SEBCM: Semantic Embedding-based Bounded Confidence Modeling

Simon Münker, Simon Werner, Achim Rettinger
Trier University, {muenker,werners,rettinger}@uni-trier.de

Abstract

Opinion dynamics models have traditionally relied on scalar representations of beliefs, limiting their ability to capture the semantic richness of real-world opinions. We propose **Semantic Embedding-based Bounded Confidence Modeling (SEBCM)**, a novel framework that integrates language model embeddings with classical bounded confidence models. Rather than treating beliefs as abstract scalars, SEBCM utilizes opinion generation combined with socio-demographic personas as explicit contextual drivers of each agent's initial stance. By representing these contextually grounded opinions as high-dimensional semantic embeddings, we enhance interpretability while enabling sociology-based opinion dynamic simulations. As a case study, we evaluate our framework on four language models across seven polarizing topics. Our results demonstrate that embedding-based bounded confidence models exhibit similar convergence patterns and homophily effects as traditional scalar-based approaches, while offering enhanced empirical realism, semantic explainability, and sensitivity to the context of participating agents. Thus, our work bridges quantitative sociology, NLP, and AI transparency research, establishing a hybrid methodology for grounded opinion simulation.

1 Introduction

Language is never de-contextualized: who speaks, from which demographic vantage point, and within which social structure fundamentally shapes what is said and how it is received. As Large Language Models (LLMs) are increasingly deployed in socially consequential settings (Törnberg et al., 2023; Li and Yao, 2025), from content recommendation to policy deliberation (Gao et al., 2024; Chuang et al., 2024; Ball et al., 2025), a question emerges: How does the socio-demographic context encoded in LLM interactions translate into collective opinion dynamics? Understanding this is critical for AI transparency (Ball et al., 2025), yet existing simulation tools occupy two disconnected extremes: classical agent-based models that are mathematically grounded but treat agents as context-free scalars (Flache et al., 2017), and LLM-based chat simulations that are textually rich but methodologically opaque (Argyle et al., 2023; Larooij and Törnberg, 2025). Neither framework accounts for the demographic and cultural identities that may drive disagreement. SEBCM bridges this gap. Grounded in the Deffuant-Weisbuch bounded confidence mechanism (Deffuant et al., 2000; Flache et al., 2017), it replaces opinion scalars with semantic embedding vectors derived from opinions generated by LLMs combined with socio-demographic personas (Hu et al., 2024). Our approach yields a transparent, auditable simulation whose dynamics are controlled by established sociological parameters while remaining sensitive to the user- and group-related context that shapes the initial opinion landscape.

From an AI transparency and contextual NLP standpoint, SEBCM makes three contributions. First, it provides observable, measurable opinion trajectories for every agent, enabling systematic audit of how socio-demographically diverse agents' views evolve under model-mediated social influence. Second, the semantic explainability pipeline translates opaque embedding into human-readable stance distributions. Showing, for instance, how the proportion of agents classified as "Strongly Oppose" or "Moderately Support" shifts from initial to final state, directly making AI-driven opinion dynamics legible without requiring domain expertise. Third, by grounding LLM behavior within a parameter-controlled sociological model, SEBCM enables counterfactual analysis: researchers can vary ϵ, μ, network topology, or the demographic composition of the agent population to probe how demographic diversity, model-specific biases in context-sensitive opinion generation, and structural factors interact to shape simulated societies.

2 Background

Bounded confidence models (BCMs), introduced by Deffuant et al. (2000); Hegselmann and Krause (2002), simulate pairwise opinion exchange: agents update their views only when the distance between their opinions falls within a threshold ϵ. Varying ϵ reproduces the full spectrum of sociological outcomes: global consensus (large ϵ), polarization (moderate ϵ), and fragmentation (small ϵ) (Flache et al., 2017). However, classical BCMs abstract away two crucial dimensions of real-world opinion formation. First, as Hegselmann (2023) highlights, representing beliefs as scalars discards the semantic content that drives real disagreement. Second, and equally important for contextually situated modeling, classical BCMs treat agents as anonymous and homogeneous: they carry no demographic identity, cultural background, or social position that would differentiate their initial stances or constrain whom they are willing to influence. Real opinion dynamics, by contrast, are fundamentally shaped by who the agents are (Flache et al., 2017).

Work in computational social science has explored LLMs as silicon samples (Argyle et al., 2023), conditioning model outputs on demographic attributes to simulate survey populations. While appealing for their diversity, such approaches lack methodological grounding and resist reproducible analysis (Münker et al., 2026; Balluff et al., 2026). Crucially, it remains unclear whose opinions LLMs reflect and how that demographic sensitivity propagates through multi-agent interaction (Santurkar et al., 2023). Our proposed approach addresses this gap by establishing a framework that preserves the structural integrity and sociological metrics of BCMs (Flache et al., 2017), incorporates the semantic interpretability of LLM-based embeddings (Luo et al., 2025; Mahajan et al., 2025), and grounds each agent's initial opinion in an explicit sociodemographic context, making the influence of user- and group-related context on collective dynamics observable and measurable.

3 Methodology

3.1 Opinion/Network Generation

Given a set of N personas $\mathcal{P} = \{p_1, \ldots, p_N\}$ and a topic τ, we generate textual opinions using a LLM (Kwon et al., 2023). Each persona p_i consists of demographic and ideological descriptors from the NVIDIA Nemotron Personas dataset (Meyer and Corneil, 2025). We construct instructions of the form: "You are `[persona description]`. Express your opinion on `[topic]` in 2-3 sentences." The language model generates opinion statements $\mathcal{O} = \{o_1, \ldots, o_N\}$, which we treat as the initial opinion distribution in our simulation.

Semantic Embedding We map each textual opinion o_i to a semantic embedding vector $\mathbf{v}_i \in \mathbb{R}^d$ using a pre-trained sentence embedding model (Reimers and Gurevych, 2019; Nussbaum et al., 2025): $\mathbf{v}_i = \text{Embed}(o_i)$, which provide semantically meaningful embeddings where cosine similarity correlates with semantic similarity. All embeddings are L2-normalized for consistency with cosine distance calculations.

Network Initialization We initialize a social network $G = (V, E)$ where $V = \{1, \ldots, N\}$ represents agents and E represents social connections and connect them via random graphs using established models from network science (Erdős-Rényi & Barabási-Albert (Newman, 2002)) to control for topological effects.

Optional: Similarity-Based Networks We apply clustering algorithms (e.g., k-means, hierarchical clustering) to $\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ to discover latent opinion structure. Edges connect agents within the same cluster or whose embeddings have high cosine similarity, creating networks with realistic homophily. This allows us to compare randomly connected populations with communities that naturally form around shared beliefs.

3.2 Bounded Confidence Dynamics

Given the network G and initial embeddings $\{\mathbf{v}_i^{(0)}\}_{i=1}^N$, we simulate opinion dynamics using an adapted bounded confidence model. At each time step t, for each agent i with neighbors $\mathcal{N}_i$, we compute:

$$\mathbf{v}_i^{(t+1)} = \mathbf{v}_i^{(t)} + \mu \sum_{j \in \mathcal{N}_i} w_{ij}^{(t)} (\mathbf{v}_j^{(t)} - \mathbf{v}_i^{(t)})$$

where the weight $w_{ij}^{(t)}$ is determined by the bounded confidence rule:

$$w_{ij}^{(t)} = \begin{cases} \frac{1}{|\{k \in \mathcal{N}_i : d(\mathbf{v}_i^{(t)}, \mathbf{v}_k^{(t)}) < \epsilon\}|} & \text{if } d(\mathbf{v}_i^{(t)}, \mathbf{v}_j^{(t)}) < \epsilon \\ 0 & \text{otherwise} \end{cases}$$

Here, $d(\mathbf{v}_i, \mathbf{v}_j) = 1 - \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|}$ is the cosine distance, $\epsilon \in [0, 2]$ is the confidence bound, and

$\mu \in (0, 1]$ is the convergence parameter. *Interpretation:* Agents only update their opinions based on neighbors whose embeddings are semantically similar (within ϵ distance). The parameter μ controls how quickly agents adjust toward similar neighbors.

Convergence Analysis We track opinion dynamics by recording the embedding history $\mathcal{H} = \{\{\mathbf{v}_i^{(t)}\}_{i=1}^N\}_{t=0}^T$ for T time steps. To quantify convergence, we compute the variance of the opinion distribution at each time step:

$$\text{Var}(t) = \sum_{k=1}^{d} \text{Var}(\{v_{i,k}^{(t)}\}_{i=1}^N)$$

where $v_{i,k}^{(t)}$ is the k-th dimension of agent i's embedding at time t. Decreasing variance indicates opinion convergence, while persistent high variance suggests polarization or fragmentation.

3.3 Semantic Explainability

To enable interpretation of the simulation results, we develop a semantic explainability method:

Keyword Labeling We instruct the language model to assign predefined keywords (e.g., "pro-choice", "pro-life", "moderate") to initial opinions $\mathcal{O}$ using constrained generation.

Classifier Training We train a multi-label classifier $f : \mathbb{R}^d \to \{0, 1\}^K$ that maps embeddings to the K keywords, using the labeled data from initial opinions.

Post-Simulation Analysis After simulation, we apply the classifier to final embeddings $\{\mathbf{v}_i^{(T)}\}$ to predict how keyword distributions have shifted, revealing semantic changes in the opinion landscape.

4 Experiment

We evaluate SEBCM across seven politically charged topics *(abortion, climate change, gun control, immigration, same-sex marriage, social media, vaccines)* using 50 agents per run. The simulation uses $\epsilon = 0.5$ and $\mu = 0.1$ over 20 steps on Erdős–Rényi graphs ($p = 0.3$). We generate opinions with four instruction-tuned LLMs (Llama-3.1-8B (Touvron et al., 2023), Mistral-7B (Liu et al., 2026), Qwen3-8B (Bai et al., 2023), Gemma2-9B (Team et al., 2024)) served via vLLM (Kwon et al., 2023) for efficient inference. We release the full implementation and results as an open-source GitHub repository: `https://github.com/cl-trier/sebcm`.

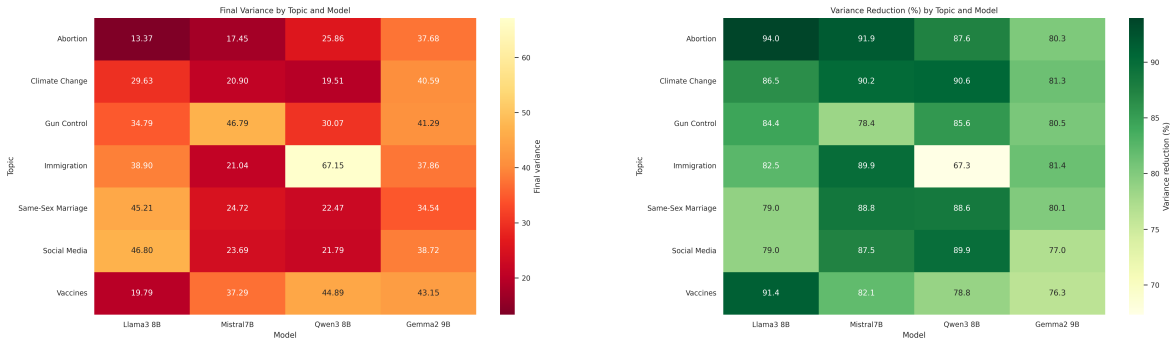
4.1 Results

Overall Convergence Across all runs (28 model/topic combinations), SEBCM achieved a mean variance reduction of 83.95% with a mean final variance of 33.07. The half-reduction step averaged between 3.3 and 4.0 across all conditions, indicating that the bulk of semantic convergence is concentrated in the early simulation steps. The mean fraction of monotonically decreasing steps exceeded 0.95 for all but one condition (Qwen3-8B on immigration, 0.65), confirming stable convergence dynamics across the embedding space.

Model-Level Performance Mistral-7B achieves the lowest mean final variance (27.41 ± 10.62) and the highest mean variance reduction ($86.98\% \pm 4.91\%$) and consensus gain ($80.50\% \pm 6.64\%$). Gemma2-9B ranks last on all three primary metrics (final variance 39.12 ± 2.83; variance reduction $79.55\% \pm 2.06\%$; consensus gain $71.78\% \pm 2.83\%$), and notably exhibits the lowest variance across topics, suggesting less sensitivity to topic difficulty. Llama-3.1-8B and Qwen3-8B occupy intermediate positions, with Qwen3-8B showing the highest within-model variability (variance reduction std 8.35%), driven largely by a single anomalous run.

Topic-Level Difficulty Figure 1 shows that topics differ in convergence difficulty. *Abortion* yields the lowest mean final variance (23.59 ± 10.73) and the highest mean variance reduction ($88.42\% \pm 6.04\%$), making it the easiest topic for all models to converge on. *Immigration* is the hardest topic (mean final variance 41.24 ± 19.12; mean variance reduction $80.29\% \pm 9.43\%$), with the largest cross-model spread, indicating that model choice matters more for contentious, multi-faceted topics. *Gun control, vaccines, and same-sex marriage* form an intermediate difficulty band.

Statistical Tests A topic-wise Friedman omnibus test (Sheldon et al., 1996) reveals no statistically significant differences across models for any of the three primary metrics: final variance ($\chi^2 = 2.83$, $p = 0.419$), variance reduction percentage ($\chi^2 = 6.60$, $p = 0.086$), and consensus gain percentage ($\chi^2 = 3.86$, $p = 0.277$). Holm-corrected Wilcoxon signed-rank tests (Zimmerman and Zumbo, 1993) for all pairwise comparisons



(a) Final variance (lower is better).

(b) Variance reduction % (higher is better).

Figure 1: Per-topic, per-model convergence metrics across all 28 runs. Mistral-7B achieves uniformly low final variance; immigration is the consistently hardest topic, with Qwen3-8B producing a pronounced outlier (67.15 final variance, 67.3% reduction).

likewise yield no significant contrasts after correction, though the raw Mistral-7B vs. Gemma2-9B comparisons approach nominal significance on both final variance ($p_{\text{raw}} = 0.031$) and variance reduction ($p_{\text{raw}} = 0.031$).

5 Discussion

SEBCM reproduces the BCM behavior in high-dimensional embedding space in our presented experiment. Variance curves decay sharply within the first three to four steps and plateau thereafter, mirroring the fast-convergence dynamics characteristic of scalar BCMs under moderate confidence bounds (Hegselmann, 2023). The consistency of the half-reduction step (≈ 3 to 4 across all 28 conditions) suggests that the Erdős–Rényi topology at $p = 0.3$, combined with $\epsilon = 0.5$, creates sufficient connectivity for early consensus formation regardless of model or topic. The persistent residual variance, may reflect semantic heterogeneity in LLM-generated opinions rather than numerical noise, a property that scalar BCMs cannot expose.

Model differences are cosmetic Mistral-7B edges ahead on all convergence metrics, while Gemma2-9B consistently lags. However, the absence of statistically significant test results across all three metrics, combined with small pairwise effect sizes after correction, indicates that model choice has a second-order influence on convergence dynamics under the current parameter regime. This is substantively informative: it suggests that SEBCM's dynamics are primarily governed by the sociological parameters ϵ and μ rather than idiosyncrasies of the underlying language model, supporting the framework's generalizability across model families. The notable exception is Gemma2-9B's unusually low cross-topic variance in convergence metrics, implying that this model generates semantically less diverse initial opinion distributions.

Topic difficulty reflects real-world polarization The topic-level results align well with empirical measures of political polarization (Baldassarri and Bearman, 2007; Fiorina and Abrams, 2008). *Immigration,* the hardest topic (highest final variance, lowest variance reduction), is consistently ranked among the most divisive issues in US public opinion research (Wright and Levy, 2020), and the Qwen3-8B anomaly on this topic (the only run with non-monotone convergence) suggests that high semantic dispersion in initial embeddings can prevent consensus formation under a fixed ϵ. Conversely, *abortion's* status as the easiest topic is initially counterintuitive given its cultural salience, but may reflect that LLM-generated opinions on abortion cluster along a single salient semantic dimension (support/oppose) that aligns well with cosine distance, enabling efficient convergence. These patterns are consistent with prior observations that framing, rather than mere valence, drives opinion divergence (Flache et al., 2017).

Conclusion SEBCM offers a principled foundation for transparent LLM-based social simulation grounded in the social sciences. Future work should extend the framework to non-US persona distributions, explore demographic disaggregation of convergence trajectories, and investigate how network topology interacts with demographic homophily, opening productive directions at the intersection of contextually situated NLP, computational sociology, and AI transparency research.

Limitations

While SEBCM provides a novel bridge between quantitative sociology and natural language processing, our current study has several important limitations that must be acknowledged:

Empirical Validation and Persona Realism Our simulation initializes the opinion landscape using LLM-generated opinions prompted by socio-demographic persona cues (Meyer and Corneil, 2025). However, current research on the validity of persona-based LLM generation is inconclusive (Larooij and Törnberg, 2025; Münker, 2025), leaving it unclear to what extent these generated stances are representative of real-world human opinions.

Embedding Semantics vs. Stance SEBCM relies on the assumption that mathematical similarity between sentence embeddings in high-dimensional vector space accurately reflects ideological alignment. However, embeddings extracted from generalized encoding models may inadvertently capture superficial lexical or stylistic similarities (Agarwal et al., 2019; Joseph and Morgan, 2020) rather than underlying opinions on a semantical level.

Vector Averaging A core assumption of SEBCM is that the linear averaging of embedding vectors, driven by the convergence parameter, corresponds to a meaningful shift in human opinion. Because we track consensus via the mathematical reduction of embedding variance, there is a significant risk that these final vectors drift into semantic "dead zones", which are positions in the sentence-level embedding space that cannot be decoded to natural text (Coenen et al., 2019). Without analyzing the manipulated vectors and decoding them into natural text, we cannot verify whether they still correspond to coherent, logical human stances.

Future Work

To address these limitations and advance SEBCM into a more rigid, better grounded, and explainable instrument, future research could focus on the following directions:

Human Grounding Future iterations of SEBCM could validate the simulated opinion dynamics against real-world opinion distributions. Aligning our experimental pipeline with established survey datasets, e.g., the American National Election Studies (Malhotra and Krosnick, 2007) or the German Longitudinal Election Study (Schmitt-Beck et al., 2010), or demographic benchmarks will be vital for verifying empirical realism.

Decoding and Coherence Verification To resolve the risk of vectors drifting into semantic "dead zones", future work could implement an inversion or generative decoding mechanism to translate post-simulation embeddings back into natural language (Kugler et al., 2024). This will enable qualitative analysis of opinion shifts from the initial to final states, verifying that the mathematical convergence reflects logically coherent human opinions.

Stance-Aware Embeddings Researchers could explore contrastive embedding models (Khosla et al., 2020) that are explicitly fine-tuned for stance detection rather than general semantic similarity. This would ensure that vector distances strictly isolate ideological disagreement. However, using specialized embeddings could prevent the decoding feature suggested previously. Thus, we think there exists a trade-off between the decodability of the sentence representations and the expressiveness in terms of sentiment or stance.

Extended Parameter Exploration Comprehensive parameter sweeps are needed to test our framework's full generalizability. This includes varying the confidence bounds and convergence rates, testing different network topologies, and simulating much larger populations over extended time horizons. The presented work on scalar-based bounded confidence modeling (Hegselmann and Krause, 2002; Flache et al., 2017; Hegselmann, 2023) could serve as a rich starting point to explore experimental setups and convergence behavior.

Acknowledgments

We thank Kai Kugler, Nils Schwager, Christoph Hau and Raghvi Baloni for our constructive discussions. This study was conducted with a financial contribution from the EU's Horizon Europe Framework (HORIZON-CL2-2022-DEMOCRACY-01-07) under grant agreement number 101095095. This work is part of the EVAS research cooperative `https://evas-research.github.io`, a community that challenges assumptions and advances methods for the evaluation of language models as reliable human proxies.

References

Oshin Agarwal, Funda Durupınar, Norman I. Badler, and Ani Nenkova. 2019. Word embeddings (also) encode human personality stereotypes. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 205–211, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Delia Baldassarri and Peter Bearman. 2007. Dynamics of Political Polarization. *American Sociological Review*, 72(5):784–811.

Sarah Ball, Simeon Allmendinger, Frauke Kreuter, and Niklas Kühl. 2025. Human preferences in large language model latent space: A technical analysis on the reliability of synthetic data in voting outcome prediction. *Preprint*, arXiv:2502.16280.

Paul Balluff, Justin Chun-ting Ho, Johannes B. Gruber, Sean Palicki, Alexis Palmer, Luca Rossi, Irina Shklovski, and Chung-hong Chan. 2026. Newer, Larger, Better? A Critique of the Unreflective LLM Adoption in Communication Research. *Political Communication*, 43(3):572–581.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating opinion dynamics with networks of LLM-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346, Mexico City, Mexico. Association for Computational Linguistics.

Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. 2019. Visualizing and measuring the geometry of BERT. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 771, pages 8594–8603. Curran Associates Inc.

Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. 2000. Mixing beliefs among interacting agents. *Advances in Complex Systems*, 03(01n04):87–98.

Morris P. Fiorina and Samuel J. Abrams. 2008. Political Polarization in the American Public. *Annual Review of Political Science*, 11(Volume 11, 2008):563–588.

Andreas Flache, Michael Mäs, Thomas Feliciani, Edmund Chattoe-Brown, Guillaume Deffuant, Sylvie Huet, and Jan Lorenz. 2017. Models of social influence: Towards the next frontiers. *Journal of Artificial Societies and Social Simulation*, 20(4):2.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large Language Models empowered agent-based modeling and simulation: A survey and perspectives. *Humanities and Social Sciences Communications*, 11(1):1259.

Rainer Hegselmann. 2023. Bounded confidence revisited: What we overlooked, underestimated, and got wrong. *Journal of Artificial Societies and Social Simulation*, 26(4):11.

Rainer Hegselmann and Ulrich Krause. 2002. Opinion dynamics and bounded confidence: models, analysis and simulation. *Journal of Artifical Societies and Social Simulation (JASSS) vol*, 5(3).

Jun Hu, Wenwen Xia, Xiaolu Zhang, Chilin Fu, Weichang Wu, Zhaoxin Huan, Ang Li, Zuoli Tang, and Jun Zhou. 2024. Enhancing sequential recommendation via LLM-based semantic embedding learning. In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 103–111, New York, NY, USA. Association for Computing Machinery.

Kenneth Joseph and Jonathan Morgan. 2020. When do word embeddings accurately reflect surveys on our beliefs about people? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4392–4415, Online. Association for Computational Linguistics.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 18661–18673. Curran Associates Inc.

Kai Kugler, Simon Münker, Johannes Höhmann, and Achim Rettinger. 2024. InvBERT: Reconstructing Text from Contextualized Word Embeddings by inverting the BERT pipeline. *Journal of Computational Literary Studies*, 2(1).

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for Large Language Model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Maik Larooij and Petter Törnberg. 2025. Validation is the central challenge for generative social simulation: A critical review of LLMs in agent-based modeling. *Artificial Intelligence Review*, 59(1):15.

Sarah Li and Ziyu Yao. 2025. Evaluating the effectiveness of persona simulation in opinion prediction with GPT-4.1. In *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 2938–2942.

Alexander H. Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, Alexandre Sablayrolles, Amélie Héliou, Amos You, Andy Ehrenberg, Andy Lo, Anton Eliseev, Antonia Calvi, Avinash Sooriyarachchi, Baptiste Bout, and 101 others. 2026. Ministral 3. *Preprint*, arXiv:2601.08584.

Jiao Luo, Yi Ping Yao, Yu Xuan Peng, and Wen Jie Tang. 2025. A transformer-enhanced stochastic bounded confidence model for opinion dynamics. In *Proceedings of the 39th ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, SIGSIMPADS '25, page 19–28, New York, NY, USA. Association for Computing Machinery.

Yash Mahajan, Matthew Freestone, Naman Bansal, Sathyanarayanan N. Aakur, and Santu Karmaker. 2025. Revisiting word embeddings in the LLM era. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2686–2717, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Neil Malhotra and Jon A. Krosnick. 2007. The effect of survey mode and sampling on inferences about political attitudes and behavior: Comparing the 2000 and 2004 ANES to internet surveys with nonprobability samples. *Political Analysis*, 15(3):286–323.

Yev Meyer and Dane Corneil. 2025. Nemotron-Personas-USA: Synthetic personas aligned to real-world distributions. *Preprint*, huggingface.co/datasets:nvidia/Nemotron-Personas-USA.

Simon Münker. 2025. Political Bias in LLMs: Unaligned Moral Values in Agent-centric Simulations. *Journal for Language Technology and Computational Linguistics*, 38(2):125–138.

Simon Münker, Nils Schwager, and Achim Rettinger. 2026. Don't trust generative agents to mimic communication on social networks unless you benchmarked their empirical realism. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1141–1151, Rabat, Morocco. Association for Computational Linguistics.

Mark E. J. Newman. 2002. *Random graphs as models of networks*, chapter 2. John Wiley & Sons, Ltd.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2025. Nomic Embed: training a reproducible long context text embedder. *Preprint*, arXiv:2402.01613.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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*, volume 202 of *ICML'23*, pages 29971–30004, Honolulu, Hawaii, USA. JMLR.org.

Rüdiger Schmitt-Beck, Hans Rattinger, Sigrid Roßteutscher, and Bernhard Weßels. 2010. *Die deutsche Wahlforschung und die German Longitudinal Election Study (GLES)*, pages 141–172. VS Verlag für Sozialwissenschaften, Wiesbaden.

Michael R. Sheldon, Michael J. Fillyaw, and W. Douglas Thompson. 1996. The use and interpretation of the Friedman test in the analysis of ordinal-scale data in repeated measures designs. *Physiotherapy Research International*, 1(4):221–228.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open Language Models at a practical size. *Preprint*, arXiv:2408.00118.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and efficient Foundation Language Models. *Preprint*, arXiv:2302.13971.

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. 2023. Simulating social media using Large Language Models to evaluate alternative news feed algorithms. *Preprint*, arXiv:2310.05984.

Matthew Wright and Morris Levy. 2020. American public opinion on immigration: Nativist, polarized, or ambivalent? *International Migration*, 58(6):77–95.

Donald W. Zimmerman and Bruno D. Zumbo. 1993. Relative Power of the Wilcoxon Test, the Friedman Test, and Repeated-Measures ANOVA on Ranks. *The Journal of Experimental Education*, 62(1):75–86.