large language models in survey research: Generating synthetic data and unlocking new possibilities. *SSRN*, 2024.
- [Mur11] Andreas Erwin Murr. “Wisdom of crowds”? A decentralised election forecasting model that uses citizens’ local expectations. *Electoral Studies*, 30(4):771–783, 2011.

- [MXYJ24] Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O. Jackson. A turing test of whether ai chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences*, 121(9):e2313925121, 2024.
- [MZ12] Tyler H McCormick and Tian Zheng. Latent demographic profile estimation in hard-to-reach groups. *The annals of applied statistics*, 6(4):1795, 2012.
- [MZ15] Tyler H McCormick and Tian Zheng. Latent surface models for networks using aggregated relational data. *Journal of the American Statistical Association*, 110(512):1684–1695, 2015.
- [NAMS⁺26] Albert Nyarko-Agyei, Scott Moser, Rowland G Seymour, Ben Brewster, Sabrina L Li, Esther Weir, Todd Landman, Emily Wyman, Christine Belle Torres, Imogen Fell, et al. A partially pooled network scale-up method model: detailed estimation of child sexual exploitation material trafficking prevalence in philippine municipalities. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 75(1):43–56, 2026.
- [Ope23] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [Ope24] OpenAI. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*, 2024.
- [PNB22] Jacob Parsons, Xiaoyue Niu, and Le Bao. A Bayesian hierarchical model for combining multiple data sources in population size estimation. *The Annals of Applied Statistics*, 16(3):1550–1562, 2022.
- [RDABF26] Juan Marcos Ramírez, Sergio Díaz-Aranda, Leonhard Balduf, and Antonio Fernández Anta. Beyond crawlers: Ipfs characterization using network scale-up method. (Submitted article), 2026.
- [SFM⁺10] MJ Salganik, D Fazito, MB Mello, et al. Estimating the number of heavy drug users in Curitiba, Brazil using multiple methods. Technical report, Technical report. UNAIDS, 2010.
- [SGAD20] Elizabeth Sully, Margaret Giorgio, and Selena Anjur-Dietrich. Estimating abortion incidence using the network scale-up method. *Demographic Research*, 43:1651–1684, 2020.
- [SMS⁺26] Ryan S Shean, Jayanth Kumar Mallapu, Tathya Shah, Haroon Adam Rasheed, David N Younessi, Yih Chung Tham, Van Nguyen, Kyle Bolo, and Benjamin Y Xu. Comparative performance of gemini 3 pro and gpt-5 family models on ophthalmology board-style questions. *Ophthalmology Science*, page 101145, 2026.

- [TPC⁺19] Alvin Kuo Jing Teo, Kiesha Prem, Mark I C Chen, Adrian Roellin, Mee Lian Wong, Hanh Hao La, and Alex R Cook. Estimating the size of key populations for HIV in singapore using the network scale-up method. *Sex Transm Infect*, 95(8):602–607, 2019.
- [UNA10] WHO UNAIDS. Guidelines on estimating the size of populations most at risk to hiv. *Geneva, Switzerland: World Health Organization*, 51, 2010.
- [VZBB26] David Villarreal-Zegarra and Luciana Bellido-Boza. Generation of synthetic data in health surveys using large language models. *medRxiv*, 2026.

A Supplementary Material

A.1 Spanish elections

A.1.1 Complete Dimension Analysis

As mentioned in 5.2.1, the tables below correspond to the complete analysis for the Spanish scenario. Note that the missing gaps are due to computational difficulties in working with GPT 4o through the Chat interface.

Model	MADRID		ANDALUSIA		VALENCIA	
	Age	Age + Income	Age	Age + Income	Age	Age + Income
<i>Panel A: API</i>						
GPT 4	10.84	11.67	10.96	11.24	11.58	10.8
GPT 4o - mini	6.86	7.66	6.65	7.37	7.12	6.91
GPT 5	3.77	3.72	3.95	3.92	4.28	4.25
<i>Panel B: CHAT</i>						
GPT 4o	3.38	3.79	–	3.48	–	3.99
GPT 5	4.10	3.46	3.86	3.28	4.21	3.42

Table SM1: Complete comparison by region and input demographic variables using RMSE and PIMLE, considering blank vote (Bold values indicate the best performing scenario in each case).

Model	MADRID		ANDALUSIA		VALENCIA	
	Original P.	Perfect P.	Original P.	Perfect P.	Original P.	Perfect P.
<i>Panel A: API</i>						
GPT 4	11.54	10.97	11.25	10.95	11.04	12.09
GPT 4o - mini	6.91	7.61	6.52	7.51	6.57	7.46
GPT 5	3.98	3.51	4.30	3.57	4.28	4.22
<i>Panel B: CHAT</i>						
GPT 4o	3.11	4.75	3.48	–	3.99	–
GPT 5	4.14	3.42	4.17	2.97	3.82	3.80

Table SM2: Complete comparison between Original and Perfect prompt across regions using RMSE and PIMLE, considering blank vote (Bold values indicate the best performing scenario for each model).

The following subsections present the experimental results for each iteration of the study, categorized by geographical region and data collection channel (API and Online). Each table provides a comparative analysis of vote intention percentages across different LLM models and NSUM estimation methods. Additionally, this section includes with the questionnaire administered to respondents regarding the Spanish general elections in the regions of Madrid,

Andalusia, and Valencia. Finally, both the original and the enhanced prompt administered to the LLM are specified for the region of Madrid.

A.1.2 Results - Madrid

Survey	PP	PSOE	VOX	SUMAR	BLANK
Real Results	40.2	27.7	13.9	15.3	2.9
Human Survey	39.2	27.6	14.6	14.7	3.8
API · P. original · Age · GPT-5 · Naive	34.0	25.0	16.7	16.4	7.8
API · P. original · Age · GPT-5 · MoS	34.1	25.0	16.8	16.4	7.8
API · P. original · Age · GPT-5 · MLE	34.2	25.0	16.8	16.2	7.8
API · P. original · Age · GPT-5 · PIMLE	34.1	25.0	16.8	16.4	7.8
API · P. original · Age and income · GPT-5 · Naive	33.9	24.9	16.4	17.0	7.8
API · P. original · Age and income · GPT-5 · MoS	33.8	25.0	16.4	17.0	7.8
API · P. original · Age and income · GPT-5 · MLE	34.0	24.9	16.4	17.0	7.8
API · P. original · Age and income · GPT-5 · PIMLE	33.9	24.9	16.4	17.0	7.8
API · P. original · Age · GPT-4 · Naive	24.9	33.1	11.3	9.8	21.0
API · P. original · Age · GPT-4 · MoS	25.0	33.2	11.3	9.8	20.7
API · P. original · Age · GPT-4 · MLE	24.9	33.0	11.2	10.0	21.0
API · P. original · Age · GPT-4 · PIMLE	25.0	33.1	11.3	9.8	20.7
API · P. original · Age and income · GPT-4 · Naive	24.3	34.9	10.8	7.9	22.1
API · P. original · Age and income · GPT-4 · MoS	24.3	35.0	10.8	8.1	21.8
API · P. original · Age and income · GPT-4 · MLE	24.5	34.5	10.9	8.5	21.6
API · P. original · Age and income · GPT-4 · PIMLE	24.4	34.9	10.8	8.1	21.8
API · P. original · Age · GPT-4o-mini · Naive	33.0	25.7	13.3	13.4	14.7
API · P. original · Age · GPT-4o-mini · MoS	32.8	25.6	13.2	13.7	14.7
API · P. original · Age · GPT-4o-mini · MLE	33.1	25.7	13.6	13.0	14.6
API · P. original · Age · GPT-4o-mini · PIMLE	32.8	25.6	13.3	13.5	14.8
API · P. original · Age and income · GPT-4o-mini · Naive	29.1	27.7	11.4	17.3	14.5
API · P. original · Age and income · GPT-4o-mini · MoS	28.9	27.6	11.5	17.5	14.5
API · P. original · Age and income · GPT-4o-mini · MLE	29.0	27.7	11.6	17.4	14.3
API · P. original · Age and income · GPT-4o-mini · PIMLE	28.8	27.6	11.4	17.6	14.5

Table SM3: Vote Intention Percentage Results Madrid (API - Original Prompt)

The acronyms of the parties in all the tables represent respectively: *Partido Popular* (PP), *Partido Socialista Obrero Español* (PSOE), Vox, Sumar.

Survey	PP	PSOE	VOX	SUMAR	BLANK
Real Results	40.2	27.7	13.9	15.3	2.9
Human Survey	39.2	27.6	14.6	14.7	3.8
API · P. perfect · Age · GPT-5 · Naive	36.6	23.4	16.7	15.4	8.0
API · P. perfect · Age · GPT-5 · MoS	36.7	23.4	16.7	15.3	7.9
API · P. perfect · Age · GPT-5 · MLE	36.6	23.4	16.7	15.3	8.0
API · P. perfect · Age · GPT-5 · PIMLE	36.6	23.4	16.7	15.3	7.9
API · P. perfect · Age and income · GPT-5 · Naive	37.3	23.1	17.0	15.2	7.5
API · P. perfect · Age and income · GPT-5 · MoS	37.3	23.1	17.0	15.1	7.4
API · P. perfect · Age and income · GPT-5 · MLE	37.5	22.9	17.1	15.1	7.4
API · P. perfect · Age and income · GPT-5 · PIMLE	37.4	23.1	17.0	15.1	7.4
API · P. perfect · Age · GPT-4 · Naive	29.1	33.0	12.3	5.4	20.2
API · P. perfect · Age · GPT-4 · MoS	28.8	33.3	12.3	5.4	20.3
API · P. perfect · Age · GPT-4 · MLE	28.8	33.0	12.5	5.6	20.2
API · P. perfect · Age · GPT-4 · PIMLE	28.9	33.2	12.3	5.4	20.3
API · P. perfect · Age and income · GPT-4 · Naive	27.6	32.5	12.1	5.4	22.4
API · P. perfect · Age and income · GPT-4 · MoS	27.8	32.3	12.2	5.6	22.2
API · P. perfect · Age and income · GPT-4 · MLE	28.0	32.3	12.2	5.5	21.9
API · P. perfect · Age and income · GPT-4 · PIMLE	27.8	32.2	12.1	5.6	22.2
API · P. perfect · Age · GPT-4o-mini · Naive	30.9	25.9	13.2	14.3	15.6
API · P. perfect · Age · GPT-4o-mini · MoS	30.4	25.6	13.0	15.3	15.7
API · P. perfect · Age · GPT-4o-mini · MLE	31.2	26.0	13.3	14.0	15.6
API · P. perfect · Age · GPT-4o-mini · PIMLE	30.2	25.8	12.9	15.4	15.7
API · P. perfect · Age and income · GPT-4o-mini · Naive	27.6	29.7	12.7	15.4	14.6
API · P. perfect · Age and income · GPT-4o-mini · MoS	27.3	29.7	12.6	15.7	14.8
API · P. perfect · Age and income · GPT-4o-mini · MLE	28.0	32.3	12.2	5.5	21.9
API · P. perfect · Age and income · GPT-4o-mini · PIMLE	27.4	29.6	12.6	15.5	14.8

Table SM4: Vote Intention Percentage Results Madrid (API - Perfect Prompt)

Survey	PP	PSOE	VOX	SUMAR	BLANK
Real Results	40.2	27.7	13.9	15.3	2.9
Human Survey	39.2	27.6	14.6	14.7	3.8
ONLINE · P. original · Age · GPT-5 · Naive	34.8	24.5	18.4	14.4	7.7
ONLINE · P. original · Age · GPT-5 · MoS	34.7	24.7	18.6	14.3	7.7
ONLINE · P. original · Age · GPT-5 · MLE	34.8	24.8	18.5	14.3	7.6
ONLINE · P. original · Age · GPT-5 · PIMLE	34.7	24.7	18.5	14.4	7.5
ONLINE · P. original · Age and income · GPT-5 · Naive	32.8	28.8	12.7	17.7	8.2
ONLINE · P. original · Age and income · GPT-5 · MoS	33.1	28.8	12.7	17.8	8.2
ONLINE · P. original · Age and income · GPT-5 · MLE	33.5	28.9	12.6	17.7	8.2
ONLINE · P. original · Age and income · GPT-5 · PIMLE	32.9	28.8	12.5	17.7	8.1
ONLINE · P. original · Age · GPT-4o · Naive	36.5	25.3	14.9	14.3	8.9
ONLINE · P. original · Age · GPT-4o · MoS	36.5	25.3	14.9	14.3	8.9
ONLINE · P. original · Age · GPT-4o · MLE	36.6	25.3	14.9	14.2	8.9
ONLINE · P. original · Age · GPT-4o · PIMLE	36.5	25.3	14.9	14.4	8.9
ONLINE · P. original · Age and income · GPT-4o · Naive	37.6	25.9	13.4	14.6	8.4
ONLINE · P. original · Age and income · GPT-4o · MoS	37.9	25.8	13.5	14.4	8.4
ONLINE · P. original · Age and income · GPT-4o · MLE	38.6	25.6	13.8	13.7	8.7
ONLINE · P. original · Age and income · GPT-4o · PIMLE	37.9	25.8	13.5	14.4	8.4
ONLINE · P. perfect · Age · GPT-5 · Naive	35.4	22.9	14.3	18.9	8.5
ONLINE · P. perfect · Age · GPT-5 · MoS	35.7	23.0	14.4	18.4	8.5
ONLINE · P. perfect · Age · GPT-5 · MLE	35.5	22.9	14.4	18.7	8.6
ONLINE · P. perfect · Age · GPT-5 · PIMLE	35.6	22.9	14.4	18.5	8.5
ONLINE · P. perfect · Age and income · GPT-5 · Naive	32.2	27.2	13.8	17.7	9.0
ONLINE · P. perfect · Age and income · GPT-5 · MoS	32.3	27.1	13.8	17.9	8.9
ONLINE · P. perfect · Age and income · GPT-5 · MLE	32.6	26.9	13.8	17.9	8.8
ONLINE · P. perfect · Age and income · GPT-5 · PIMLE	32.2	27.1	13.8	17.9	8.9
ONLINE · P. perfect · Age and income · GPT-4o · Naive	31.7	27.7	12.4	19.8	8.5
ONLINE · P. perfect · Age and income · GPT-4o · MoS	31.8	27.8	12.5	19.4	8.5
ONLINE · P. perfect · Age and income · GPT-4o · MLE	32.2	27.3	12.7	19.4	8.5
ONLINE · P. perfect · Age and income · GPT-4o · PIMLE	32.1	27.6	12.6	19.2	8.5

Table SM5: Vote Intention Percentage Results Madrid (Chat)

A.1.3 Results - Andalusia

Survey	PP	PSOE	VOX	SUMAR	BLANK
Real Results	36.4	33.5	15.3	12.0	2.8
Human Survey	35.3	28.3	15.3	12.6	8.5
ONLINE · P. perfect · Age · GPT-5 · Naive	34.1	28.3	16.9	14.2	6.5
ONLINE · P. perfect · Age · GPT-5 · MoS	34.4	28.3	17.1	13.7	6.5
ONLINE · P. perfect · Age · GPT-5 · MLE	34.8	28.1	17.2	13.4	6.5
ONLINE · P. perfect · Age · GPT-5 · PIMLE	34.5	28.3	17.1	13.7	6.5
ONLINE · P. perfect · Age and income · GPT-5 · Naive	33.8	29.6	16.2	14.2	6.3
ONLINE · P. perfect · Age and income · GPT-5 · MoS	34.2	29.5	16.3	13.9	6.2
ONLINE · P. perfect · Age and income · GPT-5 · MLE	34.3	29.2	16.3	14.0	6.2
ONLINE · P. perfect · Age and income · GPT-5 · PIMLE	34.4	29.3	16.3	13.8	6.2
ONLINE · P. original · Age · GPT-5 · Naive	33.3	26.2	18.3	14.6	7.6
ONLINE · P. original · Age · GPT-5 · MoS	33.5	26.1	18.4	14.5	7.5
ONLINE · P. original · Age · GPT-5 · MLE	33.6	26.2	18.4	14.4	7.4
ONLINE · P. original · Age · GPT-5 · PIMLE	33.4	26.1	18.4	14.6	7.5
ONLINE · P. original · Age and income · GPT-5 · Naive	33.2	27.9	16.9	14.5	7.5
ONLINE · P. original · Age and income · GPT-5 · MoS	33.3	27.8	17.0	14.5	7.4
ONLINE · P. original · Age and income · GPT-5 · MLE	33.5	27.6	17.1	14.4	7.3
ONLINE · P. original · Age and income · GPT-5 · PIMLE	33.3	27.9	17.0	14.5	7.4
ONLINE · P. original · Age and income · GPT-4o · Naive	37.4	27.5	14.7	12.9	7.5
ONLINE · P. original · Age and income · GPT-4o · MoS	37.5	27.5	14.7	12.9	7.5
ONLINE · P. original · Age and income · GPT-4o · MLE	38.5	27.0	14.9	12.2	7.5
ONLINE · P. original · Age and income · GPT-4o · PIMLE	37.6	27.5	14.7	12.8	7.5

Table SM6: Vote Intention Percentage Results Andalusia (Chat)

Survey	PP	PSOE	VOX	SUMAR	BLANK
Real Results	36.4	33.5	15.3	12.0	2.8
Human Survey	35.3	28.3	15.3	12.6	8.5
API · P. original · Age · GPT-5 · Naive	32.5	27.4	17.6	14.4	8.2
API · P. original · Age · GPT-5 · MoS	32.5	27.4	17.6	14.3	8.2
API · P. original · Age · GPT-5 · MLE	32.5	27.4	17.5	14.4	8.2
API · P. original · Age · GPT-5 · PIMLE	32.5	27.4	17.6	14.3	8.1
API · P. original · Age and income · GPT-5 · Naive	32.3	27.4	17.6	14.6	8.1
API · P. original · Age and income · GPT-5 · MoS	32.2	27.4	17.6	14.8	8.1
API · P. original · Age and income · GPT-5 · MLE	32.4	27.3	17.6	14.5	8.1
API · P. original · Age and income · GPT-5 · PIMLE	32.3	27.4	17.6	14.6	8.1
API · P. original · Age · GPT-4 · Naive	21.8	39.2	11.2	6.5	21.4
API · P. original · Age · GPT-4 · MoS	21.8	38.9	11.3	6.7	21.3
API · P. original · Age · GPT-4 · MLE	22.1	38.7	11.2	6.8	21.3
API · P. original · Age · GPT-4 · PIMLE	21.7	39.0	11.3	6.7	21.3
API · P. original · Age and income · GPT-4 · Naive	23.1	37.5	11.3	7.7	20.5
API · P. original · Age and income · GPT-4 · MoS	23.2	37.1	11.2	8.0	20.5
API · P. original · Age and income · GPT-4 · MLE	23.1	36.8	11.4	8.3	20.4
API · P. original · Age and income · GPT-4 · PIMLE	23.1	37.2	11.2	7.9	20.6
API · P. original · Age · GPT-4o-mini · Naive	33.6	26.2	13.9	11.4	14.9
API · P. original · Age · GPT-4o-mini · MoS	33.4	26.3	14.0	11.3	15.0
API · P. original · Age · GPT-4o-mini · MLE	33.6	26.3	14.2	11.0	14.9
API · P. original · Age · GPT-4o-mini · PIMLE	33.4	26.2	14.0	11.4	15.0
API · P. original · Age and income · GPT-4o-mini · Naive	29.8	28.0	12.4	15.0	14.8
API · P. original · Age and income · GPT-4o-mini · MoS	29.3	28.3	12.5	15.2	14.7
API · P. original · Age and income · GPT-4o-mini · MLE	29.4	28.2	12.7	15.2	14.5
API · P. original · Age and income · GPT-4o-mini · PIMLE	29.5	28.1	12.5	15.2	14.7

Table SM7: Vote Intention Percentage Results Andalusia (API - Original Prompt)

Survey	PP	PSOE	VOX	SUMAR	BLANK
Real Results	36.4	33.5	15.3	12.0	2.8
Human Survey	35.3	28.3	15.3	12.6	8.5
API · P. perfect · Age · GPT-5 · Naive	34.6	27.8	16.7	12.7	8.2
API · P. perfect · Age · GPT-5 · MoS	34.6	27.9	16.7	12.7	8.1
API · P. perfect · Age · GPT-5 · MLE	34.7	27.9	16.7	12.6	8.1
API · P. perfect · Age · GPT-5 · PIMLE	34.7	27.8	16.7	12.7	8.1
API · P. perfect · Age and income · GPT-5 · Naive	34.2	28.1	17.1	12.9	7.7
API · P. perfect · Age and income · GPT-5 · MoS	34.3	28.1	17.1	12.8	7.7
API · P. perfect · Age and income · GPT-5 · MLE	34.5	28.0	17.1	12.8	7.7
API · P. perfect · Age and income · GPT-5 · PIMLE	34.3	28.1	17.1	12.8	7.7
API · P. perfect · Age · GPT-4 · Naive	23.3	39.6	12.2	4.6	20.4
API · P. perfect · Age · GPT-4 · MoS	23.4	39.3	12.3	4.7	20.4
API · P. perfect · Age · GPT-4 · MLE	23.3	39.3	12.3	4.8	20.3
API · P. perfect · Age · GPT-4 · PIMLE	23.4	39.4	12.3	4.7	20.3
API · P. perfect · Age and income · GPT-4 · Naive	21.9	41.2	12.2	4.5	20.2
API · P. perfect · Age and income · GPT-4 · MoS	22.1	41.3	12.2	4.4	20.0
API · P. perfect · Age and income · GPT-4 · MLE	22.2	40.3	12.5	4.8	20.2
API · P. perfect · Age and income · GPT-4 · PIMLE	22.0	41.1	12.3	4.5	20.2
API · P. perfect · Age · GPT-4o-mini · Naive	36.0	24.0	14.9	9.6	15.5
API · P. perfect · Age · GPT-4o-mini · MoS	35.6	24.0	14.7	10.3	15.5
API · P. perfect · Age · GPT-4o-mini · MLE	36	24.0	14.9	9.5	15.5
API · P. perfect · Age · GPT-4o-mini · PIMLE	35.6	24.0	14.7	10.2	15.5
API · P. perfect · Age and income · GPT-4o-mini · Naive	23.8	36.2	11.9	15.4	12.7
API · P. perfect · Age and income · GPT-4o-mini · MoS	23.2	35.5	11.6	17.1	12.7
API · P. perfect · Age and income · GPT-4o-mini · MLE	23.3	35.7	11.7	16.7	12.7
API · P. perfect · Age and income · GPT-4o-mini · PIMLE	23.3	35.7	11.7	16.7	12.7

Table SM8: Vote Intention Percentage Results Andalusia (API - Perfect Prompt)

A.1.4 Results - Valencia

Survey	PP	PSOE	VOX	SUMAR	BLANK
Real Results	34.9	32.1	15.7	15.2	2.2
Human Survey	32.3	27.7	16.5	15.0	8.5
ONLINE · P. original · Age · GPT-5 · Naive	32.8	25.7	17.6	16.7	7.3
ONLINE · P. original · Age · GPT-5 · MoS	32.7	25.8	17.6	16.7	7.3
ONLINE · P. original · Age · GPT-5 · MLE	33.1	25.7	17.7	16.4	7.1
ONLINE · P. original · Age · GPT-5 · PIMLE	32.6	25.8	17.5	16.8	7.3
ONLINE · P. original · Age and income · GPT-5 · Naive	31.1	28.0	15.5	17.6	7.9
ONLINE · P. original · Age and income · GPT-5 · MoS	31.0	28.1	15.5	17.5	7.9
ONLINE · P. original · Age and income · GPT-5 · MLE	31.3	27.7	15.4	17.8	7.8
ONLINE · P. original · Age and income · GPT-5 · PIMLE	31.0	28.1	15.5	17.6	7.9
ONLINE · P. perfect · Age · GPT-5 · Naive	34.9	24.0	15.8	17.3	8.0
ONLINE · P. perfect · Age · GPT-5 · MoS	35.3	24.2	16.0	16.7	7.8
ONLINE · P. perfect · Age · GPT-5 · MLE	34.5	24.1	15.7	17.7	8.1
ONLINE · P. perfect · Age · GPT-5 · PIMLE	35.2	24.2	15.9	16.8	8.0
ONLINE · P. perfect · Age and income · GPT-5 · Naive	33.8	27.4	15.5	16.5	6.9
ONLINE · P. perfect · Age and income · GPT-5 · MoS	34.3	27.2	15.6	16.1	6.9
ONLINE · P. perfect · Age and income · GPT-5 · MLE	33.8	27.4	15.5	16.6	6.8
ONLINE · P. perfect · Age and income · GPT-5 · PIMLE	34.5	27.1	15.6	16.0	6.8
ONLINE · P. original · Age and income · GPT-4o · Naive	36.5	26.2	13.6	15.4	8.3
ONLINE · P. original · Age and income · GPT-4o · MoS	36.8	26.0	13.7	15.3	8.2
ONLINE · P. original · Age and income · GPT-4o · MLE	37.8	25.8	14.0	14.4	8.0
ONLINE · P. original · Age and income · GPT-4o · PIMLE	36.8	26.1	13.7	15.2	8.2

Table SM9: Vote Intention Percentage Results Valencia (Chat)

Survey	PP	PSOE	VOX	SUMAR	BLANK
Real Results	34.9	32.1	15.7	15.2	2.2
Human Survey	32.3	27.7	16.5	15.0	8.5
API · P. original · Age · GPT-5 · Naive	31.3	26.0	17.5	17.0	8.2
API · P. original · Age · GPT-5 · MoS	31.4	26.0	17.6	16.9	8.2
API · P. original · Age · GPT-5 · MLE	31.4	26.1	17.6	16.9	8.1
API · P. original · Age · GPT-5 · PIMLE	31.3	26.0	17.5	16.9	8.2
API · P. original · Age and income · GPT-5 · Naive	32.2	25.4	17.5	16.8	8.1
API · P. original · Age and income · GPT-5 · MoS	32.2	25.4	17.5	16.8	8.1
API · P. original · Age and income · GPT-5 · MLE	32.3	25.4	17.5	16.7	8.1
API · P. original · Age and income · GPT-5 · PIMLE	32.3	25.4	17.5	16.8	8.1
API · P. original · Age · GPT-4 · Naive	22.5	34.5	10.9	9.4	22.7
API · P. original · Age · GPT-4 · MoS	22.4	34.3	10.8	9.6	22.9
API · P. original · Age · GPT-4 · MLE	22.7	34.3	11.0	9.7	22.4
API · P. original · Age · GPT-4 · PIMLE	22.5	34.5	10.9	9.4	22.7
API · P. original · Age and income · GPT-4 · Naive	22.9	34.7	11.2	9.3	21.9
API · P. original · Age and income · GPT-4 · MoS	23.0	34.7	11.2	9.6	21.7
API · P. original · Age and income · GPT-4 · MLE	23.2	34.2	11.1	9.7	21.8
API · P. original · Age and income · GPT-4 · PIMLE	23.0	34.6	11.1	9.5	21.7
API · P. original · Age · GPT-4o-mini · Naive	34.5	24.6	13.1	13.7	14.1
API · P. original · Age · GPT-4o-mini · MoS	34.3	24.6	13.1	13.5	14.5
API · P. original · Age · GPT-4o-mini · MLE	34.5	24.6	13.4	13.5	14.2
API · P. original · Age · GPT-4o-mini · PIMLE	34.5	24.5	13.2	13.7	14.2
API · P. original · Age and income · GPT-4o-mini · Naive	29.1	27.4	11.5	18.1	13.9
API · P. original · Age and income · GPT-4o-mini · MoS	28.8	27.4	11.6	18.3	13.9
API · P. original · Age and income · GPT-4o-mini · MLE	28.9	27.5	11.9	18.1	13.6
API · P. original · Age and income · GPT-4o-mini · PIMLE	28.9	27.4	11.6	18.3	13.8

Table SM10: Vote Intention Percentage Results Valencia (API - Original Prompt)

Survey	PP	PSOE	VOX	SUMAR	BLANK
Real Results	34.9	32.1	15.7	15.2	2.2
Human Survey	32.3	27.7	16.5	15.0	8.5
API · P. perfect · Age · GPT-5 · Naive	32.6	25.3	17.6	16.2	8.3
API · P. perfect · Age · GPT-5 · MoS	32.6	25.4	17.5	16.2	8.3
API · P. perfect · Age · GPT-5 · MLE	32.7	25.3	17.6	16.1	8.4
API · P. perfect · Age · GPT-5 · PIMLE	32.7	25.4	17.6	16.1	8.3
API · P. perfect · Age and income · GPT-5 · Naive	33.5	24.9	17.8	16.1	7.8
API · P. perfect · Age and income · GPT-5 · MoS	33.7	24.8	17.8	15.9	7.8
API · P. perfect · Age and income · GPT-5 · MLE	33.8	24.7	17.9	16.0	7.7
API · P. perfect · Age and income · GPT-5 · PIMLE	33.6	24.9	17.8	16.0	7.7
API · P. perfect · Age · GPT-4 · Naive	25.1	34.1	11.2	5.5	24.1
API · P. perfect · Age · GPT-4 · MoS	25.0	33.9	11.3	5.6	24.1
API · P. perfect · Age · GPT-4 · MLE	25.3	33.9	11.3	5.8	23.7
API · P. perfect · Age · GPT-4 · PIMLE	25.0	33.9	11.2	5.6	24.3
API · P. perfect · Age and income · GPT-4 · Naive	23.7	35.3	10.3	5.9	24.8
API · P. perfect · Age and income · GPT-4 · MoS	23.4	35.4	10.3	6.2	24.7
API · P. perfect · Age and income · GPT-4 · MLE	24.1	34.8	10.6	6.1	24.4
API · P. perfect · Age and income · GPT-4 · PIMLE	23.5	35.4	10.3	6.2	24.7
API · P. perfect · Age · GPT-4o-mini · Naive	24.3	32.0	12.4	16.3	15.0
API · P. perfect · Age · GPT-4o-mini · MoS	24.4	31.4	12.3	16.7	15.2
API · P. perfect · Age · GPT-4o-mini · MLE	24.5	32.0	12.5	15.8	15.2
API · P. perfect · Age · GPT-4o-mini · PIMLE	24.1	31.7	12.3	16.7	15.2
API · P. perfect · Age and income · GPT-4o-mini · Naive	25.7	31.3	12.4	16.2	14.4
API · P. perfect · Age and income · GPT-4o-mini · MoS	25.2	31.3	12.3	16.8	14.6
API · P. perfect · Age and income · GPT-4o-mini · MLE	26.1	31.2	12.5	15.6	14.6
API · P. perfect · Age and income · GPT-4o-mini · PIMLE	25.4	31.3	12.3	16.4	14.6

Table SM11: Vote Intention Percentage Results Valencia (API - Perfect Prompt)

A.1.5 Questionnaire for the General Election in Madrid, Andalusia and Valencia

- **Q1. Scientific Study Information**

We are a team of computer and statistics researchers. We collect data to estimate the vote intention in different Spain regions, with the goal of improving statistical estimation techniques. The following survey contains questions regarding the voting intention of the people that you know for the Spanish general election on July 23, 2023. These contacts can include family, friends, and colleagues whose voting intention you know.

- Responses are anonymous. The researchers cannot identify the participants.
- Data cannot be modified or deleted once the survey is completed.
- The data collected may be made public in aggregated form under the *Open Data Commons Attribution License*.
- The results of this study will not be published before the end of the Spanish general elections on July 23, 2023.

You can participate in the survey even if you have answered it before.

Are you willing to take part in the following survey?

- ☐ Yes, I'm above the age of 18, and I'm giving you my consent.
- ☐ No, I do not wish to take part.

Consider only your contacts of voting age in Spain's general election on 23 July 2023.

- **Q2.** How many of your contacts will vote for PP (Partido Popular) for the Spanish general election?
- Q3.** How many of your contacts will vote for PSOE (Partido Socialista Obrero Español) for the Spanish general election?
- Q4.** How many of your contacts will vote for Vox for the Spanish general election?
- Q5.** How many of your contacts will vote for Sumar for the Spanish general election?
- Q6.** How many of your contacts will vote for other parties or blank for the Spanish general election?
- Q7.** How many of your contacts will not vote or you do not know their vote for the Spanish general election?
- Q8.** How many of your contacts considered above are between 18 and 34 years old?
- Q9.** How many of your contacts considered above are between 35 and 54 years old?

- Q10. How many of your contacts considered above are over 54 years old?
- Q11. How many of your contacts considered above are medical doctors?
- Q12. How many of your contacts considered above are autonomous workers?
- Q13. How many of your contacts considered above are unemployed?

A.1.6 Original Prompt for Madrid considering Age

Role and Context

You are a person between 18 and 24 years old from Madrid. You have read and accepted to participate in a survey conducted by a team of statistics and computer science researchers who are collecting data on the voting intention for the Spanish general elections of July 23, 2023. The group of contacts may include family, friends, or work colleagues whose voting intention you know.

- **Q1. Scientific Study Information** We are a team of statistics and computer science researchers. We collect data to estimate the vote intention in the 2023 Spanish general elections. The following survey contains questions regarding the voting intention of the people that you know for the 2023 Spanish general election. These contacts can include family, friends, or colleagues whose voting intention you know.
 - Responses are anonymous. The researchers cannot identify the participants.
 - Data cannot be modified or deleted once submitted.
 - The data may be published in aggregated form under the *Open Data Commons Attribution License*.
 - The results will not be published before the 2023 Spanish general elections have concluded.

Please respond to the following questions considering that you have given your consent to participate in the survey, keeping in mind that it is July 2023.

- **Q2.** How many of your contacts will vote for *Partido Popular* (PP) in the Spanish general election?
- **Q3.** How many of your contacts will vote for *Partido Socialista Obrero Español* (PSOE) in the Spanish general election?
- **Q4.** How many of your contacts will vote for Vox in the Spanish general election?
- **Q5.** How many of your contacts will vote for Sumar in the Spanish general election?

- **Q6.** How many of your contacts will vote for other parties or cast a blank vote in the Spanish general election?
- **Q7.** How many of your contacts will not vote or you do not know their vote in the Spanish general election?
- **Q8.** How many of your contacts considered previously are between 18 and 34 years old?
- **Q9.** How many of your contacts considered previously are between 35 and 54 years old?
- **Q10.** How many of your contacts considered previously are over 54 years old?
- **Q11.** How many of your contacts considered previously are self-employed workers?
- **Q12.** How many of your contacts considered previously are medical doctors?
- **Q13.** How many of your contacts considered previously are unemployed?

A.1.7 Perfect Prompt for Madrid considering Age and income

It is July 2023. As a participant in a survey on voting intention for the Spanish general elections on July 23, 2023, provide numerical responses in a single column taking into account the given context. Assume you are a person living in Madrid, between (age range) years old, with an annual income of (income range) euros. Provide numerical answers for the following questions regarding your contacts:

- **Q1.** Number of your contacts who will vote for *Partido Popular* (PP).
- **Q2.** Number of your contacts who will vote for *Partido Socialista Obrero Español* (PSOE).
- **Q3.** Number of your contacts who will vote for Vox.
- **Q4.** Number of your contacts who will vote for Sumar.
- **Q5.** Number of your contacts who will vote for other parties or cast a blank vote.
- **Q6.** Number of your contacts who will not vote or whose vote is unknown.
- **Q7.** Number of your contacts with ages between 18 and 34 years.
- **Q8.** Number of your contacts with ages between 35 and 54 years.
- **Q9.** Number of your contacts older than 54 years.

- **Q10.** Number of your contacts who are self-employed workers.
- **Q11.** Number of your contacts who are medical doctors.
- **Q12.** Number of your contacts who are unemployed.

Response Format:

[Number of contacts voting for PP]
 [Number of contacts voting for PSOE]
 [Number of contacts voting for Vox]
 [Number of contacts voting for Sumar]
 [Number of contacts voting for others/blank]
 [Number of contacts not voting/unknown]
 [Number of contacts aged 18-34]
 [Number of contacts aged 35-54]
 [Number of contacts aged >54]
 [Number of contacts self-employed]
 [Number of contacts doctors]
 [Number of contacts unemployed]

The information provided will be used anonymously and will only be published in aggregate after the elections. We thank you for your participation in this study.

A.1.8 Correlation Matrices

The following tables show the correlations between the other two regions commented in section 5.2.3.

Table SM12: Correlation matrix: Andalusia, Human Surveys

	PP	PSOE	Vox	Sumar
PP	1.0000	0.1238	0.1366	0.0369
PSOE	0.1238	1.0000	-0.0616	0.7994
Vox	0.1366	-0.0616	1.0000	-0.0609
Sumar	0.0369	0.7994	-0.0609	1.0000

Table SM13: Correlation matrices: Andalusia (GPT 5 - P. Perfect - Age and income)

(a) Chat Interface					(b) API Interface				
	PP	PSOE	VOX	SUMAR		PP	PSOE	VOX	SUMAR
PP	1.0000	-0.0664	0.8330	-0.2575	PP	1.0000	0.6812	0.8821	0.2270
PSOE	-0.0664	1.0000	-0.0124	0.0935	PSOE	0.6812	1.0000	0.6753	0.5064
VOX	0.8330	-0.0124	1.0000	-0.1365	VOX	0.8821	0.6753	1.0000	0.4515
SUMAR	-0.2575	0.0935	-0.1365	1.0000	SUMAR	0.2270	0.5064	0.4515	1.0000

Table SM14: Correlation matrix: Valencia, Human Surveys

	PP	PSOE	Vox	Sumar
PP	1.0000	0.2647	0.4569	-0.0215
PSOE	0.2647	1.0000	0.1564	0.6778
Vox	0.4569	0.1564	1.0000	0.0013
Sumar	-0.0215	0.6778	0.0013	1.0000

Table SM15: Correlation matrices: Valencia (GPT 5 - P. Perfect - Age and income)

(a) Chat Interface					(b) API Interface				
	PP	PSOE	VOX	SUMAR		PP	PSOE	VOX	SUMAR
PP	1.0000	0.0405	0.8486	-0.3034	PP	1.0000	0.6548	0.8690	0.0997
PSOE	0.0405	1.0000	0.1446	0.4660	PSOE	0.6548	1.0000	0.6140	0.4573
VOX	0.8486	0.1446	1.0000	-0.1037	VOX	0.8690	0.6140	1.0000	0.2692
SUMAR	-0.3034	0.4660	-0.1037	1.0000	SUMAR	0.0997	0.4573	0.2692	1.0000

A.2 Chilean Elections

Analogous to the Spanish study, the tables below display the vote estimates for every iteration using the four NSUM estimators. The results are categorized by round and temporal stage. Furthermore, we have included both the Original and Perfect prompts to illustrate their respective impacts on the LLM-generated data.

A.2.1 Results - First Round

Survey	FP	JJ	MEO	JK	JAK	EA	EM	HM	BLANK
Real Results	19.7	26.9	1.2	13.9	23.9	0.7	12.5	1.3	1.1
Human Surveys	7.1	23.3	2.6	13.5	26.8	1.9	13.3	2.9	8.6
Date: 08/11									
API · P.original · GPT 4 · Naive	15.2	12.8	13.8	2.6	21.8	1.9	16.2	6.7	9.0
API · P.original · GPT 4 · MOS	15.2	12.7	14.1	2.7	21.6	1.8	16.2	6.7	8.8
API · P.original · GPT 4 · MLE	14.9	12.5	13.9	2.9	21.7	2.0	16.2	6.8	9.2
API · P.original · GPT 4 · PIMLE	15.1	12.7	14.1	2.8	21.7	1.8	16.2	6.7	8.8
API · P.perfect · GPT 4 · Naive	14.3	13.9	15.4	2.8	17.4	1.5	15.2	6.7	12.7
API · P.perfect · GPT 4 · MOS	14.3	13.9	15.5	2.9	17.4	1.5	15.3	6.7	12.6
API · P.perfect · GPT 4 · MLE	13.9	14.1	15.4	3.1	17.2	1.6	15.3	6.9	12.6
API · P.perfect · GPT 4 · PIMLE	14.3	13.9	15.5	2.9	17.4	1.5	15.3	6.7	12.6
API · P.original · GPT 5 · Naive	10.2	12.5	7.9	3.9	20.3	3.9	27.9	2.7	10.6
API · P.original · GPT 5 · MOS	10.2	12.6	8.0	3.9	20.3	3.9	28.0	2.7	10.5
API · P.original · GPT 5 · MLE	10.1	12.6	7.9	3.8	20.3	3.8	28.0	2.7	10.6
API · P.original · GPT 5 · PIMLE	10.2	12.6	8.0	3.9	20.3	3.9	28.0	2.7	10.5
API · P.perfect · GPT 5 · Naive	10.0	12.6	7.8	3.7	20.5	3.7	29.2	2.6	10.0
API · P.perfect · GPT 5 · MOS	10.0	12.7	7.8	3.7	20.4	3.7	29.2	2.7	9.9
API · P.perfect · GPT 5 · MLE	10.0	12.7	7.8	3.7	20.4	3.7	29.2	2.7	9.9
API · P.perfect · GPT 5 · PIMLE	10.0	12.7	7.8	3.7	20.4	3.7	29.2	2.7	9.9
API · P.original · GPT 4o mini · Naive	22.2	11.1	0.0	0.0	33.3	0.0	11.1	0.0	22.2
API · P.original · GPT 4o mini · MOS	22.2	11.1	0.0	0.0	33.3	0.0	11.1	0.0	22.2
API · P.original · GPT 4o mini · MLE	22.2	11.1	0.0	0.0	33.3	0.0	11.1	0.0	22.2
API · P.original · GPT 4o mini · PIMLE	22.2	11.1	0.0	0.0	33.3	0.0	11.1	0.0	22.2
API · P.perfect · GPT 4o mini · Naive	22.2	11.1	0.0	0.0	33.3	0.1	11.1	0.0	22.2
API · P.perfect · GPT 4o mini · MOS	22.2	11.1	0.0	0.0	33.3	0.1	11.1	0.0	22.2
API · P.perfect · GPT 4o mini · MLE	22.2	11.1	0.0	0.0	33.3	0.1	11.1	0.0	22.2
API · P.perfect · GPT 4o mini · PIMLE	22.2	11.1	0.0	0.0	33.3	0.1	11.1	0.0	22.2
CHAT · P.original · GPT 5 · Naive	20.4	11.7	7.5	0.9	26.9	0.1	19.6	6.0	6.9
CHAT · P.original · GPT 5 · MOS	20.3	11.8	7.5	0.9	26.8	0.1	19.8	6.0	6.9
CHAT · P.original · GPT 5 · MLE	20.1	11.6	7.7	0.9	26.6	0.1	20.2	6.0	6.9
CHAT · P.original · GPT 5 · PIMLE	20.2	11.8	7.5	0.9	26.8	0.1	19.8	6.0	6.9
CHAT · P.perfect · GPT 5 · Naive	14.4	14.1	6.0	0.6	23.6	2.6	24.9	3.8	10.0
CHAT · P.perfect · GPT 5 · MOS	13.9	14.4	5.9	0.7	23.6	2.6	25.1	3.8	9.9
CHAT · P.perfect · GPT 5 · MLE	13.7	13.9	5.8	0.7	23.0	2.9	25.9	4.2	9.9
CHAT · P.perfect · GPT 5 · PIMLE	14.0	14.4	6.0	0.7	23.7	2.5	24.9	3.8	9.9
Date: 21/11									
CHAT · P.original · GPT 5 - Post · Naive	13.7	10.7	6.2	0.9	22.7	0.3	29.8	3.8	12.0
CHAT · P.original · GPT 5 - Post · MOS	14.8	10.6	6.6	0.9	22.2	0.3	28.9	4.1	11.6
CHAT · P.original · GPT 5 - Post · MLE	15.7	10.2	6.9	1.2	21.6	0.4	28.0	4.5	11.4
CHAT · P.original · GPT 5 - Post · PIMLE	14.8	10.6	6.6	0.9	22.3	0.3	28.9	4.1	11.6
CHAT · P.perfect · GPT 5 - Post · Naive	16.5	17.6	8.1	1.7	24.3	2.2	18.2	1.9	9.5
CHAT · P.perfect · GPT 5 - Post · MOS	16.3	18.8	8.4	2.0	24.4	2.2	17.0	1.8	9.2
CHAT · P.perfect · GPT 5 - Post · MLE	16.2	18.0	8.2	2.0	23.6	2.5	18.0	2.1	9.5
CHAT · P.perfect · GPT 5 - Post · PIMLE	16.3	18.6	8.3	1.9	24.3	2.2	17.2	1.8	9.3

Table SM16: Vote Intention Percentage Results Chile (First Round). The first iterations were conducted before the first round (16/11) whereas the second set was generated after the primary election. (FP: Parisi, JJ: Jara, MEO: Ominami, JK: Kaiser, JAK: Kast, EA: Artés, EM: Matthei, HM: Mayne-Nicholls)

A.2.2 Results - Second Round

Survey	JAK	JJ	BLANK
Real Results	57.39	41.29	1.33
Human Surveys	48.00	38.00	12.00
Date: 24/11			
API · P.original · GPT 4 · Naive	38.60	43.66	17.74
API · P.original · GPT 4 · MOS	38.52	43.64	17.85
API · P.original · GPT 4 · MLE	38.43	43.35	18.22
API · P.original · GPT 4 · PIMLE	38.56	43.66	17.78
API · P.perfect · GPT 4 · Naive	40.33	39.77	19.89
API · P.perfect · GPT 4 · MOS	40.89	39.47	19.63
API · P.perfect · GPT 4 · MLE	40.85	39.33	19.82
API · P.perfect · GPT 4 · PIMLE	41.00	39.42	19.58
API · P.original · GPT 5 · Naive	44.96	38.68	16.36
API · P.original · GPT 5 · MOS	45.01	38.65	16.34
API · P.original · GPT 5 · MLE	45.12	38.53	16.35
API · P.original · GPT 5 · PIMLE	45.00	38.65	16.35
API · P.perfect · GPT 5 · Naive	45.82	37.94	16.24
API · P.perfect · GPT 5 · MOS	45.86	37.99	16.15
API · P.perfect · GPT 5 · MLE	46.00	37.97	16.02
API · P.perfect · GPT 5 · PIMLE	45.86	38.00	16.15
API · P.original · GPT 4o mini · Naive	33.33	66.67	0.00
API · P.original · GPT 4o mini · MOS	33.33	66.67	0.00
API · P.original · GPT 4o mini · MLE	33.33	66.67	0.00
API · P.original · GPT 4o mini · PIMLE	33.33	66.67	0.00
API · P.perfect · GPT 4o mini · Naive	33.33	56.75	9.92
API · P.perfect · GPT 4o mini · MOS	33.33	54.16	12.51
API · P.perfect · GPT 4o mini · MLE	33.33	54.23	12.43
API · P.perfect · GPT 4o mini · PIMLE	33.33	54.15	12.52
CHAT · P.original · GPT 5 · Naive	47.24	34.70	18.05
CHAT · P.original · GPT 5 · MOS	47.50	34.62	17.87
CHAT · P.original · GPT 5 · MLE	47.54	34.61	17.85
CHAT · P.original · GPT 5 · PIMLE	47.61	34.61	17.78
CHAT · P.perfect · GPT 5 · Naive	45.72	40.61	13.67
CHAT · P.perfect · GPT 5 · MOS	46.69	39.88	13.43
CHAT · P.perfect · GPT 5 · MLE	46.44	40.00	13.56
CHAT · P.perfect · GPT 5 · PIMLE	46.77	39.81	13.43

Table SM17: Vote Intention Percentage Results Chile (Second Round, First Iteration). (JAK: Kast, JJ: Jara)

Survey	JAK	JJ	BLANK
Real Results	57.39	41.29	1.33
Date: 04/12			
Human Surveys	47.98	38.23	12.79
CHAT · P.original · GPT 5 · Naive	47.15	35.99	16.85
CHAT · P.original · GPT 5 · MOS	47.30	36.02	16.68
CHAT · P.original · GPT 5 · MLE	47.46	36.05	16.48
CHAT · P.original · GPT 5 · PIMLE	47.31	35.99	16.70
CHAT · P.perfect · GPT 5 · Naive	45.32	39.92	14.76
CHAT · P.perfect · GPT 5 · MOS	46.53	38.94	14.53
CHAT · P.perfect · GPT 5 · MLE	46.14	39.07	14.80
CHAT · P.perfect · GPT 5 · PIMLE	46.52	38.97	14.51
Date: 12/12			
Human Surveys	46.53	39.73	13.73
API · P.original · GPT 4 · Naive	38.80	42.42	18.77
API · P.original · GPT 4 · MOS	39.25	42.42	18.33
API · P.original · GPT 4 · MLE	39.11	42.13	18.77
API · P.original · GPT 4 · PIMLE	39.36	42.39	18.25
API · P.perfect · GPT 4 · Naive	37.16	41.75	21.09
API · P.perfect · GPT 4 · MOS	37.02	42.08	20.91
API · P.perfect · GPT 4 · MLE	37.15	41.72	21.13
API · P.perfect · GPT 4 · PIMLE	37.18	42.11	20.71
API · P.original · GPT 5 · Naive	45.24	38.53	16.23
API · P.original · GPT 5 · MOS	45.30	38.51	16.20
API · P.original · GPT 5 · MLE	45.40	38.46	16.14
API · P.original · GPT 5 · PIMLE	45.31	38.51	16.18
API · P.perfect · GPT 5 · Naive	45.37	37.60	17.03
API · P.perfect · GPT 5 · MOS	45.39	37.64	16.97
API · P.perfect · GPT 5 · MLE	45.40	37.64	16.96
API · P.perfect · GPT 5 · PIMLE	45.39	37.64	16.97
API · P.original · GPT 4o mini · Naive	33.33	66.67	0.00
API · P.original · GPT 4o mini · MOS	33.33	66.67	0.00
API · P.original · GPT 4o mini · MLE	33.33	66.67	0.00
API · P.original · GPT 4o mini · PIMLE	33.33	66.67	0.00
API · P.perfect · GPT 4o mini · Naive	33.33	58.25	8.42
API · P.perfect · GPT 4o mini · MOS	33.33	55.40	11.26
API · P.perfect · GPT 4o mini · MLE	33.33	55.48	11.18
API · P.perfect · GPT 4o mini · PIMLE	33.33	55.41	11.26
CHAT · P.original · GPT 5 · Naive	40.15	40.29	19.55
CHAT · P.original · GPT 5 · MOS	40.09	40.37	19.53
CHAT · P.original · GPT 5 · MLE	39.90	40.42	19.69
CHAT · P.original · GPT 5 · PIMLE	40.16	40.35	19.49
CHAT · P.perfect · GPT 5 · Naive	45.82	40.67	13.51
CHAT · P.perfect · GPT 5 · MOS	47.12	39.70	13.18
CHAT · P.perfect · GPT 5 · MLE	46.81	39.89	13.30
CHAT · P.perfect · GPT 5 · PIMLE	47.10	39.73	13.17

Table SM18: Vote Intention Percentage Results Chile (Second Round, Second and Third Iterations). (JAK: Kast, JJ18Jara)

A.2.3 Original Prompt - First Round

Role and Context

You are a person between (age range) years old from Chile. You have read and accepted to participate in a survey conducted by a team of statistics and computer science researchers who are collecting data on the voting intention for the 2025 Chilean presidential elections. The group of contacts may include family, friends, or work colleagues whose voting intention you know.

- **Q1. Scientific Study Information**

We are a team of statistics and computer science researchers. We collect data to estimate the vote intention in the 2025 Chilean presidential elections. The following survey contains questions regarding the voting intention of the people that you know for the 2025 Chilean presidential election. These contacts can include family, friends, or colleagues whose voting intention you know.

- Responses are anonymous. The researchers cannot identify the participants.
- Data cannot be modified or deleted once submitted.
- The data may be published in aggregated form under the *Open Data Commons Attribution License*.
- The results will not be published before the 2025 Chilean presidential elections have concluded.

Please respond to the following questions considering that you have given your consent to participate in the survey, keeping in mind that it is early November 2025.

- **Q2.** How many of your contacts will vote for Franco Parisi Fernández in the Chilean presidential election?
- **Q3.** How many of your contacts will vote for Jeannette Jara Román in the Chilean presidential election?
- **Q4.** How many of your contacts will vote for Marco Antonio Enríquez - Ominami Gumucio in the Chilean presidential election?
- **Q5.** How many of your contacts will vote for Johannes Kaisser Barents - Von Hohenhagen in the Chilean presidential election?
- **Q6.** How many of your contacts will vote for José Antonio Kast Rist in the Chilean presidential election?
- **Q7.** How many of your contacts will vote for Eduardo Antonio Artés Brichetti in the Chilean presidential election?
- **Q8.** How many of your contacts will vote for Evelyn Matthei Fornet in the Chilean presidential election?

- **Q9.** How many of your contacts will vote for Harold Mayne - Nicholls Secul in the Chilean presidential election?
- **Q10.** How many of your contacts will vote null or blank in the Chilean presidential election?
- **Q11.** How many of your contacts will not vote or you do not know their vote for the Chilean presidential election?
- **Q12.** How many of your contacts considered previously are between 18 and 34 years old?
- **Q13.** How many of your contacts considered previously are between 35 and 54 years old?
- **Q14.** How many of your contacts considered previously are over 54 years old?
- **Q15.** How many of your contacts considered previously are independent workers?
- **Q16.** How many of your contacts considered previously are unemployed?
- **Q17.** How many of your contacts considered previously are medical doctors?

A.2.4 Perfect Prompt - Second Round

It is November 2025. As a participant in a survey on voting intention for the Chilean presidential elections on December 14, 2025, provide numerical responses in a single column taking into account the given context. Assume you are a person living in Chile, between (age range) years old. Provide numerical answers for the following questions regarding your contacts:

- **Q1.** Number of your contacts who will vote for Jeannette Jara Román.
- **Q2.** Number of your contacts who will vote for José Antonio Kast Rist.
- **Q3.** Number of your contacts who will vote null or blank.
- **Q4.** Number of your contacts who will not vote or whose vote is unknown.
- **Q5.** Number of your contacts with ages between 18 and 34 years.
- **Q6.** Number of your contacts with ages between 35 and 54 years.
- **Q7.** Number of your contacts older than 54 years.
- **Q8.** Number of your contacts who are independent workers.
- **Q9.** Number of your contacts who are unemployed.

- **Q10.** Number of your contacts who are medical doctors.
- **Q11.** Number of your contacts who live abroad.

Response Format:

[Number of contacts voting for Jeannette Jara Román]
[Number of contacts voting for José Antonio Kast Rist]
[Number of contacts voting null/blank]
[Number of contacts not voting/unknown]
[Number of contacts aged 18-34]
[Number of contacts aged 35-54]
[Number of contacts aged >54]
[Number of contacts independent workers]
[Number of contacts unemployed]
[Number of contacts doctors]
[Number of contacts living abroad]

The information provided will be used anonymously and will only be published in aggregate after the elections. We thank you for your participation in this study.