Towards a framework for peer-supported conditioning in LLM survey response simulation

Anonymous ACL submission

Abstract

Can LLMs simulate how people respond to surveys more accurately when given testimony from people like them? We explore this question by retrieving interview excerpts from demographically similar individuals and incorporating them into the prompt, rather than relying solely on demographics. Testing this on Chilean political surveys, we find that none of our variants improves individual-level accuracy over the demographics-only baseline. Reasoning-based prompts do produce more varied responses, but without predicting the right answers better, more variance, and not more accuracy. We diagnose why: the approach requires the model to already have a baseline with attitudinal priors aligned with the target population, a qualitative corpus carrying stance evidence on the target items, and a prompting scheme that preserves inter-peer variance. We show none holds cleanly in this setting, clarifying directions for future work.

1 Introduction

During the last few years, survey response rates have declined steadily, raising operational costs and motivating interest in LLM-based survey simulation (Luiten et al., 2020; Acharán et al., 2024; Sanders et al., 2023). LLMs encode statistical regularities from their pretraining data, including biases correlated with demographic identities (Hu et al., 2025; Argyle et al., 2023); persona prompting exploits this to steer generation toward a target subgroup. Argyle et al. (Argyle et al., 2023) introduced the paradigm as *algorithmic fidelity*, simulating ANES voting behaviour with persona-conditioned GPT-3. Persona-only prompting has documented limitations: it pulls outputs toward identity-correlated pretraining text (Kang et al., 2025), compresses within-group variance (Li et al., 2025), and underperforms outside English-speaking settings. On Chile's CEP survey, previous work on survey response simulation report poor predictive and distributional alignment, with heterogeneity across questions and subpopulations (González-Bustamante et al., 2025; Abeliuk et al., 2025).

Recent work address these limitations by conditioning simulation on interview transcripts. Park et al. (Park et al., 2024) show that two-hour interview transcripts of the target individual outperform persona-only prompting across the GSS, Big Five, and behavioural-economics replications, with gains preserved under transcript truncation. Zhang et al. (Zhang et al., 2025) report similar gains for the BREQ instrument across GPT-4.1, Gemini 2.0 Flash, and Claude 3.7 Sonnet. Both setups require interview material from the target persona, rarely available in practice. An earlier marketing study by Arora et al. (Arora et al., 2025) used retrieval-augmented generation to condition synthetic survey responses on qualitative interview material from comparable consumers, finding that added context improved response heterogeneity and reliability.

This paper presents a framework that uses qualitative interview data as context for LLM survey simulation via retrieval-augmented generation (RAG). Instead of using an individual's *own* interview to simulate that same individual, we retrieve testimony from *demographically similar others* (peers) to model target respondents who have never been interviewed, thereby informing simulation across a survey population without one-to-one interviewee–respondent correspondence. Figure 1 illustrates the retrieval process for an example respondent.

We test the framework on a Chilean political survey, where prior work on survey simulation has shown lower predictive power than in US baselines (Abeliuk et al., 2025; González-Bustamante et al., 2025), using qualitative data from the VOCES study[1]. Our contribution is diagnostic. Three con-

[1] Still ongoing; see Appendix A for study details.

ditions appear necessary for peer-supported simulation to improve over persona-only prompting: priors aligned with the target population, a corpus carrying stance evidence on the target items, and a prompting scheme that preserves inter-peer variance. We show none holds cleanly, clarifying directions for future work.

2 Data

We combine two data sources. The first is the CEP survey, a closed-ended public-opinion survey conducted in Chile since 1987; we use waves 93 (March–April 2025) (Centro de Estudios Públicos, 2025a) and 94 (May–June 2025) (Centro de Estudios Públicos, 2025b), restricted to La Araucanía and yielding 65 and 75 respondents respectively. The second is the VOCES study, 82 semi-structured interviews conducted in 2025 in Padre Las Casas (49) and Purén (33) focused on the Chilean state–Mapuche conflict. VOCES interviews were conducted in the same region and within the same time window as the CEP waves, thus making them plausible peers of CEP respondents.

Six sociodemographic attributes (age, sex, political self-placement, religious affiliation, indigenous identification, and educational attainment) were retained based on broad availability and prior use in similar work. Attributes were recoded to reduce category sparsity: political self-placement was collapsed from a 1–10 scale into five categories (left, centre-left, centre, centre-right, right); educational attainment into six levels (none, primary, secondary, technical, undergraduate, postgraduate); and indigenous identification kept negative responses as a category distinct from missing.

Eight CEP items were selected as targets, all opinion-based with unordered nominal responses: stance on *abortion*, preference for *democracy*, *evaluation* of President Boric, *perception* of Chile's current state, expectations for the national *economy* over the next twelve months, the impact of digital technology on *jobs*, and comparisons of current housing conditions to those of the respondent's parents (*housing-past*) and to their children's future (*housing-future*). A "don't know" option was retained throughout. Three items appear in both waves; the rest are wave-specific. Usable sample sizes range from 62 to 139. Full question wordings and response sets are in the Appendix.

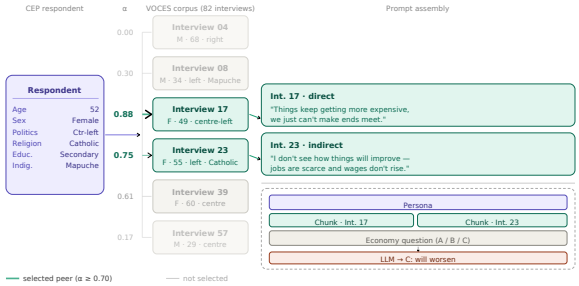

Figure 1: Peer-supported conditioning: matching a CEP respondent to VOCES peers.

3 Methodology

3.1 Simulation setup

Each CEP respondent is described by a *persona* built from their available sociodemographic attributes. Simulations are generated at the level of unique attribute configurations and mapped back to the full CEP sample for evaluation, with 4 repetitions per (variant, question, persona) tuple to account for provider nondeterminism. We compare a baseline with only sociodemographic attributes (Persona, **P**) to three variants: chain-of-thought (CoT, **C**), retrieval-augmented (RAG, **R**), and their combination (CoT+RAG, **CR**). All simulations are produced by Qwen-3 32B (Team, 2025), served through OpenRouter on DeepInfra's fp8 endpoint with a temperature of 0.2 and top-k of 1, with its native thinking mode disabled.

3.2 Peer matching

For each CEP respondent c and interview participant t, an attribute-based similarity score combines attribute-level similarities $s_{t|c,k}$ over the six sociodemographic attributes in $\mathcal{A}$. Let $E(c) \subseteq \mathcal{A}$ denote the attributes available for c and $C(c,t) \subseteq E(c)$ those also available for t. The score is

$$\alpha_{t|c} = \frac{\sum_{k \in C(c,t)} w_k \, s_{t|c,k}}{\sum_{k \in C(c,t)} w_k} \times \left(\frac{\sum_{k \in C(c,t)} w_k}{\sum_{k \in E(c)} w_k} \right)^{\gamma} \quad (1)$$

where the factor at the left penalises for missing attributes on the interviewee side. Political identity is treated ordinally on a 1–5 scale ($s = 1$ for exact match, 0.5 for adjacent, 0 otherwise); age is matched by exact membership in one of four Chilean political cohorts informed by Maureira (2008); Bargsted et al. (2019) (Appendix B); all other attributes use exact matching. Weights are $w_{\text{pol}} = 3$, $w_{\text{age}} = 2$, $w_k = 1$ elsewhere, and $\gamma = 2/3$.

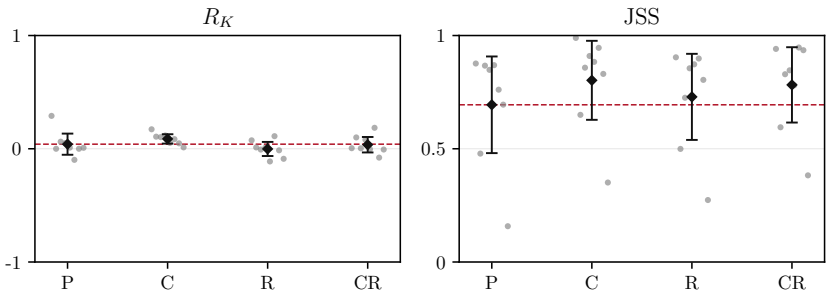

Figure 2: Average R_K and Jensen–Shannon similarity (JSS) across the eight CEP items, by prompt variant: persona-only baseline (**P**), CoT (**C**), RAG (**R**), and CoT+RAG (**CR**). Grey points show the item-level values, black diamonds show the equal-weighted mean across items, and vertical bars show 95% small-sample t confidence intervals over the eight item-level values.

3.3 Hybrid retrieval

Interviews are split into overlapping four-turn windows with a stride of two. Each chunk is encoded with BGE-M3 (Chen et al., 2024), which produces three complementary representations from a single forward pass: a dense vector for semantic similarity, sparse term weights for lexical overlap, and ColBERT-style multi-vector embeddings for fine-grained token-level alignment. Per-modality chunk–question similarities are normalised and combined as a weighted sum with weights 0.4 (dense), 0.2 (sparse), and 0.4 (ColBERT).

To improve fragment relevance, we introduce an auxiliary LLM judge. For each survey item, hybrid retrieval surfaces top chunks from the highest-ranked peer interviews, which are then submitted to the judge. The judge labels each chunk with a usefulness score in $\{0, 1, 2, 3\}$ and a label (0 *irrelevant*, 1 *on-topic* but uninformative, 2 useful *indirect* or *background* evidence, 3 *direct* stance-bearing evidence), based on the chunk's relevance to the target question and the stance-bearing content it carries. Only chunks with usefulness ≥ 2 are eligible. Appendix Figure 5 reports the average fragment composition of prompts, broken down by judge-assigned label.

For each (CEP respondent, question) pair, peer interviews are ranked by $\alpha_{t|c}$. The peer window starts at the top 4 (with ties at the cut-off) and expands stepwise to 6, 8, and 10 whenever fewer than four useful chunks are available. Within the final pool, up to four highest-scoring useful chunks across all eligible interviews enter the prompt; multiple chunks from the same interview are permitted, since the strongest stance-bearing chunk from a peer is often not its top-ranked one. If no useful chunks are found within the widest window, the prompt is assembled sparsely.

3.4 Evaluation

We report two complementary measures. The K-category correlation coefficient R_K (Gorodkin, 2004), given by

$$R_K = \frac{cs - \sum_k p_k t_k}{\sqrt{(s^2 - \sum_k p_k^2)(s^2 - \sum_k t_k^2)}} \quad (2)$$

where, for the $K \times K$ confusion matrix C with C_{ij} counting predictions of class j for true class i, $t_k = \sum_i C_{ik}$ and $p_k = \sum_i C_{ki}$ are the true and predicted instance counts for class k, $c = \sum_k C_{kk}$ counts correct predictions, and $s = \sum_{ij} C_{ij}$ is the total. $R_K \in [-1, 1]$ with $R_K = 1$ at perfect classification.

Distributional fidelity is reported as the complement of the Jensen–Shannon divergence:

$$\mathrm{JSS}(P\|Q) = 1 - \tfrac{1}{2}\left[D_{\mathrm{KL}}(P\|M) + D_{\mathrm{KL}}(Q\|M)\right], \quad (3)$$

with $M = \frac{1}{2}(P + Q)$ and JSS $\in [0, 1]$. JSS values near 1 can arise even from trivial predictors that match marginals without recovering individual rows.

Confidence intervals on ΔR_K and Δ JSS relative to **P** are obtained by stratified bootstrap with $B = 10{,}000$ replicates, resampling rows with replacement within each wave and pooling across waves, with two-sided percentile intervals at $\alpha = 0.01$.

We refer to the answer distribution under **P** as the model's *prior* for that (persona, item) pair: the ingrained concept associations it brings to the task before any added context.

4 Results

Figure 2 reports R_K and JSS averaged across questions for each prompt variant. On average, the baseline performs near chance ($R_K \approx 0$), and no variant produces substantive gains in R_K. Distributional similarity (JSS) increases most under the reasoning variants **C** and **CR**.

Per-question results with bootstrap CIs are reported in Figure 3 in the Appendix. With-context variants failed to improve individual-level performance on seven of the eight questions and led to a statistically significant decrease in performance on the abortion question. The *digital jobs* is the only item for which R_K improves under **C** and **CR**, with the latter showing a greater improvement over the baseline, which sits below chance.

Figure 4 shows the answer distributions obtained for each variant, alongside the empirical distribution. We observe strict modal collapse ($\geq 99\%$ on a single alternative) in three cases: **P** on *housing-past*, and both **P** and **R** on *democracy*; in all three, the simulated mode does not align with the empirical mode. A near-collapse pattern arises on the *housing future* question: the model assigns 96.5% probability to category *C*, while its empirical mass is only 3.1%; the empirical mode *A* (84.4%) receives no predictions. Low variance in LLM survey simulations is a well-studied phenomenon, and we find that requesting reasoning aids in counteracting modal collapse. We interpret increases in JSS associated with C and CR variants associated with no improvement in R_K as pure variance gains.

5 Discussion

The framework relies on two conditions holding jointly for peer-supported context to improve over persona-only prompting: the baseline must encode attitudinal regularities aligned with the target population, and the model must be responsive to added context in a way that moves predictions toward the empirical mode rather than merely adding variance. Our results suggest that neither condition is cleanly satisfied in this setting.

Model challenges At baseline, the model exhibits strong priors that frequently deviate from the empirical distribution, and prompt variants barely shift them. The *democracy* item is the clearest illustration: the prior heavily favours the pro-democracy alternative, whereas the empirical distribution is less concentrated, with the neutral stance emerging as the dominant choice. Several mechanisms plausibly contribute: priors anchored on the values of English-speaking, high-resource pretraining corpora; alignment training that suppresses politically charged stances; and a more general attenuation of algorithmic fidelity as base models are post-trained into assistants.

Sample challenges Prior interview-based simulation work (Park et al., 2024; Arora et al., 2025; Zhang et al., 2025) used interviews designed alongside the survey instrument that map onto the response sets. While VOCES is political in nature, direct stance-bearing evidence towards CEP items is sparse. Political survey simulation as a class is plausibly harder than the attitudinal tasks the cited literature has addressed: the items concern topics under censorship pressure for mainstream models, and the alternatives are value-laden rather than descriptive.

As a diagnostic of whether the model can recover explicitly stated stances from the interviews, we map LLM-judged direct-evidence fragments as ground truths and run a Park et al.-style simulation conditioned on the full interview and a set of expert reflections, four times per case. We find that synthetic responses agreed with the extracted ground truths 47.5% of the time, or 63.6%, counting 2-2 ties, raising concerns about the applicability of one-to-one interview-based silicon sampling in this specific domain.

Reasoning as context aggregation We find that our multi-fragment prompting setup leads the model to act as an aggregator across narratives, often weighing them toward middle-ground positions or defaulting to the majority. This behaviour effectively reproduces the low-variance phenomenon observed in LLM survey simulations, especially in settings with scarce peer evidence.

From the diagnosis above, at least three requirements must be met for a peer-supported LLM survey simulation to be feasible. First, the underlying model must be able to produce the full range of possible answers to politically sensitive items without censorship or refusal; otherwise, no amount of contextual anchoring will recover individual variation. Second, the qualitative corpus must carry a stance-bearing signal on the items being simulated. Third, the prompting scheme must preserve inter-peer differences rather than averaging across them, or the framework will reproduce the low-variance behaviour it was designed to overcome.

6 Limitations

Scope of testing We evaluated a single model, Qwen-3 32B, on a single regional subsample (La Araucanía, CEP waves 93 and 94) of a single national survey. Other models were tested during development (Gemma 3 27B, Mistral Small 3.2) but not included here; prior work reports substantial performance heterogeneity across model–question pairings (González-Bustamante et al., 2025), but we found comparable behaviour between the models tested. LLaMA 3.1 models were also tested but were dropped due to a 100% refusal rate. Larger models may yield higher baselines, although it is unclear whether their response to prompt variants would differ in kind rather than degree.

Disentanglability of misalignment sources Our setup cannot separate three plausible contributors to the baseline misalignment with the CEP sample: (i) the model's concept social associations not extending to contexts outside the training-corpus; (ii) the political nature of the survey, which involves topics more sensitive than the personality or lifestyle domains covered in prior interview-based work; and (iii) attenuation of algorithmic fidelity through post-training that pushes models toward neutral, assistant-like outputs. These likely act jointly, with weights that a single-model study cannot identify.

Corpus design The VOCES interviews were not designed to be compatible with the CEP survey. They are political in nature but do not cover the CEP target items uniformly, and direct stance evidence is unevenly distributed. A corpus designed in tandem with the target instrument would offer a cleaner test of peer-to-population transfer and would help isolate framework-level findings from sample-specific ones.

Risks Synthetic survey responses risk being mistaken for, or substituted into, real opinion data, particularly in low-cost-of-fieldwork settings where the appeal of simulation is highest. Reported aggregate similarity (JSS) can hide poor individual-level performance, as our results illustrate. We caution against using peer-supported simulation as a substitute for survey data without independent validation against held-out human responses.

References

Andrés Abeliuk, Vanessa Gaete, and Naim Bro. 2025. Fairness in LLM-Generated Surveys. *arXiv preprint arXiv:2501.15351*.

Joao Acharán, Paulina Valenzuela, Jorge Fábrega, and Maddalena Fiore. 2024. Clima de opinión hacia las encuestas en Chile: factores determinantes de la desconfianza. *Revista mexicana de opinión pública*, (37):115–133.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis*, 31(3):337–351.

Neeraj Arora, Ishita Chakraborty, and Yohei Nishimura. 2025. AI–Human Hybrids for Marketing Research: Leveraging Large Language Models (LLMs) as Collaborators. *Journal of Marketing*, 89(2):43–70.

Matías Bargsted, Nicolás M. Somma, and Benjamín Muñoz-Rojas. 2019. Electoral Turnout in Chile: An Age-Period-Cohort Approach. *Revista de ciencia política (Santiago)*, 39(1):75–98.

Centro de Estudios Públicos. 2025a. Estudio nacional de opinión pública n° 93, marzo–abril 2025. Data set.

Centro de Estudios Públicos. 2025b. Estudio nacional de opinión pública n° 94, mayo–junio 2025. Data set.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. _eprint: 2402.03216.

Bastián González-Bustamante, Nando Verelst, and Carla Cisternas. 2025. Emulating Public Opinion: A Proof-of-Concept of AI-Generated Synthetic Survey Responses for the Chilean Case. ArXiv:2509.09871 [cs].

J. Gorodkin. 2004. Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry*, 28(5):367–374.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2025. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75.

Minwoo Kang, Suhong Moon, Seung Hyeong Lee, Ayush Raj, Joseph Suh, David M. Chan, and John Canny. 2025. Deep Binding of Language Model Virtual Personas: a Study on Approximating Political Partisan Misperceptions. *arXiv preprint*. ArXiv:2504.11673 [cs] version: 5.

Dai Li, Linzhuo Li, and Huilian Sophie Qiu. 2025. ChatGPT is not A Man but Das Man: Representativeness and Structural Consistency of Silicon Samples Generated by Large Language Models. _eprint: 2507.02919.

Annemieke Luiten, Joop Hox, and Edith de Leeuw. 2020. Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys. *Journal of Official Statistics*, 36(3):469–487.

Sergio Toro Maureira. 2008. De lo épico a lo cotidiano: Jóvenes y generaciones políticas en Chile. *Revista de ciencia política (Santiago)*, 28(2):143–160.

Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative Agent Simulations of 1,000 People. *arXiv preprint*. ArXiv:2411.10109 [cs].

Nathan E. Sanders, Alex Ulinich, and Bruce Schneier. 2023. Demonstrations of the Potential of AI-based Political Issue Polling. *Harvard Data Science Review*, 5(4).

Qwen Team. 2025. Qwen3 Technical Report. _eprint: 2505.09388.

Jihong Zhang, Xinya Liang, Anqi Deng, Nicole Bonge, Lin Tan, Ling Zhang, and Nicole Zarrett. 2025. Leveraging Interview-Informed LLMs to Model Survey Responses: Comparative Insights from AI-Generated and Human Data. *arXiv preprint*. ArXiv:2505.21997 [cs].

A The VOCES study

VOCES (*Visiones, Opiniones y Conversaciones en el Espacio Social*) is an ongoing study funded by the Instituto Milenio para la Investigación en Violencia y Democracia, the Centro de Estudios Interculturales e Indígenas, and the Instituto Milenio Fundamentos de los Datos. The study aims to characterise the principal dimensions of the conflict between the Chilean state and the Mapuche through semi-structured interviews of up to one hour.

The interview protocol is organised into thematic blocks. It begins with a general characterisation of the participant (including basic sociodemographic information), then elicits political self-placement and values, and later moves to perceptions of local problems, the region and the conflict, actors and institutions, and possible paths forward. The script encourages interviewers to request elaboration on selected portions of the interview and to ask for clarification when an answer is deemed insufficient.

The transcripts used in this paper were provided by the VOCES team and correspond to interviews conducted during the second half of 2025 in two designated sectors of the La Araucanía region: 49 from Padre Las Casas and 33 from Purén ($n = 82$). The VOCES team obtained informed consent from all participants. VOCES contributed only the transcripts; the present analysis is independent of the VOCES study.

B Political cohort definitions

Age is operationalised by dividing respondents into four Chilean political generational cohorts, adapted from Maureira (2008); Bargsted et al. (2019) and extended to incorporate respondents who reached voting age during or after the 2019 *estallido social*:

- **Pre-plebiscite**: born in or before 1970.
- **Post-plebiscite transition**: born between 1971 and 1978 inclusive.
- **Pre-*estallido social***: born between 1979 and 2000.
- **Post-*estallido***: born in or after 2001.

Age similarity between a CEP respondent c and an interview participant t is defined as exact membership in the same cohort: $s_{t|c,\text{age}} = \mathbf{1}\big(g(c) = g(t)\big)$, where $g(k)$ maps age onto its cohort.

C A note on performance measure choices

We forgo standard classification metrics such as accuracy, F1, and macro-averaged recall. Because the empirical class distributions are markedly unbalanced, these metrics can reward a predictor whose modal preference happens to align with the majority class. These metrics also fail to account for the shape of predictions by collapsing the problem into a false dichotomy. We select the K-category correlation coefficient R_K (Gorodkin, 2004) because it accounts for the full confusion matrix and is robust to class imbalance. We note that, for the categorical case, a negative R_K cannot be clearly interpreted as negative correlation.

As a sanity check on our prompt setup, we compared our persona-only accuracy on the abortion question against the closest prior result on a Chilean survey simulation: 0.55 for our setup, compared with 0.52 for ChatGPT-4 and LLaMA-13B (Abeliuk et al., 2025). We find our baseline is in range.

Item	Question	Alternatives	W93	W94	Both
Abortion `religion_14`	Which of these ideas best expresses your view regarding abortion?	`A`: Abortion should always be prohibited; `B`: Abortion should be allowed only in some cases; `C`: Abortion should be a free choice for women	65	0	65
Democracy `democracia_21`	Which of the following statements do you most agree with?	`A`: Democracy is always preferable; `B`: Sometimes an authoritarian regime may be preferable; `C`: It makes no difference whether the regime is democratic or authoritarian	62	0	62
Country state `percepcion_4`	Do you think that, at the present time, Chile is progressing, stagnant, or in decline?	`A`: Progressing; `B`: Stagnant; `C`: In decline	62	75	137
Government `eval_gob_1`	Regardless of your political position, do you approve or disapprove of the way President Gabriel Boric is leading the government?	`A`: Approve; `B`: Disapprove; `C`: Neither approve nor disapprove	63	73	136
Economy `percepcion_3`	Do you think that, over the next 12 months, the country's economic situation will improve, stay the same, or worsen?	`A`: Will improve; `B`: Will stay the same; `C`: Will worsen	65	74	139
Digital jobs `trabajo_70`	Thinking about the impact that the use of new digital technologies will have on Chileans' work, would you say that more jobs will be created than lost, that a similar number of jobs will be created and lost, or that more jobs will be lost than created?	`A`: More jobs will be created than lost; `B`: A similar number of jobs will be created and lost; `C`: More jobs will be lost than created	0	75	75
Housing past `pobreza_75_f`	Compared with your parents when they were your current age, would you say your housing conditions are better, the same, or worse than theirs?	`A`: Better; `B`: The same; `C`: Worse	65	0	65
Housing future `pobreza_76_f`	Thinking about the future and the situation your children will have at your age, or would have if you had them, do you think your children's housing conditions will be better, the same, or worse than yours?	`A`: Better; `B`: The same; `C`: Worse	64	0	64

Table 1: CEP items used in this paper. Support columns report the number of valid responses in wave 93, wave 94, and both waves combined. The `D` alternative ("don't know") is omitted from the alternatives column since it has identical wording across items.

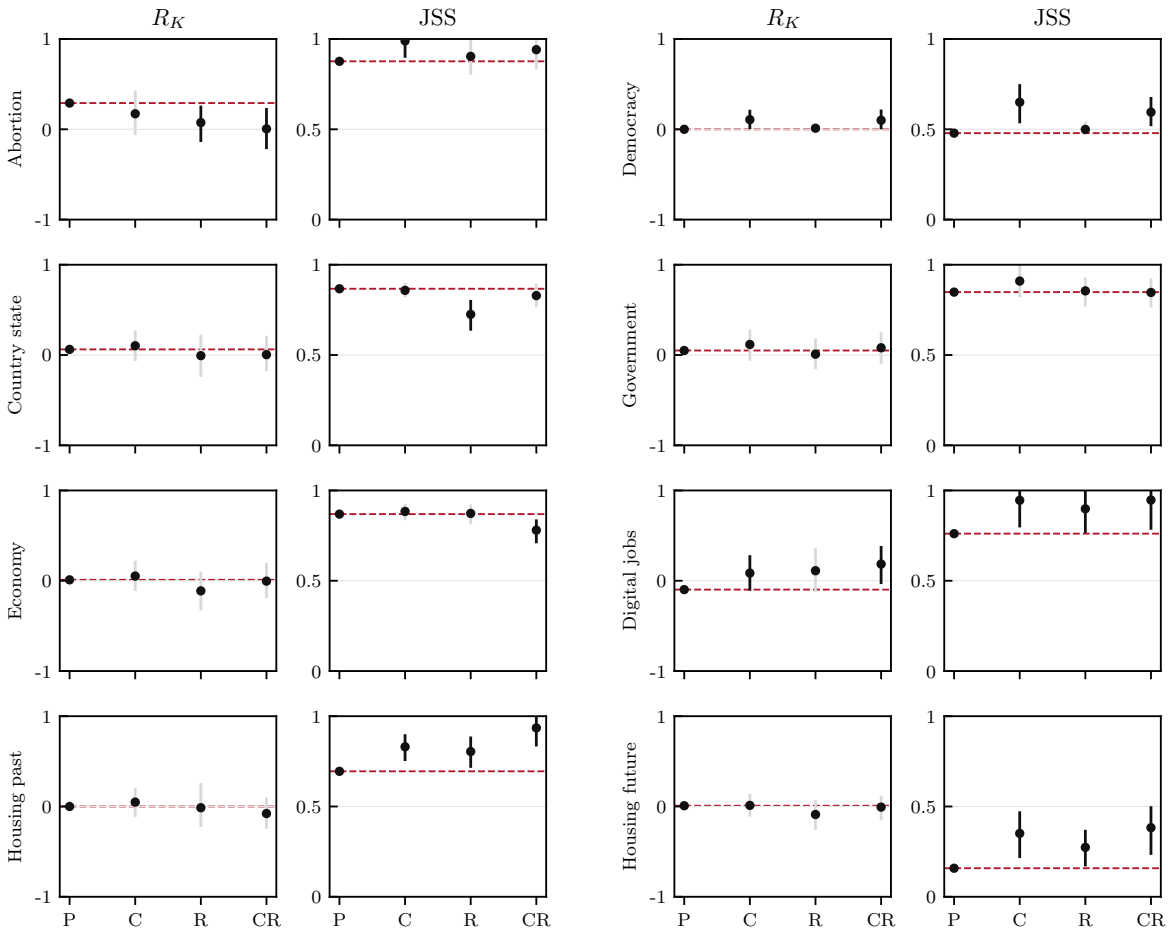

Figure 3: Per-question R_K and JSS for each prompt variant. Each CEP item appears in paired panels: individual-level agreement (R_K, left) and distributional similarity (JSS, right). Points show observed values for persona-only (P), CoT (C), RAG (R), and CoT+RAG (CR). Dashed horizontal lines mark the persona-only baseline. Bars for the other variants give bootstrap 99% CIs for the difference from baseline. Higher values indicate better agreement with the empirical CEP responses.

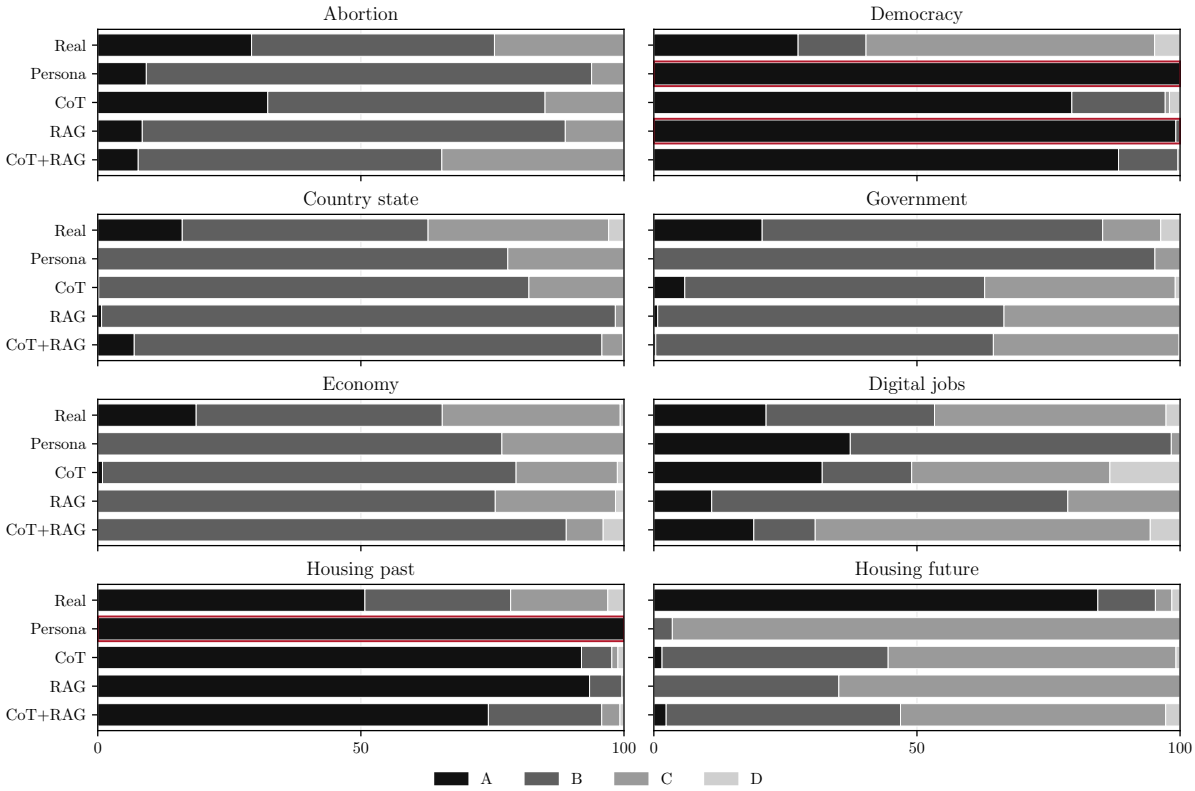

Figure 4: Empirical and simulated response distributions for each CEP item, with Qwen-3 32B. Within each item, the top row reports the observed CEP distribution and subsequent rows report simulated responses under Persona (P), CoT (C), RAG (R), and CoT+RAG (CR). Horizontal bars show the share assigned to each answer alternative (A–D, with D the "don't know" option). Variants with modal collapse ($\geq 99\%$ share) are highlighted in red.

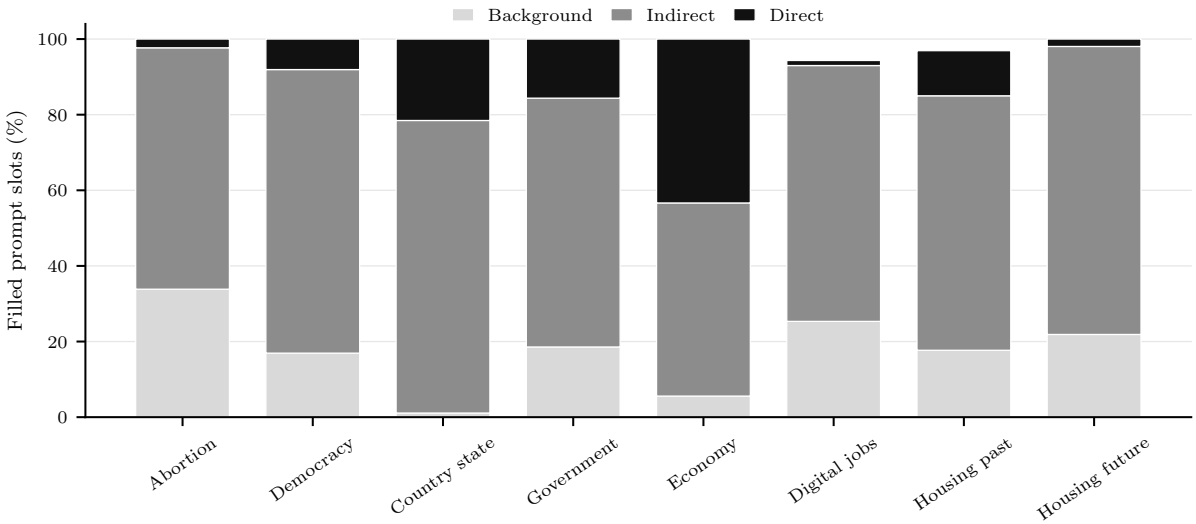

Figure 5: Stacked bars show the composition of interview fragments included in each prompt. Bar height is the average share of available prompt slots filled by retrieved fragments; segments show the judge-assigned label distribution among included fragments: background, indirect, and direct evidence.