Let Them Speak: Finding Who Policy Leaves Behind via Blind-Spot-Targeted LLM Simulation

Jaywoong Jeong
KAIST
Daejeon, Republic of Korea

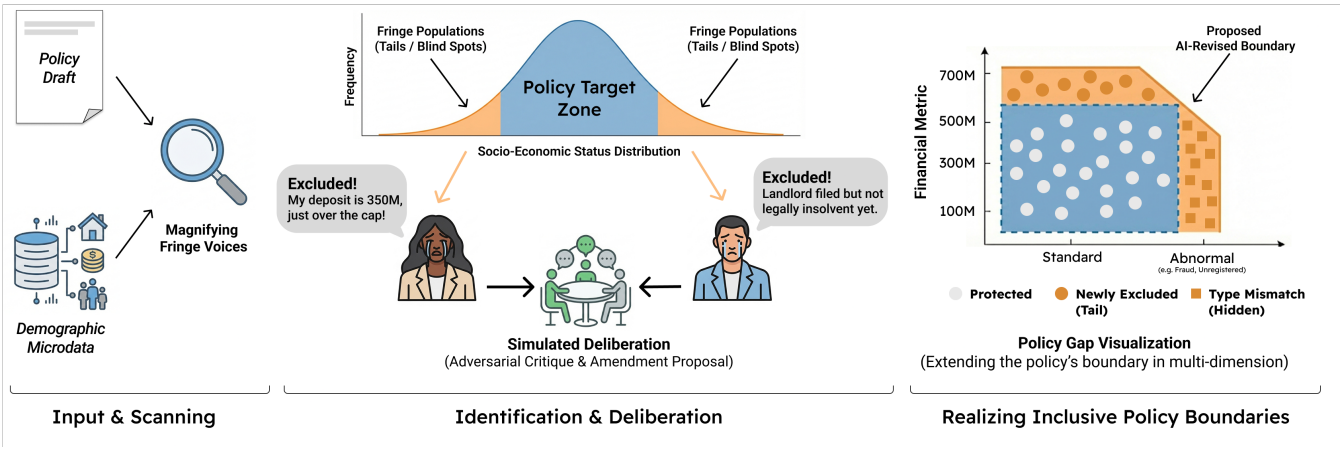

Figure 1: The *Let Them Speak* workflow: discovering policy blind spots by targeting individuals at the boundary of eligibility and generating their voices to inform legislative revision.

Abstract

LLM-based social simulation has gained attention as a tool for predicting policy outcomes, yet growing evidence suggests that LLMs cannot faithfully represent the full diversity of society. We turn this limitation into a feature. Rather than simulating all citizens, we propose *Let Them Speak*, a framework that deliberately identifies the people a policy structurally excludes and generates their voices. Given a policy text and demographic microdata, the system automatically discovers individuals at the boundary of eligibility—those excluded by narrow numerical margins or by gaps the law fails to anticipate—and instantiates them as LLM personas who articulate policy critiques endogenously, without researcher-injected hypotheses. Applied to South Korea's Jeonse Fraud Special Act, the framework's proposed amendment matched 5 of 8 criteria in the actual 2026 legislative revision, compared to 1 of 8 from a randomly sampled control group. These results suggest that targeted blind-spot discovery can yield more actionable legislative recommendations than untargeted simulation.

PoliSim '26, Barcelona, Spain

ACM ISBN 978-1-4503-XXXX-X/26/04
https://doi.org/XXXXXXX.XXXXXXX

CCS Concepts

• **Computing methodologies** → **Multi-agent systems**; • **Applied computing** → *Law*.

Keywords

LLM Simulation, Policy Auditing, Legislative Blind Spots, Agent-Based Modeling, Algorithmic Fairness

ACM Reference Format:
Jaywoong Jeong. 2026. Let Them Speak: Finding Who Policy Leaves Behind via Blind-Spot-Targeted LLM Simulation. In *Proceedings of PoliSim@CHI 2026: LLM Agent Simulation for Policy (PoliSim '26).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

LLM-based social simulation has emerged as a compelling paradigm for studying social phenomena at scale. A language model conditioned on a persona description can elicit responses that closely mirror those of the demographic group it represents [2]. Scaling this to hundreds of agents in an interactive environment yields a virtual micro-society—one that can be queried, stressed, and iterated upon without the burdens of human-subject experimentation [14].

This paradigm has been applied to macroeconomic modeling [12], public health [9], tax policy [10], and legal proceedings [8]. Its potential as a *policy simulation* tool—predicting the effects of legislation before deployment—has attracted particular attention.

However, recent studies have raised concerns about this approach. LLMs prompted with identity-group personas may produce *out-group imitations*—reflecting stereotypes rather than authentic

perspectives [19]. Training data that over-represents majority viewpoints can leave marginalized populations inadequately captured [6, 16]. Furthermore, minor prompt variations can meaningfully influence simulation outcomes [5]. Even optimistic assessments of LLM-based simulation acknowledge vulnerabilities like sycophancy, prompt sensitivity, and poor generalization [1].

Existing work has primarily addressed these limitations by making LLM outputs more closely match real human responses—through richer prompting, fine-tuning on social science data, and demographic calibration [11]. Efforts to incorporate minority or plural perspectives have also been ongoing [4, 7, 17]. However, these approaches share the goal of better representing the overall population. In the context of policy simulation, even faithful reproduction of population-level averages leaves a gap: the voices at the statistical margins—where policy blind spots arise—remain structurally inaudible.

We ask a different question: if LLMs cannot faithfully represent everyone, can we instead *deliberately sample those most likely to be excluded* and amplify their voices? Just as algorithmic auditing uses controlled synthetic identities—"sock puppets"—to expose discrimination invisible to aggregate analysis [3, 15], we apply this audit logic to legislative text. By identifying individuals who fall just outside a policy's eligibility boundary—and by surfacing blind spots in legislative design that no stakeholder anticipated—we can stress-test a policy's inclusiveness before it takes effect.

We present ***Let Them Speak***, a framework for ex-ante policy auditing through blind-spot-targeted LLM simulation. Our approach translates a policy's eligibility rules into machine-executable expressions and identifies individuals whom the policy structurally excludes—whether by a narrow numerical margin or by a gap the law fails to anticipate. These individuals are instantiated as LLM personas who critique the policy and propose amendments through adversarial deliberation. We apply this framework to South Korea's Jeonse Fraud Special Act and show that targeted simulation surfaces structural exclusions that conventional analysis misses. Rather than replacing human deliberation, *Let Them Speak* augments it: providing legislators with a computationally generated "minority report" before a policy takes effect.

2 Framework

We propose a framework for automatically identifying structural blind spots in draft legislation. Given a policy text and demographic microdata as input, the framework discovers who the policy excludes and produces a concrete proposed amendment. Traditionally, discovering such blind spots and gathering stakeholder perspectives has required costly, time-intensive processes—town hall meetings, congressional hearings, and public comment periods. Our framework performs this function computationally.

The pipeline operates as follows. An LLM first reads the policy text and converts its eligibility criteria into machine-executable conditional expressions, which are then mapped to the corresponding columns in the microdata. This enables two directions of blind spot discovery.

First, individuals who fail the conditions—those excluded by a narrow numerical margin or by an administrative classification that diverges from reality—are filtered directly from the data.

Second, standardized microdata lacks the granular edge cases and unexpected real-world situations that a drafted policy may fail to anticipate. Some individuals appear to satisfy the conditions based on available data columns but may fall outside the policy's actual scope. To surface these invisible blind spots, our system employs **counterfactual prompting**. By forcing a base persona into a contradictory scenario—acting as a genuine victim who is explicitly excluded from the draft—the LLM must deduce a plausible logical narrative explaining their exclusion. In this process, the LLM generates new conditions and edge cases that the drafted law failed to anticipate.

A key design principle is that the LLM is not informed of the exclusion in advance. It receives only the individual's raw data and the legislative text. This ensures that the policy critique emerges endogenously from the model's own legal reasoning, rather than from hypotheses injected by the researcher.

The resulting statements are aggregated by a legislator-role LLM, which synthesizes them into specific article-level amendments to the original policy. We validate the framework's predictive accuracy through a natural experiment. The pipeline is applied to a *past* version of a law, and the proposed amendment is compared against the revision that the legislature actually enacted afterward. Because the model's training data has a knowledge cutoff prior to the real amendment, the possibility that the model memorized the answer is ruled out.

3 Case Study: Jeonse Fraud Special Act

We applied the framework to the South Korean *Jeonse Fraud Special Act*, enacted in June 2023 to protect tenants who lost deposits to fraudulent landlords. We compared the system's output against the actual amended Act passed in January 2026. The model used was GPT-4o [13], whose training data has a knowledge cutoff of October 2023, ensuring it could not have encountered the 2026 amendment. The demographic microdata source was the Survey of Household Finances and Living Conditions [18].

The Act defines victim eligibility through four conjunctive conditions: (1) tenant status with legal standing, (2) deposit $\leq$ 300M KRW, (3) insolvency of the landlord with multiple affected tenants, and (4) evidence of fraudulent intent. Since the survey captures household financial status rather than external events, conditions (3) and (4)—which depend on landlord behavior and criminal proceedings—cannot be evaluated from the data. The mapping therefore focuses on occupancy type, housing classification, and deposit amount. This is precisely the gap the LLM is designed to address: the data identifies *who could be affected*, while the LLM reasons about what the law fails to cover.

Based on the framework in Section 3, we constructed a targeted group of N=100. First, following the direct filtering approach, we sampled N=50 tenants narrowly failing numerical conditions (e.g., deposits just over 300M KRW). Second, to surface invisible blind spots, we sampled N=50 individuals whose baseline data appeared eligible. We subjected them to counterfactual prompting, forcing the base persona into a contradictory scenario: acting as a genuine victim who is explicitly excluded from the draft. This scenario forced the LLM to hypothesize edge cases absent from both the data columns and the bill text, such as tenants facing "trust fraud"

(complex ownership structures) or living in "illegal buildings" (non-residential commercial registries). A control group of N=100 was randomly sampled from the general *jeonse* tenant population without the counterfactual constraints. Both groups were processed through the pipeline.

We extracted eight ground-truth criteria by having the LLM systematically compare the conditional structure of the 2023 and 2026 versions of the Act. Table 1 summarizes the results. The targeted group predicted 5 of 8 amendment criteria, while the control group predicted only 1.

Table 1: Proposed Amendment vs. Actual 2026 Amendment.

	Criterion	Type	Target	Control
g_1	Raise deposit cap (300M → 500M, max 700M KRW)	Num.	✓	×
g_2	Define "multiple victims" as ≥2	Num.	×	×
g_3	Extend law validity (2 → 4 years)	Num.	✓	×
g_4	Include illegal buildings in purchase scope	Cat.	✓	✓
g_5	Include trust-fraud housing in purchase scope	Cat.	✓	×
g_6	Introduce auction-surplus rent subsidy mechanism	Logic	×	×
g_7	Set support application deadline (3yr / 1yr)	Logic	×	×
g_8	Allow streamlined non-victim decisions	Logic	✓	×
Total			**5/8**	1/8

The three criteria missed by the targeted group—g_2, g_6, and g_7—represent administrative upgrades and novel institutional mechanisms rather than direct coverage expansions. For instance, g_6 (auction-surplus subsidy) is an innovative financing mechanism that redirects auction price gaps into rent support. Such design does not emerge from individual complaint data, but requires human policy engineering. Its absence confirms our design principle: the framework surfaces *who is excluded* and *why*, leaving *how to fund solutions* to human expertise.

Conversely, the control group only predicted g_4 (illegal buildings). Random sampling captures some tenants in non-residential registries, making this issue visible without targeted discovery. However, remaining criteria like trust-fraud inclusions and deposit cap adjustments required the direct filtering of boundary cases and counterfactual prompting that untargeted random sampling missed.

4 Discussion and Conclusion

Traditional policy evaluation asks: "Did this law achieve its intended outcomes?" This question is asked after enactment, after budgets are allocated, and—critically—after real people have already fallen through the cracks. The framework presented here inverts this timeline. By computationally generating the voices of the excluded *before* a policy takes effect, *Let Them Speak* reframes LLM simulation not as a predictive model that must be "accurate," but as a stress-testing tool that must be *adversarial*—deliberately seeking failure modes rather than confirming expected outcomes.

We emphasize that boundary personas are not substitutes for real stakeholder engagement. They are synthetic constructs whose narratives, however grounded in microdata, remain products of a language model with its own biases. Furthermore, by deliberately targeting the boundaries of a policy, this framework risks over-amplifying noise—identifying hyper-specific edge cases whose actual probability is infinitesimally small or whose legislative accommodation would be prohibitively costly. The value of this framework therefore lies not in dictating policy, but in providing the *systematic coverage* of exclusion zones that human deliberation may overlook. We advocate for its use as a complement to—not a replacement for—public hearings and participatory processes. A natural extension of this work is to apply this framework to future legislation where systematic analysis of blind spots is critical, such as welfare and subsidy programs.

Our case study on the Jeonse Fraud Special Act demonstrates that targeted simulation surfaces actionable exclusions that conventional analysis misses, predicting 5 of 8 actual amendment criteria compared to 1 of 8 from random sampling. We hope this work contributes to a future where no policy is enacted without first hearing from those it would leave behind.

References

[1] Jacy Reese Anthis, Johnie Kim, and Janet Z. Yang. 2025. Position: LLM Social Simulations Are a Promising Research Method. *Proceedings of the 42nd International Conference on Machine Learning (ICML)* (2025).

[2] 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 (2023), 337–351.

[3] Joshua Asplund, Motahhare Eslami, Hari Sundaram, Christian Sandvig, and Karrie Karahalios. 2020. Auditing Race and Gender Discrimination in Online Housing Markets. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, Vol. 14. 24–35.

[4] Michiel A. Bakker, Martin J. Chadwick, Hannah R. Sheahan, Michael Henry Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matthew M. Botvinick, and Christopher Summerfield. 2022. Fine-tuning Language Models to Find Agreement Among Humans with Diverse Preferences. *arXiv [cs.LG]* (2022).

[5] Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza-del Arco, Johannes B Gruber, and Dirk Hovy. 2025. Large language model hacking: Quantifying the hidden risks of using LLMs for text annotation. *arXiv [cs.CL]* (Oct. 2025).

[6] Esin Durmus, Karina Nyugen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards Measuring the Representation of Subgroups in Language Models. *arXiv preprint arXiv:2310.00321* (2024).

[7] Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA.

[8] Zhitao He, Pengfei Cao, Chenhao Wang, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Kang Liu, and Jun Zhao. 2024. AgentsCourt: Building judicial decision-making agents with court debate simulation and legal knowledge augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, Stroudsburg, PA, USA, 9399–9416.

[9] Abe Bohan Hou, Hongru Du, Yichen Wang, Jingyu Zhang, Zixiao Wang, Paul Pu Liang, Daniel Khashabi, Lauren Gardner, and Tianxing He. 2025. Can A society of generative agents simulate human behavior and inform public health policy? A case study on vaccine hesitancy. *arXiv [cs.MA]* (July 2025).

[10] Seth Karten, Wenzhe Li, Zihan Ding, Samuel Kleiner, Yu Bai, and Chi Jin. 2025. LLM Economist: Large Population Models and Mechanism Design in Multi-Agent Generative Simulacra. In *NeurIPS 2025 Workshop on Algorithmic Collective Action*.

[11] Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S Bernstein. 2025. Finetuning LLMs for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, 30084–30099.

[12] Nian Li, Chen Gao, Mingyu Li, Yong Li, and Qingmin Liao. 2023. EconAgent: Large language model-empowered agents for simulating macroeconomic activities. *arXiv [cs.AI]* (Oct. 2023).

[13] OpenAI. 2024. GPT-4o System Card. https://platform.openai.com/docs/models/gpt-4o

[14] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. ACM, New York, NY, USA.

[15] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. In *Proceedings of "Data and Discrimination: Converting Critical Concerns into Productive Inquiry," a preconference at the 64th Annual Meeting of the International Communication Association*. Seattle, WA.

[16] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect?. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

[17] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. A Roadmap to Pluralistic Alignment. arXiv:2402.05070 [cs.AI] https://arxiv.org/abs/2402.05070

[18] Statistics Korea, Bank of Korea, and Financial Supervisory Service. 2023. Survey of Household Finances and Living Conditions, 2023. https://mods.go.kr

[19] Angelina Wang, Jamie Morgenstern, and John P. Dickerson. 2024. Large Language Models that Replace Human Participants Can Harmfully Misportray and Flatten Identity Groups. *arXiv preprint arXiv:2402.01908* (2024).