

LARGE LANGUAGE MODEL-GENERATED BEHAVIORAL PATTERNS FOR EPIDEMIOLOGICAL MODELING: A MULTI-BENCHMARK EVALUATION FRAMEWORK

Sophia Mao¹, Keshav Gandhi^{1,2}, Alexander Gutfraind¹

¹*Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago, 1603 W. Taylor St., Chicago, IL, USA;* ²*Program for Experimental and Theoretical Modeling, Division of Hepatology, Department of Medicine, Loyola University Chicago, Maywood, IL 60153, USA*

ABSTRACT

Representing individual-level behavioral heterogeneity is challenging in epidemiological modeling, particularly when empirical behavioral data are sparse. We assess the potential of large language model (LLM)-generated behavioral outputs in addressing this challenge, using hepatitis C virus-related behaviors among people who inject drugs as a case study. Correspondence with epidemiological patterns was evaluated using external benchmarks: literature-informed association directions, correlations from an agent-based model (ABM), and survey associations. Across three LLMs, unconditional directional recovery across 28 retained literature-informed associations was 46–68%, compared with 82–93% agreement when LLMs were directly asked to classify the same associations. Mean absolute differences from ABM and survey correlation patterns were 0.11–0.13 and 0.15–0.18, respectively. Together, these findings show partial alignment between LLM-generated behavioral relationships and epidemiological evidence, despite higher knowledge-probe agreement. More systematic evaluation is needed before incorporating LLM-generated behavioral outputs into epidemiological models that simulate transmission dynamics.

Keywords large language models · epidemiological modeling · generative agent-based modeling · behavioral heterogeneity · infectious disease transmission · hepatitis C virus · people who inject drugs

I. Introduction

Large language models (LLMs) have introduced a new possibility for epidemiological modeling: generating individual-level behavioral outputs that can supplement sparse behavioral data. This behavioral heterogeneity is important for epidemiological models that aim to capture individual-level variation in infectious disease transmission dynamics. Established epidemiological modeling approaches, such as microsimulation and agent-based models (ABMs), represent populations of simulated agents, each with individual characteristics and behaviors that require a large number of parameters based on empirical data and/or literature-derived estimates to define their behaviors (Chen et al., 2024; Niedzielewski et al., 2024). However, human behaviors are complex and diverse, making their behavioral parametrization and validation difficult. These requirements are especially challenging for populations where behavioral, biological, and network data are limited (Gutfraind et al., 2015). The empirical characterization of these behaviors and contact structures may also be subject to reporting biases, and can be costly or impractical at the population level (Eames et al., 2015).

One emerging approach for incorporating LLMs into computational modeling is generative agent-based modeling (GABM), which uses LLMs to generate agent behaviors and interactions within simulated environments rather than relying exclusively on predefined rules. This approach has been proposed as a way to represent aspects of human decision-making in computational simulations (Lu et al., 2024; Adornetto et al., 2025). For example, in an epidemic GABM study, LLM-driven agents made context-dependent behavioral decisions, such as whether to self-quarantine when symptomatic or self-isolate as case counts increased. These decisions were fed into an epidemic simulation to see whether the resulting epidemic dynamics qualitatively resembled known epidemiological patterns (Williams et al.,

2023). However, aggregate simulation outcomes that seem plausible do not necessarily imply that the underlying agent behaviors reproduce relationships important for epidemiological modeling.

This challenge is particularly relevant for PWID, among whom HCV transmission risk is shaped by interconnected behavioral, social, and contextual factors, including injection behaviors and housing instability (Arum et al., 2021). Moreover, different sources of behavioral information may capture complementary aspects of human interaction rather than providing interchangeable representations of the same underlying behavior (Thi Phuong et al., 2026). Thus, evaluating LLM-generated behavioral relationships against multiple external benchmarks, rather than relying on a single reference source, may provide a broader assessment of their alignment with existing evidence.

Accordingly, we used a multi-benchmark approach to evaluate LLM-generated behavioral relationships. Agreement between LLMs and external sources was quantified using (1) reproduction of association directions from the epidemiological literature, (2) similarity to ABM correlations, and (3) similarity to empirical survey associations (Figure 1). Profile-level bootstrap resampling was used to quantify uncertainty across benchmarks. Sensitivity was assessed across three LLM families and a semantically equivalent alternative prompt, without seeking to identify a universally optimal LLM. Together, these analyses provide a structured approach to evaluating LLM-generated behavioral relationships before their implementation in epidemiological models that simulate transmission dynamics.

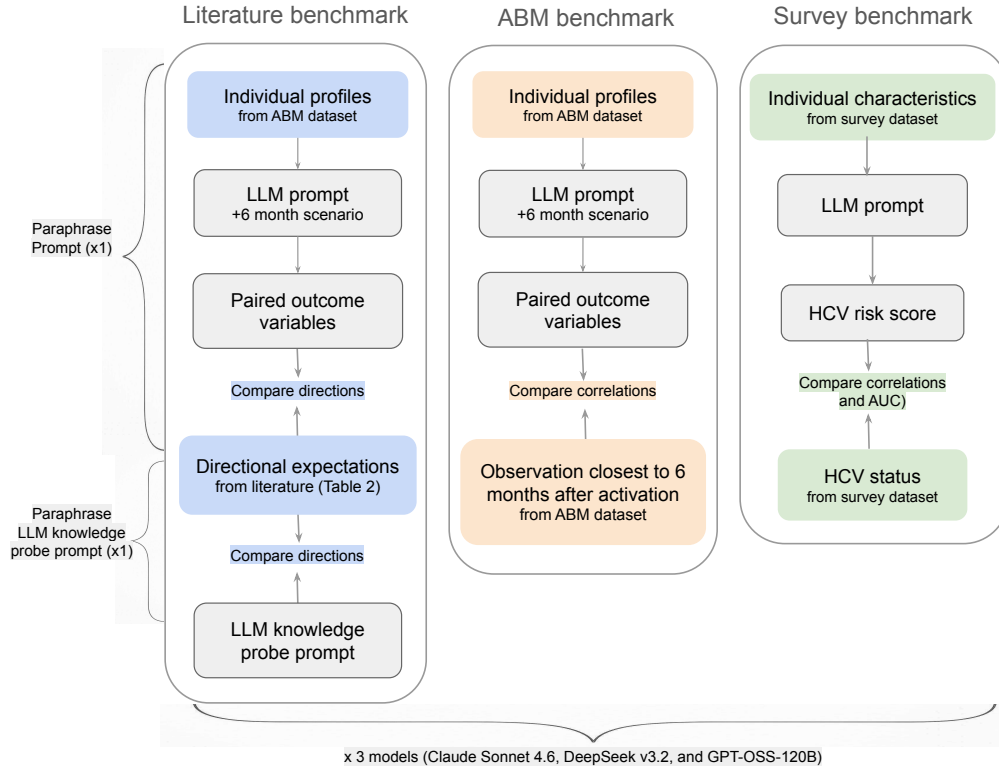


Figure 1: Overview of the multi-benchmark evaluation framework

II. Methods

1. Paired within-individual design

We used a within-individual design in which the same individual profiles were evaluated under two housing conditions. Individual profiles were derived from a published ABM dataset of PWID in Chicago (Gutfraind et al., 2015). The ABM’s synthetic population was constructed primarily using data from the Community Outreach Intervention Projects Needle Exchange Program survey. Characteristics of each agent at the model’s initial time point were provided to the LLM.

224 unique agents were selected through stratified random sampling by sex, race/ethnicity, and injection experience (<3 vs. ≥ 3 years), ensuring equal representation across predefined strata. Sample inclusion criteria were age ≥ 18 years, HCV-susceptible status at baseline, non-zero injection frequency, and at least two recorded time points from the ABM.

Each profile was assigned two housing conditions: stable for at least one year vs. homeless for at least one month. These thresholds were chosen to distinguish long-term housing stability from current or recent homelessness. All other baseline attributes were held constant across housing conditions. The prompt scenario specified a 6-month period in 2026 during which individuals were aware that illicitly manufactured fentanyl was present in the drug supply. LLMs were instructed to generate behavioral outcomes representing a typical 4-week period at the end of this 6-month period.

This design yielded 448 sets of behavioral outputs (224 individuals under two housing scenarios). Prompts were executed independently. Let X_{ic} denote the baseline characteristics for individual profile i under housing condition $c \in \text{stable}, \text{homeless}$, and let S denote the shared prompt context, including the scenario and task instructions. For each prompt, the LLM generated multiple outcome variables (Table 1):

$$Y_{ic}^* = (Y_{ic1}^*, \dots, Y_{icK}^*) \mid X_{ic}, S, \quad (1)$$

In Equation 1, i indicates individual profiles and $k = 1, \dots, K$ indicates the outcomes.

Table 1: LLM-generated outcome variables (4-week recall period)

Outcome variable	Definition	Type
Injection frequency	Average weekly injections	Count
Network in-degree	Unique individuals from whom the person received a used syringe	Count
Network out-degree	Unique individuals to whom the person gave a used syringe	Count
Partners	Unique individuals with at least one shared injection event	Count
Non-injection routes	Non-injection use (smoking or snorting)	Count
Overdose	Non-fatal overdoses	Count
Receptive sharing	Injections involving reuse of a syringe previously used by another person	Proportion
Partners with HCV	Injection partners who are HCV positive	Proportion
Public injecting	Injections occurring in public settings (parks, streets, shelters)	Proportion
Treatment	Enrolled in opioid use disorder treatment program	Binary
HCV status	Current HCV state, restricted to susceptible or acute infection	Binary

Note. Variable definitions and measurements varied across studies and did not necessarily correspond exactly to the outcome variable specified by the LLM prompt.

Three models (Claude Sonnet 4.6, DeepSeek v3.2, and GPT-OSS-120B) were evaluated and are hereafter referred to as Claude, DeepSeek, and GPT-OSS, respectively. LLM temperature was set to zero to minimize stochastic variation across generations for the same input. Thus, the analysis focuses on variation across profiles, models, and the evaluated prompt wording rather than variation from repeated stochastic generations.

2. Evaluation against external benchmarks

a. Reproduction of association directions from the literature

We first assessed whether LLMs could reproduce the directions of associations reported in the epidemiological literature on PWID (Table 2; Supplementary Table S1). 32 literature-informed associations were identified, four of which had mixed or inconclusive prespecified directions. Reported effect estimates (e.g., adjusted odds ratios and risk ratios) from the literature were included but were not quantitatively compared with LLM-generated estimates, as magnitudes vary across populations, settings, and methodological approaches even when the general association is consistent.

Table 2: Prespecified literature-informed directions for associations between housing status and outcome variable

Variable pairs		Direction	Evidence
Homeless	Public injecting	↑	OR = 4.23 (3.17–5.65) (Khezri et al., 2025) (meta-analysis) Positive association reported (Arum et al., 2021) (systematic review)
Homeless	Injection frequency	↑	OR = 1.40 (1.08–1.82) (Ballard et al., 2025) Positive association reported (Arum et al., 2021) (systematic review)
Homeless	Receptive sharing	↑	Positive association reported (Arum et al., 2021) (systematic review) aOR = 1.93 (1.67–2.23) (Hotton et al., 2021) aPR = 1.26 (1.15–1.39) (Kan et al., 2026) (cohort) RR = 1.18 (0.60–2.39) (Morris et al., 2014) (cohort)
Homeless	Positive HCV status	↑	aHR = 1.86 (1.32–2.55) (Morris et al., 2020) (cohort) aRR = 1.64 (1.43–1.89) (Arum et al., 2021) (meta-analysis) aHR = 1.46 (1.21–1.76) (Gahrton et al., 2024) (cohort) aOR = 1.55 (1.50–1.61), currently homeless versus at risk of homelessness among veterans (Byrne et al., 2019) HR = 1.20 (0.60–2.30) (Hahn et al., 2002) (cohort)
Homeless	Overdose	↑	OR = 1.64 (1.49–1.81) (Dunne et al., 2025) (meta-analysis) aPR = 1.64 (1.49–1.79) (Marcus et al., 2022) aOR = 1.36 (0.70–2.63) (Gunaratne et al., 2026) aOR = 1.29 (0.82–2.04) (Guarino et al., 2023) Positive association reported (Arum et al., 2021; Austin et al., 2021) (systematic reviews)
Homeless	Injection partners	↑	aRR = 1.48 (1.25–1.76) (Hotton et al., 2021)
Homeless	Partners with HCV	↔	No direct estimate for this variable pair; contextual evidence from Arum et al. (2021) and Degenhardt et al. (2023) (systematic reviews)
Homeless	Treatment	↔	Lower initiation and adherence to opioid substitution treatment reported (Arum et al., 2021) (systematic review) Mixed evidence for receipt and completion of substance-use treatment (Austin et al., 2021) (systematic review)
Homeless	Non-injection routes	↓	aPR = 0.84 (0.83–0.85) for preferred non-injection route (Karandinos et al., 2024) Indirect evidence on homelessness and transition toward injecting (Rahimi-Movaghar et al., 2015) (systematic review)

Note. ↔ indicates no fixed directional benchmark; these pairs are excluded from directional scoring. Intervals accompanying estimates are 95% confidence intervals. OR = odds ratio; RR = risk ratio; HR = hazard ratio; PR = prevalence ratio. The prefix 'a' denotes adjustment; aRR in Hotton et al. denotes an adjusted rate ratio. Estimates refer to the source studies' own definitions and comparisons.

For profile i and outcome k , the LLM-generated housing condition difference was defined as

$$\Delta_{ik} = Y_{ik}^{*(homeless)} - Y_{ik}^{*(stable)}. \quad (2)$$

To enable comparison across variables measured on different scales, the within-profile differences defined in Equation 2 were summarized using the paired standardized mean difference d_z , calculated as the mean paired difference divided by the standard deviation of the paired differences. We refer to this as paired Cohen's d_z .

Because LLMs may generate invariant outcomes for which effects or correlations cannot be estimated due to insufficient variation in the paired outputs, estimability was incorporated into the evaluation framework. An outcome was considered invariant when it had no variation across profiles, and correlations involving an invariant variable were considered non-estimable. Housing effects were considered non-estimable when the paired differences had zero standard deviation.

For model m and literature-informed target g , let E_{mg} indicate whether the corresponding generated association is estimable and let A_{mg} indicate whether its estimated direction matches the literature-informed expectation. For G targets with prespecified directional expectations,

$$C_m = \frac{\sum_{g=1}^G E_{mg} A_{mg}}{\sum_{g=1}^G E_{mg}}, \quad U_m = \frac{1}{G} \sum_{g=1}^G E_{mg} A_{mg}. \quad (3)$$

In Equation 3, C_m measures directional recovery conditional on estimability, whereas U_m measures directional recovery across all prespecified targets, retaining non-estimable targets in the denominator. Mixed directional findings in the literature were not classified as directional successes or failures.

b. Similarity to ABM correlations

Our second external benchmark was the same ABM used to construct the LLM prompt profiles. The dataset contained longitudinal records of agents at different timepoints. The first timepoint was derived from PWID survey data, whereas later timepoints were simulated independently of the LLM. Because housing status and fentanyl-supply exposure are not part of the ABM, the ABM was used only to compare relationships among outcome variables rather than to assess whether LLMs reproduced the ABM’s agent-level outcome values.

Comparisons were restricted to variables available in both the LLM and ABM outputs: weekly injection frequency, receptive syringe sharing, network in-degree, network out-degree, and partners with HCV. Because the LLM prompt specified a 6-month scenario, each agent’s available ABM follow-up record closest to six months after activation was selected. No maximum deviation from this target was imposed; selected follow-up times had a median of 3.51 years after activation (IQR 1.59–6.19 years).

For each LLM, Spearman correlations were calculated across pooled stable- and homeless-condition outputs, with each profile’s ABM snapshot aligned with both housing-condition rows. Differences between corresponding LLM and ABM correlations were summarized using mean absolute difference (MAD), with lower values indicating closer agreement on the correlation scale. Supplementary Figure S2 presents condition-specific correlations. We also assessed sensitivity to progressively narrower follow-up windows around the six-month target (Supplementary Figure S3).

c. Similarity to empirical survey associations

As the third external benchmark, we evaluated whether LLMs reproduced HCV-related associations observed in the 2004 New York City Health and Nutrition Examination Survey (New York City Department of Health and Mental Hygiene and National Center for Health Statistics, 2011). LLMs were provided with individual-level profiles from this survey, which included variables indicating injection drug use and street drug use but did not include housing status. Profile characteristics were selected based on their availability among a small number of HCV antibody-positive participants in the survey. The sample included all 35 HCV antibody-positive participants and a random sample of 35 HCV antibody-negative participants meeting the required data-completeness criteria for profile construction. Missing variables for one positive participant was presented as ‘Not reported.’ Each profile was constructed using available demographic, socioeconomic, behavioral, and health characteristics, excluding HCV status. For each profile, LLMs generated an HCV risk score from 0 to 100.

To assess recovery of empirical associations, we compared (1) Spearman correlations between individual characteristics and HCV antibody status with (2) correlations between those same individual characteristics and LLM-generated HCV risk scores. Differences between empirical and LLM correlations were summarized using MAD (Supplementary Section III). Survey MAD compares covariate associations with a continuous score and a binary antibody result; it is not a calibration measure. LLM-generated HCV risk scores were also evaluated for their ability to distinguish participants with positive versus negative HCV antibody results. Discriminatory performance was assessed using the area under the receiver operating characteristic curve (AUC), treating the generated risk score as a continuous predictor of HCV antibody status.

d. Bootstrap uncertainty

Across the three benchmarks, we used 1,000 nonparametric bootstrap resamples to quantify uncertainty for selected measures. Stable and homeless predictions were kept paired within each individual when estimating housing effects. In the ABM comparison, agents were resampled with both housing-condition outputs and their corresponding ABM outcomes kept together. For the empirical HCV benchmark, individuals were resampled separately within the HCV antibody-positive and HCV antibody-negative groups. Percentile 95% bootstrap confidence intervals were calculated

for d_z , ABM MAD, empirical MAD, and AUC when the corresponding statistic was estimable. These intervals quantify uncertainty from resampling the evaluated profiles conditional on the generated LLM outputs and benchmark data.

3. Knowledge probes

Knowledge probes prompted the LLMs to directly identify the expected direction of association between predefined variable pairs. These outputs were compared with the corresponding literature-informed expectations (Table 2 and Supplementary Table S1). Knowledge probes are a secondary diagnostic, used to assess whether an LLM can identify an epidemiological relationship, whereas behavioral generation assesses whether individual-level outputs from a prompt express that relationship.

4. Sensitivity analyses

LLM-generated responses may vary with model choice and prompt wording. We therefore evaluated two potential sources of variation: (1) sensitivity to model choice, which was broadly assessed by comparing results across LLMs, and (2) sensitivity to prompt wording, which compared the original prompt with a semantically equivalent paraphrase—a prompt with the same meaning but different wording. For prompt sensitivity, we compared both (i) the direction of housing-condition effects between individual-level prompts and (ii) the responses to the knowledge probe prompts.

III. Results

The homeless condition consistently corresponded with higher levels of public injecting in LLM-generated outputs ($d_z = 1.48$ – 5.90), aligning with literature-informed expectations. Other housing effects were smaller ($|d_z| = 0.01$ – 0.65) and varied in direction. Across the 28 retained associations with literature-informed directions, conditional directional recovery was 81.8% for Claude, 76.5% for DeepSeek, and 67.9% for GPT-OSS. Unconditional directional recovery, which included non-estimable targets in the denominator, was 64.3%, 46.4%, and 67.9%, respectively (Figure 2; Supplementary Figure S1).

Variable pairs	Expected Direction	LLM Knowledge Probe			LLM-Generated Outputs		
		Claude Sonnet 4.6	DeepSeek v3.2	GPT-OSS-120B	Claude Sonnet 4.6	DeepSeek v3.2	GPT-OSS-120B
		direction + LLM-rated evidence strength + LLM-derived association estimate			Cohen's d_z (standardized mean differences) + 95% bootstrap CI		
Homelessness ~ public injecting	↑	↑ strong aOR 2.87 (0.59–3.94)	↑ strong	↑ moderate	↑ 5.90 (5.39–6.59)	↑ 1.48 (1.29–1.71)	↑ 2.27 (2.05–2.50)
Homelessness ~ injection frequency	↑	↑ moderate aOR 1.57 (1.12–2.20)	↑ moderate	↑ moderate	↑ 0.03	↓ -0.07	↓ -0.41 (-0.53–0.28)
Homelessness ~ receptive sharing	↑	↑ moderate aOR 1.72 (1.23–2.40)	↑ moderate	↑ moderate	↑ 0.37 (0.30–0.44)	↑ 0.01 (-0.12–0.14)	↓ -0.13 (-0.25–0.00)
Homelessness ~ positive HCV status	↑	↑ moderate	↑ strong aOR 2.40 (1.85–3.20)	↑ moderate	↑ 0.25 (0.18–0.32)	↑ 0.17 (0.05–0.27)	↓ -0.06 (-0.21–0.05)
Homelessness ~ non-fatal overdose	↑	↑ moderate aOR 1.86 (1.19–2.90)	↑ moderate	↑ moderate	↑ 0.65 (0.51–0.79)	—	↑ 0.19 (0.06–0.33)
Homelessness ~ injection partners	↑	↑ moderate	↑ moderate	↑ moderate	↑ 0.04 (-0.06–0.53)	↑ 0.08 (-0.03–0.15)	↑ 0.01 (-0.12–0.16)
Homelessness ~ partners with HCV	↔	↑ weak	↑ moderate	↑ moderate	↑ 0.07	—	↓ -0.10 (-0.23–0.03)
Homelessness ~ treatment	↔	↓ moderate	↓ moderate	↑ moderate	—	—	↓ -0.57 (-0.72–0.45)
Homelessness ~ non-injection routes	↓	↓ weak	↑ moderate	↑ moderate	↓ -0.24 (-0.36–0.13)	↑ 0.15 (0.02–0.28)	↓ -0.05 (-0.18–0.08)

Matches expected direction
Opposite of expected direction
No variation between conditions
Expected direction is mixed

Figure 2: Comparison of literature-informed association directions with LLM knowledge-probe and generated-output directions.

The MAD between ABM and LLM correlation matrices was 0.11 for Claude, 0.13 for DeepSeek, and 0.12 for GPT-OSS. MADs between empirical HCV antibody associations in the survey data and the corresponding associations from LLM-generated outputs were 0.17, 0.15, and 0.18, respectively. LLM-generated HCV risk scores also distinguished HCV-antibody-positive from HCV-antibody-negative survey participants, with AUCs of 0.85 for Claude, 0.83 for DeepSeek, and 0.78 for GPT-OSS. The 95% confidence intervals for these metrics are reported in Table 3.

Table 3: Evaluation metrics used across benchmarks

Metric	Claude Sonnet 4.6	DeepSeek v3.2	GPT-OSS-120B
<i>LLM outputs vs. literature</i>			
Estimable targets	22/28 (78.6%)	17/28 (60.7%)	28/28 (100.0%)
Conditional directional recovery	18/22 (81.8%)	13/17 (76.5%)	19/28 (67.9%)
Unconditional directional recovery	18/28 (64.3%)	13/28 (46.4%)	19/28 (67.9%)

Metric	Claude Sonnet 4.6	DeepSeek v3.2	GPT-OSS-120B
Knowledge-probe agreement	26/28 (92.9%)	25/28 (89.3%)	23/28 (82.1%)
<i>LLM outputs vs. ABM data</i>			
Correlation MAD	0.11 (0.08–0.14)	0.13 (0.10–0.17)	0.12 (0.10–0.16)
<i>LLM outputs vs. survey data</i>			
Correlation MAD	0.17 (0.11–0.27)	0.15 (0.10–0.26)	0.18 (0.11–0.28)
HCV-score AUC	0.85 (0.76–0.93)	0.83 (0.73–0.91)	0.78 (0.67–0.88)
<i>Prompt sensitivity</i>			
Housing-direction agreement	1,670/2,016 (82.8%)	1,881/2,016 (93.3%)	1,155/2,016 (57.3%)
Knowledge-probe stability	31/32 (96.9%)	31/32 (96.9%)	28/32 (87.5%)

Direct knowledge-probe agreement with the literature was 26/28 (92.9%) for Claude, 25/28 (89.3%) for DeepSeek, and 23/28 (82.1%) for GPT-OSS, whereas unconditional recovery in LLM-generated behavioral outputs ranged from 46.4% to 67.9%. The gap between knowledge-probe agreement and unconditional recovery was 28.6 percentage points for Claude, 42.9 for DeepSeek, and 14.3 for GPT-OSS.

Sensitivity to prompt wording was assessed by comparing original prompts with alternative phrasings. Across 2,016 profile–outcome pairs per model (224 profiles $\times$ 9 numeric outcomes), the direction of the housing difference was unchanged between prompt versions for 82.8% of profile–outcome pairs for Claude, 93.3% for DeepSeek, and 57.3% for GPT-OSS. Across the 32 originally queried variable pairs, Claude and DeepSeek assigned the same association classification to 31 pairs (96.9%) across prompt versions, and GPT-OSS did so for 28 pairs (87.5%) (Supplementary Table S2).

IV. Discussion

Previous studies have used LLMs to generate individual-level behaviors in agent-based models, including in epidemiological applications (Williams et al., 2023; Lu et al., 2024; Adornetto et al., 2025). Such applications highlight the need to evaluate whether generated behavioral relationships align with multiple sources of external evidence before LLM outputs are used to complement empirical data. Although these generative models cannot fully overcome persistent challenges related to missing data and underrepresented populations, they may be particularly useful for hard-to-study populations, such as people experiencing homelessness, and in changing contexts where empirical data may be limited, such as a rapidly evolving drug landscape. If LLM-generated outputs can reliably reproduce epidemiologically important behavioral relationships, they could potentially help explore questions in understudied populations and contexts where empirical data are limited (Chen et al., 2024; Von Hoene et al., 2025).

Overall, our evaluation finds that LLM-generated relationships reproduced the literature-informed direction for 46–68% of prespecified targets, with differences from ABM and empirical correlation patterns summarized by MADs of 0.11–0.13 and 0.15–0.18, respectively. While these descriptive comparisons do not establish that generated data represent the behavior of a target population, they show partial correspondence with relationships observed across multiple external benchmarks. The three LLMs also differed in the types of limitations observed. Claude and DeepSeek generated the same values for certain behavioral outcomes across all 224 sampled agents, leaving no variation with which to evaluate relationships with other variables, whereas GPT-OSS was most sensitive to prompt wording. Similar variation across LLMs has been reported previously (Anderson et al., 2026), and as new models continue to be developed and released, standardized criteria for evaluating their suitability for particular applications remain underdeveloped (Lu et al., 2024; Adornetto et al., 2025).

In the separate knowledge-probe analysis, LLMs correctly identified most prespecified literature-informed association directions and showed high stability to prompt rewording, whereas generated behavioral outputs were less stable across semantically equivalent prompts. These comparisons suggest that (1) an LLM’s ability to correctly identify known epidemiological relationships does not necessarily translate into reproducing them in individual-level behavioral outputs and (2) LLMs were generally less sensitive to prompt wording when directly identifying epidemiological relationships than when generating behavioral relationships.

This distinction is particularly important because existing LLM evaluations have largely emphasized tasks such as factual knowledge, question answering, reasoning, and other isolated task-level benchmarks (Chang et al., 2024). However, performance on these benchmarks does not establish whether LLM-generated individual-level behavioral outputs reproduce relationships important for epidemiological modeling. Such evaluation is important for infectious disease transmission modeling, where individual behaviors and interactions can directly influence downstream transmission

and epidemiological outcomes in simulations (Abuelezzam et al., 2013; Gutfraind et al., 2015; Nguyen, Megiddo and Howick, 2020). Using multiple benchmarks may therefore provide a more informative assessment than requiring exact correspondence with a single reference source, particularly given variability in LLM outputs and sampling variability in the reference data (Chen et al., 2024).

Future work could establish whether knowledge probes predict behavioral-generation performance across a broader range of models and prompts. Another direction would be to test the incorporation of LLM-generated behaviors as predefined inputs in agent-based epidemiological models, rather than generating behaviors dynamically during simulations. This approach could make generated behaviors easier to trace and allow researchers to assess how specific inputs affect downstream epidemiological outcomes while reducing opportunities for error propagation. Additionally, more extensive prompt-perturbation experiments could determine whether sensitivity to prompt wording is consistent across epidemiological conditions or varies by characteristic or domain. Although previous studies have shown that LLM outputs are sensitive to prompt wording (Chang et al., 2024), variation across models and prompts could potentially be used to characterize uncertainty when multiple model-prompt combinations produce epidemiologically plausible outputs.

Together, these findings illustrate the value of evaluating LLM-generated behavioral relationships against multiple external benchmarks, revealing partial alignment with epidemiological evidence as well as limitations in behavioral variation and sensitivity to prompt wording that may not be apparent from knowledge-based evaluations alone. Such limitations may also be overlooked when GABMs are evaluated solely through aggregate simulation outcomes (Adornetto et al., 2025). It is therefore important that evaluations of LLM-based epidemiological models consider not only downstream simulation outcomes but also the generated behaviors that contribute to them, particularly when those behavioral outputs are used to simulate transmission dynamics.

V. Data Availability Statement

The data, LLM-generated outputs, and code used in this study are publicly available at https://github.com/mayonais/hcv_llm.

VI. Declaration of Generative AI

Claude (Anthropic), ChatGPT (OpenAI), and Codex (OpenAI) were used in an assistive capacity to develop code for analyses and figures and to suggest revisions to the manuscript. The authors made the final methodological, interpretive, and editorial decisions and take full responsibility for the work.

VII. Declarations

The work was supported in part by National Institutes of Health [grant number R01GM121600]. In addition to his affiliation with the University of Illinois, A.G. is employed by Amazon.com but his employer did not have any part in this study

References

- Abuelezzam, N.N., Rough, K., Seage III, G.R., 2013. Individual-Based Simulation Models of HIV Transmission: Reporting Quality and Recommendations. *PLOS ONE* 8, e75624. <https://doi.org/10.1371/journal.pone.0075624>
- Adornetto, C., Mora, A., Hu, K., Garcia, L.I., Atchade-Adelomou, P., Greco, G., Pastor, L.A.A., Larson, K., 2025. Generative Agents in Agent-Based Modeling: Overview, Validation, and Emerging Challenges. *IEEE Transactions on Artificial Intelligence* 6, 3165–3183. <https://doi.org/10.1109/TAI.2025.3566362>
- Anderson, T., Von Hoene, S., Cinar, O.Y., Von Hoene, E., Roess, A., Crooks, A., Kavak, H., 2026. Using Zero-Shot LLM-Generated Survey Data for Geographically Explicit Population Synthesis [WWW Document]. *arXiv.org*. URL <https://arxiv.org/abs/2605.27401v1> (accessed 7.10.26).
- Arum, C., Fraser, H., Artenie, A.A., Bivegete, S., Trickey, A., Alary, M., Astemborski, J., Iversen, J., Lim, A.G., MacGregor, L., Morris, M., Ong, J.J., Platt, L., Sack-Davis, R., van Santen, D.K., Solomon, S.S., Sypsa, V., Valencia, J., van den Boom, W., Walker, J.G., Ward, Z., Stone, J., Vickerman, P., Homelessness, HIV, and HCV Review Collaborative Group, 2021. Homelessness, unstable housing, and risk of HIV and hepatitis C virus acquisition among people who inject drugs: a systematic review and meta-analysis. *The Lancet Public Health* 6, e309–e323. [https://doi.org/10.1016/S2468-2667\(21\)00013-X](https://doi.org/10.1016/S2468-2667(21)00013-X)

- Austin, A.E., Shiue, K.Y., Naumann, R.B., Figgatt, M.C., Gest, C., Shanahan, M.E., 2021. Associations of housing stress with later substance use outcomes: A systematic review. *Addict Behav* 123, 107076. <https://doi.org/10.1016/j.addbeh.2021.107076>
- Ballard, A.M., Kesich, Z., Crane, H.M., Feinberg, J., Friedmann, P.D., Go, V.F., Jenkins, W.D., Korthuis, P.T., Miller, W.C., Pho, M.T., Seal, D.W., Smith, G.S., Stopka, T.J., Westergaard, R.P., Zule, W.A., Young, A.M., Cooper, H.L.F., 2025. Rural houselessness among people who use drugs in the United States: Results from the National Rural Opioid Initiative. *Drug Alcohol Depend* 266, 112498. <https://doi.org/10.1016/j.drugalcdep.2024.112498>
- Byrne, T., Troszak, L., Midboe, A.M., Fincke, B.G., Schwartz, M., Gifford, A.L., McInnes, D.K., 2019. A Novel Measure to Assess Variation in Hepatitis C Prevalence Among Homeless and Unstably Housed Veterans, 2011-2016. *Public Health Rep* 134, 126–131. <https://doi.org/10.1177/0033354918821071>
- Chang, Yupeng, Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Yi, Yu, P.S., Yang, Q., Xie, X., 2024. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* 15, 39:1-39:45. <https://doi.org/10.1145/3641289>
- Chen, J., Bhattacharya, P., Hoops, S., Machi, D., Adiga, A., Mortveit, H., Venkatramanan, S., Lewis, B., Marathe, M., 2024. Role of heterogeneity: National scale data-driven agent-based modeling for the US COVID-19 Scenario Modeling Hub. *Epidemics* 48, 100779. <https://doi.org/10.1016/j.epidem.2024.100779>
- Degenhardt, L., Webb, P., Colledge-Frisby, S., Ireland, J., Wheeler, A., Ottaviano, S., Willing, A., Kairouz, A., Cunningham, E.B., Hajarizadeh, B., Leung, J., Tran, L.T., Price, O., Peacock, A., Vickerman, P., Farrell, M., Dore, G.J., Hickman, M., Grebely, J., 2023. Epidemiology of injecting drug use, prevalence of injecting-related harm, and exposure to behavioural and environmental risks among people who inject drugs: a systematic review. *The Lancet Global Health* 11, e659–e672. [https://doi.org/10.1016/S2214-109X\(23\)00057-8](https://doi.org/10.1016/S2214-109X(23)00057-8)
- Dunne, N., Casey, M.-B., Ivers, J.-H., 2025. What are the risk factors of non-fatal overdose among persons who use opioids? A systematic review and meta-analysis. *Journal of Substance Use and Addiction Treatment* 171, 209630. <https://doi.org/10.1016/j.josat.2025.209630>
- Eames, K., Bansal, S., Frost, S., Riley, S., 2015. Six challenges in measuring contact networks for use in modelling. *Epidemics* 10, 72–77. <https://doi.org/10.1016/j.epidem.2014.08.006>
- Gahrton, C., Navér, G., Warnqvist, A., Dalgard, O., Aleman, S., Kåberg, M., 2024. Changes in hepatitis C virus prevalence and incidence among people who inject drugs in the direct acting antiviral era. *International Journal of Drug Policy* 128, 104433. <https://doi.org/10.1016/j.drugpo.2024.104433>
- Guarino, H., Frank, D., Quinn, K., Kim, D., Gile, K., Ruggles, K., Friedman, S.R., Mateu-Gelabert, P., 2023. Syndemic factors associated with non-fatal overdose among young opioid users in New York City. *Front. Public Health* 11, 1195657. <https://doi.org/10.3389/fpubh.2023.1195657>
- Gunaratne, M.P., Loeb, T.A., McFall, A.M., Ganapathi, L., Baishya, J.J., Kedar, A., Sinha, A.K., Srikrishnan, A.K., Solomon, S.S., Lucas, G.M., Mehta, S.H., 2026. Characterizing non-fatal overdose among people who inject drugs in India: prevalence and risk factor analysis, 2022-2024. *International Journal of Drug Policy* 147, 105073. <https://doi.org/10.1016/j.drugpo.2025.105073>
- Gutfraind, A., Boodram, B., Prachand, N., Hailegiorgis, A., Dahari, H., Major, M.E., 2015. Agent-Based Model Forecasts Aging of the Population of People Who Inject Drugs in Metropolitan Chicago and Changing Prevalence of Hepatitis C Infections. *PLOS ONE* 10, e0137993. <https://doi.org/10.1371/journal.pone.0137993>
- Hahn, J.A., Page-Shafer, K., Lum, P.J., Bourgois, P., Stein, E., Evans, J.L., Busch, M.P., Tobler, L.H., Phelps, B., Moss, A.R., 2002. Hepatitis C virus seroconversion among young injection drug users: relationships and risks. *J Infect Dis* 186, 1558–1564. <https://doi.org/10.1086/345554>
- Hotton, A., Mackesy-Amity, M.-E., Boodram, B., 2021. Trends in homelessness and injection practices among young urban and suburban people who inject drugs: 1997-2017. *Drug and Alcohol Dependence* 225, 108797. <https://doi.org/10.1016/j.drugalcdep.2021.108797>
- Kan, M., Holmén, E., Brynte, C., Ekström, A.M., Kåberg, M., 2026. Needle and syringe coverage per injection and factors associated with needle sharing among clients of the Stockholm Needle and Syringe Program, 2013–2023. *International Journal of Drug Policy* 153, 105299. <https://doi.org/10.1016/j.drugpo.2026.105299>
- Karandinos, G., Unick, J., Ondocsin, J., Holm, N., Mars, S., Montero, F., Rosenblum, D., Ciccarone, D., 2024. Decrease in injection and rise in smoking and snorting of heroin and synthetic opioids, 2000-2021. *Drug Alcohol Depend* 263, 111419. <https://doi.org/10.1016/j.drugalcdep.2024.111419>
- Khezri, M., Kimball, S., McKnight, C., Rouhani, S., Bunting, A.M., Karamouzian, M., Ompad, D.C., Des Jarlais, D., 2025. Harms associated with injecting in public spaces: a global systematic review and meta-analysis. *Int J Drug Policy* 140, 104819. <https://doi.org/10.1016/j.drugpo.2025.104819>
- Lu, Y., Aleta, A., Du, C., Shi, L., Moreno, Y., 2024. LLMs and generative agent-based models for complex systems research. *Physics of Life Reviews* 51, 283–293. <https://doi.org/10.1016/j.plprev.2024.10.013>
- Marcus, R., Cha, S., Sionean, C., Kanny, D., for the National HIV Behavioral Surveillance Study Group, 2022. HIV Injection Risk Behaviors among HIV-Negative People Who Inject Drugs Experiencing Homelessness, 23 U.S. Cities. *J Soc Distress Homeless* 31, 96–104. <https://doi.org/10.1080/10530789.2021.1892931>

- Morris, M.D., Evans, J., Montgomery, M., Yu, M., Briceno, A., Page, K., Hahn, J.A., 2014. Intimate Injection Partnerships Are at Elevated Risk of High-Risk Injecting: A Multi-Level Longitudinal Study of HCV-Serodiscordant Injection Partnerships in San Francisco, CA. *PLOS ONE* 9, e109282. <https://doi.org/10.1371/journal.pone.0109282>
- Morris, M.D., Yen, I.H., Shiboski, S., Evans, J.L., Page, K., 2020. Housing Stability and Hepatitis C Infection for Young Adults Who Inject Drugs: Examining the Relationship of Consistent and Intermittent Housing Status on HCV Infection Risk. *J Urban Health* 97, 831–844. <https://doi.org/10.1007/s11524-020-00445-7>
- New York City Department of Health and Mental Hygiene and United States Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics (2011). *New York City Health and Nutrition Examination Survey (NYC HANES), 2004*. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. <https://doi.org/10.3886/ICPSR31421.v1>.
- Nguyen, L.K.N., Megiddo, I., Howick, S., 2020. Simulation models for transmission of health care-associated infection: A systematic review. *American Journal of Infection Control* 48, 810–821. <https://doi.org/10.1016/j.ajic.2019.11.005>
- Niedzielewski, K., Bartzuk, R.P., Bielczyk, N., Bogucki, D., Dreger, F., Dudziuk, G., Górski, Ł., Gruziel-Słomka, M., Haman, J., Kaczorek, A., Kisielewski, J., Krupa, B., Moszyński, A., Nowosielski, J.M., Radwan, M., Semeniuk, M., Tymoszek, U., Zieliński, J., Rakowski, F., 2024. Forecasting SARS-CoV-2 epidemic dynamic in Poland with the pDyn agent-based model. *Epidemics* 49, 100801. <https://doi.org/10.1016/j.epidem.2024.100801>
- Rahimi-Movaghar, A., Amin-Esmaili, M., Shadloo, B., Noroozi, A., Malekinejad, M., 2015. Transition to injecting drug use in Iran: A systematic review of qualitative and quantitative evidence. *International Journal of Drug Policy* 26, 808–819. <https://doi.org/10.1016/j.drugpo.2015.04.018>
- Thi Phuong, H., Suer, J., Belik, V., Rincón Hidalgo, A., Jarynowski, A.K., Pastor, R., Schulz, S., Thampi, A., Xu, C., Zambrano, M., Mikolajczyk, R., Karch, A., Jaeger, V.K., on behalf of the OptimAgent Consortium, 2026. Social contact patterns derived from an epidemiological survey and GPS-based co-location data – A systematic comparison using parallel data collections during the COVID-19 pandemic in Germany. *Epidemics* 54, 100886. <https://doi.org/10.1016/j.epidem.2026.100886>
- Von Hoene, E., Roess, A., Kavak, H., Anderson, T., 2025. Synthetic population generation with public health characteristics for spatial agent-based models. *PLOS Computational Biology* 21, e1012439. <https://doi.org/10.1371/journal.pcbi.1012439>
- Williams, R., Hosseinichimeh, N., Majumdar, A., Ghaffarzadegan, N., 2023. Epidemic Modeling with Generative Agents. <https://doi.org/10.48550/arXiv.2307.04986>

Supplementary Information

I. Reproduction of association directions from the literature

a Comparing literature-informed directions with LLM-generated correlations

In addition to the 9 housing-related associations (Table 2), we identified 23 literature-informed association directions between outcome variables and compared them to correlations among large language model (LLM)-generated outcome variables. This produced a total of 32 expected directions. Four had mixed or inconclusive prespecified directions and were thus excluded from directional-recovery calculations.

Table S1: Prespecified literature-informed directions for associations between variable pairs

Variable pairs		Direction	Evidence
Public injecting	Positive HCV status	↑	OR = 1.55 (1.18–2.02) (Khezri et al., 2025)(meta-analysis) aOR = 1.91 (1.33–2.75) (Eckhardt et al., 2017)
Public injecting	Overdose	↑	OR = 2.51 (2.01–3.13) (Khezri et al., 2025) (meta-analysis)
Public injecting	Receptive sharing	↑	OR = 2.41 (1.97–2.94) (Khezri et al., 2025)(meta-analysis)
Public injecting	Injection frequency	↑	OR range across studies = 1.96–5.34 (Khezri et al., 2025) (systematic review)
Public injecting	Treatment	↓	OR = 0.61 (0.48–0.76) for methadone maintenance treatment (longitudinal study reported in Khezri et al., 2025)
Receptive sharing	Positive HCV status	↑	HR = 2.1 (1.2–3.7) (Hahn et al., 2002) (cohort) PRR = 1.94 (1.53–2.46) (Pouget et al., 2012) (meta-analysis) aHR = 1.79 (1.44–2.22) (Gahrton et al., 2024) (cohort) Indirect evidence: aRR = 1.53 (1.01–2.31) for syringe sharing, without distinguishing receptive from distributive sharing (Mackesy-Amiti et al., 2023; baseline analysis) aPR = 1.20 (1.09–1.33) (Kan et al., 2026) (cohort)
Receptive sharing	Overdose	↑	Indirect evidence: aOR = 2.65 (1.43–4.94) for sharing syringes and/or other injection paraphernalia (Jenkins et al., 2011) aOR = 1.70 (1.03–2.81) (Guarino et al., 2023)
Treatment	Overdose	↓	OR = 0.72 (0.64–0.81) (Dunne et al., 2025) (meta-analysis)
Treatment	Positive HCV status	↓	aHR = 0.74 (0.56–0.97) (Gahrton et al., 2024) (cohort)
Treatment	Receptive sharing	↓	aOR = 0.046 (0.034–0.061) (Walker et al., 2025) Negative association reported (Karlsson et al., 2021) (cohort)
Treatment	Non-injection route	↓	Indirect evidence: annual decline in any non-injection opioid use after OAT initiation; aOR = 0.89 (0.85–0.92) per year (Dong et al., 2020) (cohort)
Treatment	Injection frequency	↓	aRR = 0.21 (0.12–0.36) (Walker et al., 2025)
Injection frequency	Positive HCV status	↑	HR = 2.7 (1.0–7.6) (Hahn et al., 2002) (cohort) aOR = 1.26 (1.09–1.45) (Eckhardt et al., 2017)
Injection frequency	Overdose	↑	aOR = 1.55 (1.00–2.40) (Gunaratne et al., 2026) OR = 1.19 (0.73–1.93) (Des Jarlais et al., 2023) aOR = 1.05 (1.02–1.08) (Rafful et al., 2018) (cohort)
Injection frequency	Receptive sharing	↑	aPR = 1.39 (1.22–1.58) (Kan et al., 2026) (cohort) OR = 1.02 (0.62–1.69) (Hahn et al., 2010) Positive association reported (Tran et al., 2020) (meta-analysis)
Injection frequency	Non-injection routes	↓	Negative association reported (Rahimi-Movaghar et al., 2015) and (Kral et al., 2021) (cohort)
Non-injection routes	Positive HCV status	↓	aOR = 0.72 (0.52–0.99) (Des Jarlais et al., 2018)
Non-injection routes	Overdose	↔	OR = 0.63 (0.14–1.95) (Des Jarlais et al., 2023)
Partners	Receptive sharing	↓	OR = 0.60 (0.28–1.30) (Hahn et al., 2010) RR = 0.59 (0.25–1.43) (Morris et al., 2014) (cohort) PR = 0.49 (0.34–0.70) (Rudolph and Rhodes, 2023)
Partners	Injection frequency	↑	aHR = 1.41 (1.21–1.64) (Spelman et al., 2019) (cohort)
Partners	Positive HCV status	↑	aHR = 1.05 (1.02–1.13) (Spelman et al., 2019) (cohort)
Partners with HCV	Receptive sharing	↔	PR = 1.02 (0.52–1.97) (Rudolph and Rhodes, 2023) OR = 0.49 (0.25–0.95) (Hahn et al., 2010)

Positive HCV status Overdose ↑ aOR = 3.24 (2.60–4.03) (Gunaratne et al., 2026)
 OR = 2.69 (1.98–3.66) (Eckhardt et al., 2017)
 aOR = 2.64 (1.47–4.74) (Guarino et al., 2023)
 aOR = 2.44 (1.20–4.97) for chronic HCV infection (Schiavon et al., 2018)

Note. Abbreviations: OR = odds ratio; RR = risk ratio; HR = hazard ratio; PR = prevalence ratio; IRR = incidence rate ratio; aOR, aRR, aPR, aHR = adjusted measures. For Walker et al. (2025), aRR denotes an adjusted rate ratio.

Variable pairs	Expected Direction	LLM Knowledge Probe			LLM-Generated Outputs		
		Claude Sonnet 4.6	DeepSeek v3.2	GPT-OSS-120B	Claude Sonnet 4.6	DeepSeek v3.2	GPT-OSS-120B
		direction + LLM-rated evidence strength + LLM-derived association estimate			Spearman's rank correlation		
Public injecting ~ positive HCV status	↑	↑ moderate aOR 1.74 (1.13–2.68)	↑ moderate	↑ strong	↑ 0.13	↑ 0.18	↑ 0.07
Public injecting ~ non-fatal overdose	↑	↑ moderate aOR 1.67 (1.22–2.29)	↑ moderate	↑ moderate	↑ 0.26	—	↑ 0.12
Public injecting ~ receptive sharing	↑	↑ moderate aOR 1.84 (1.32–2.57)	↑ moderate	↑ moderate	↑ 0.11	↑ 0.24	↑ 0.06
Public injecting ~ injection frequency	↑	↑ moderate aOR 1.98 (1.23–2.93)	↑ moderate	↑ moderate	↓ -0.04	↑ 0.11	↓ -0.02
Public injecting ~ treatment	↓	↑ moderate aOR 0.58 (0.40–0.85)	↓ moderate	↓ moderate	—	—	↓ -0.39
Receptive sharing ~ positive HCV status	↑	↑ strong aOR 1.79 (1.38–2.34)	↑ strong	↑ strong aOR 2.00 (1.50–2.60)	↑ 0.21	↑ 0.06	↑ 0.27
Receptive sharing ~ non-fatal overdose	↑	↑ moderate	↑ moderate aOR 1.46 (1.20–1.78)	↔ weak	↑ 0.38	—	↑ 0.16
Treatment ~ non-fatal overdose	↓	↓ moderate	↓ strong aOR 2.47 (0.33–0.87)	↓ moderate	—	—	↓ -0.11
Treatment ~ positive HCV status	↓	↓ moderate aOR 0.40 (0.26–0.62)	↓ strong aOR 0.50 (0.40–0.63)	↑ strong RR 0.58 (0.44–0.71)	—	—	↑ 0.00
Treatment ~ receptive sharing	↓	↓ moderate	↓ strong	↓ strong	—	—	↓ -0.22
Treatment ~ non-injection routes	↓	↑ moderate	↑ moderate	↔ moderate	—	—	↑ 0.08
Treatment ~ injection frequency	↓	↓ strong aOR 2.44 (0.33–0.89)	↓ strong	↓ strong	—	—	↓ -0.02
Injection frequency ~ positive HCV status	↑	↑ strong aOR 1.77 (1.45–2.16)	↑ strong	↑ strong	↑ 0.13	↑ 0.11	↑ 0.18
Injection frequency ~ non-fatal overdose	↑	↑ moderate aOR 1.00 (1.01–1.05)	↑ moderate	↑ strong	↑ 0.65	—	↑ 0.42
Injection frequency ~ receptive sharing	↑	↑ moderate	↑ strong aOR 1.02 (1.01–1.03)	↑ strong	↑ 0.10	↑ 0.21	↑ 0.07
Injection frequency ~ non-injection routes	↓	↓ moderate	↓ moderate	↓ moderate	↓ 0.85	↓ -0.07	↑ 0.45
Non-injection routes ~ positive HCV status	↓	↓ moderate aOR 0.42 (0.23–0.77)	↓ moderate aOR 0.40 (0.20–0.80)	↓ moderate	↑ 0.05	↑ 0.11	↑ 0.10
Non-injection routes ~ non-fatal overdose	↔	↑ moderate aOR 0.50 (0.36–0.75)	↑ moderate aOR 1.38 (1.10–1.73)	↑ moderate	↑ 0.62	—	↑ 0.16
Injection partners ~ receptive sharing	↓	↑ strong aOR 1.47 (1.28–1.69)	↑ moderate	↑ strong	↑ 0.49	↑ 0.56	↑ 0.43
Injection partners ~ injection frequency	↑	↑ moderate	↑ moderate	↑ moderate	↑ 0.30	↑ 0.28	↑ 0.35
Injection partners ~ positive HCV status	↑	↑ moderate	↑ moderate	↑ strong	↑ 0.17	↑ 0.20	↑ 0.49
Partners with HCV ~ receptive sharing	↔	↑ moderate	↑ moderate	↑ moderate	↑ 0.11	↑ 0.11	↑ 0.10
Positive HCV status ~ non-fatal overdose	↑	↑ moderate aOR 1.74 (1.19–2.55)	↑ moderate	— weak	↑ 0.23	—	↑ 0.12

 Matches expected direction
 Opposite of expected direction
 NaN correlation
 Mixed (↔) or uncertain (—) direction
 Expected direction is mixed

Figure S1: Comparison of literature-based association directions with LLM knowledge-probe directions and LLM-generated-output correlation directions.

b Knowledge-prompt discrepancies

Of the 32 variable pairs evaluated, Claude Sonnet 4.6 and DeepSeek v3.2 each produced different responses between the original and paraphrased knowledge prompts for one pair, while GPT-OSS-120B differed for four pairs (Table S2). Any change in direction category, including changes involving “mixed” or “uncertain,” counted as a discrepancy. When differences involving these classifications were excluded, Claude Sonnet 4.6 had no discrepancies, while DeepSeek v3.2 and GPT-OSS-120B each had one.

Table S2: Variable pairs with different knowledge directions under different prompt wordings.

Model	Variable pair	Original	Paraphrased
Claude Sonnet 4.6	Homelessness / Non-injection routes	decrease	uncertain
DeepSeek v3.2	Non-injection routes / HCV status	decrease	increase
GPT-OSS-120B	Treatment / Non-injection routes	mixed	increase
GPT-OSS-120B	Non-injection routes / Overdose	increase	decrease
GPT-OSS-120B	HCV status / Overdose	uncertain	mixed
GPT-OSS-120B	Receptive sharing / Overdose	mixed	increase

II. Similarity to ABM correlations

a ABM correlation discrepancy

The MAD reported in main-text Table 3 was calculated from correlations across both housing conditions; Figure S2 displays the correlations separately by housing condition. For model m , MAD was defined as

$$MAD_m = \frac{1}{|\mathcal{P}_m|} \sum_{(j,k) \in \mathcal{P}_m} |\rho_{jk,m}^{LLM} - \rho_{jk,m}^{ABM}|,$$

where $\mathcal{P}_m$ contains variable pairs with defined correlations in both datasets, and $\rho_{jk,m}^{LLM}$ and $\rho_{jk,m}^{ABM}$ denote the Spearman correlations between variables j and k in the LLM and ABM data, respectively. For each pair, both correlations used the same matched observations. All ten pairs contributed to each model's observed MAD, as well as to each of the 1,000 bootstrap MAD estimates.

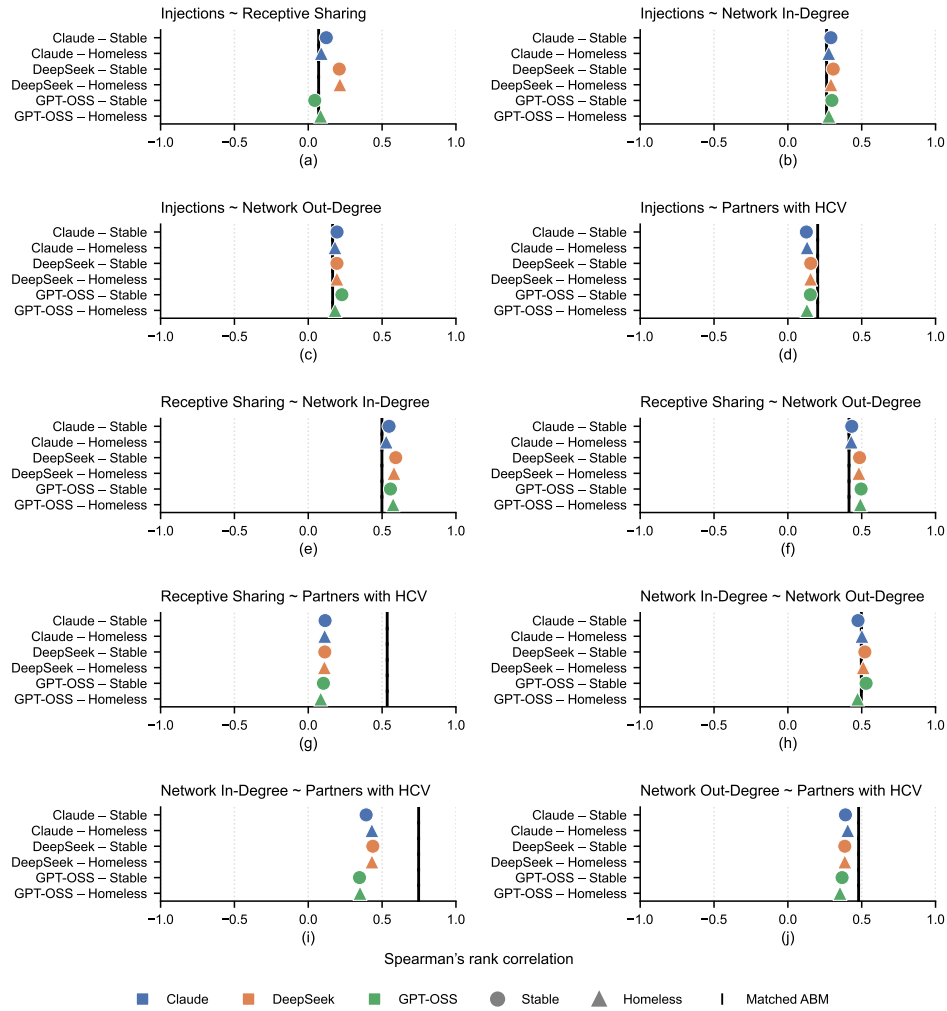


Figure S2: Comparison of ABM and LLM correlations

b Sensitivity of LLM-ABM Correlation MAD to Observation Timing

Because ABM observations were recorded at discrete simulation timepoints, we assessed the sensitivity of LLM-ABM correlation agreement to the timing of the selected ABM observations. For each model, we calculated the MAD between corresponding LLM and ABM Spearman correlations across the ten pairs formed by five shared variables.

For each ABM agent i , let t_i^0 denote the earliest activation time and t_{ir} the time of a follow-up record r , measured in years. The record closest to six months after activation was selected as

$$r_i^* = \arg \min_{r: t_{ir} > t_i^0} |t_{ir} - (t_i^0 + 0.5)|.$$

No time restriction was imposed when selecting the observation closest to the six-month target. We repeated the correlation comparison across progressively narrower tolerance windows around the six-month target, ranging from ± 10.0 to ± 0.1 years. At each tolerance, agents were retained only if their selected ABM record fell within the window. The number of retained agents decreased from 224 under the widest window to 14 under the narrowest window. Because narrower windows also changed the included agents, differences in MAD reflect both observation timing restrictions and changes in sample composition.

MAD was lowest at intermediate tolerance windows (± 0.5 years for Claude Sonnet 4.6 and DeepSeek v3.2 and ± 1 year for GPT-OSS-120B), with the highest MAD observed at the narrowest tolerance (± 0.1 years) across all three model (Figure S3).

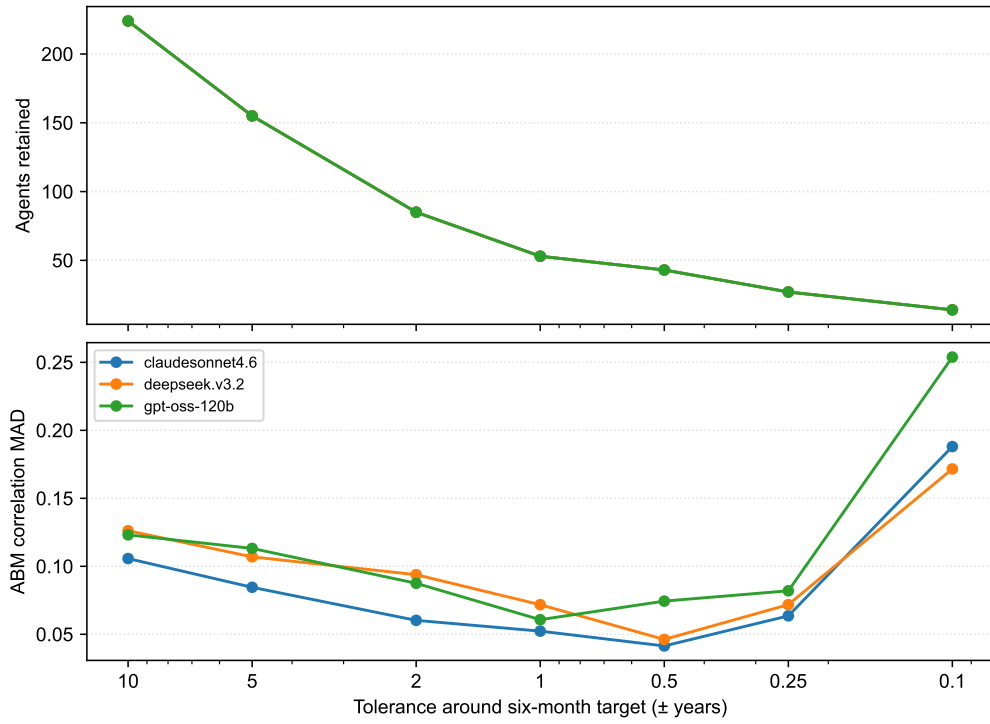


Figure S3: Sensitivity of LLM-ABM Correlation MAD to Follow-up Tolerance

III. Similarity to empirical survey associations

a Covariate coding and correlation discrepancy

The empirical correlation comparison used eight continuous variables: age, BMI, waist circumference, systolic pressure, diastolic pressure, HbA1c, total cholesterol, and HDL. The four binary variables were coded as female versus male, current daily or some-day smoking versus no current smoking, ever street-drug use versus no use, and ever injection-drug use versus no use. Unmapped or unknown binary responses were treated as missing.

For model m and covariate X_v , empirical MAD was defined as

$$M_m^{emp} = \frac{1}{|\mathcal{V}_m|} \sum_{v \in \mathcal{V}_m} |\rho_S(X_v, B) - \rho_S(X_v, L_m)|,$$

where B denotes observed binary HCV antibody status, L_m denotes the LLM-generated HCV risk score, and ρ_S denotes Spearman's rank correlation. $\mathcal{V}_m$ contains covariates with both correlations defined.

IV. Sensitivity analysis

a Directional agreement under prompt paraphrasing

Let $\Delta_{imk}^{(p)}$ denote the housing-related difference for profile i , model m , outcome k , and prompt version p , where $p = 0$ represents the original prompt and $p = 1$ the paraphrased prompt. The proportion of profiles with matching directions was calculated as

$$Q_{mk} = \frac{1}{N} \sum_{i=1}^N \mathbf{1} \left\{ \text{sign} \left(\Delta_{imk}^{(0)} \right) = \text{sign} \left(\Delta_{imk}^{(1)} \right) \right\}, \quad \text{sign}(0) = 0.$$

Here, $N = 224$ denotes the number of profiles, and Q_{mk} represents directional agreement for every model and outcome. Profiles with zero housing-related changes under both prompt versions were counted as agreement.

The model-level summary reported in the main paper is the arithmetic mean of Q_{mk} across nine numeric outcomes. Treatment and active HCV status were excluded. Overall directional agreement varied substantially across models and outcomes, ranging from 32.1% to 100% (Table S3).

Table S3: Outcome-specific directional agreement (%) between the original and paraphrased prompts across 224 profiles per model.

Outcome	Claude Sonnet 4.6	DeepSeek v3.2	GPT-OSS-120B
Overdose	49.1	100.0	32.1
Public injecting	100.0	74.6	98.7
Receptive sharing	69.6	94.6	40.2
Injection partners	71.4	90.6	35.3
Network in-degree	97.8	96.9	71.9
Network out-degree	97.3	99.1	67.4
Partner HCV proportion	97.8	100.0	88.4
Injection frequency	97.3	99.6	33.5
Non-injection routes	65.2	84.4	48.2

References

- Des Jarlais, D.C., Arasteh, K., Feelemyer, J., McKnight, C., Barnes, D.M., Perlman, D.C., Uuskula, A., Cooper, H.L.F., Tross, S., 2018. Hepatitis C virus prevalence and estimated incidence among new injectors during the opioid epidemic in New York City, 2000–2017: Protective effects of non-injecting drug use. *Drug Alcohol Depend* 192, 74–79. <https://doi.org/10.1016/j.drugalcdep.2018.07.034>
- Des Jarlais, D., Weng, C.A., Feelemyer, J., McKnight, C., 2023. Non-fatal drug overdose among persons who inject drugs during first two years of the COVID-19 pandemic in New York City: Prevalence, risk factors. *Drug Alcohol Depend Rep* 8, 100171. <https://doi.org/10.1016/j.dadr.2023.100171>
- Dong, H., Hayashi, K., Milloy, M.-J., DeBeck, K., Singer, J., Wong, H., Wood, E., Kerr, T., 2020. Changes in substance use in relation to opioid agonist therapy among people who use drugs in a Canadian setting. *Drug Alcohol Depend* 212, 108005. <https://doi.org/10.1016/j.drugalcdep.2020.108005>
- Dunne, N., Casey, M.-B., Ivers, J.-H., 2025. What are the risk factors of non-fatal overdose among persons who use opioids? A systematic review and meta-analysis. *Journal of Substance Use and Addiction Treatment* 171, 209630. <https://doi.org/10.1016/j.josat.2025.209630>
- Eckhardt, B., Winkelstein, E.R., Shu, M.A., Carden, M.R., McKnight, C., Des Jarlais, D.C., Glesby, M.J., Marks, K., Edlin, B.R., 2017. Risk factors for hepatitis C seropositivity among young people who inject drugs in New York City: Implications for prevention. *PLOS ONE* 12, e0177341. <https://doi.org/10.1371/journal.pone.0177341>
- Gahrton, C., Navér, G., Warnqvist, A., Dalgard, O., Aleman, S., Kåberg, M., 2024. Changes in hepatitis C virus prevalence and incidence among people who inject drugs in the direct acting antiviral era. *International Journal of Drug Policy* 128, 104433. <https://doi.org/10.1016/j.drugpo.2024.104433>

Guarino, H., Frank, D., Quinn, K., Kim, D., Gile, K., Ruggles, K., Friedman, S.R., Mateu-Gelabert, P., 2023. Syndemic factors associated with non-fatal overdose among young opioid users in New York City. *Front. Public Health* 11, 1195657. <https://doi.org/10.3389/fpubh.2023.1195657>

Gunaratne, M.P., Loeb, T.A., McFall, A.M., Ganapathi, L., Baishya, J.J., Kedar, A., Sinha, A.K., Srikrishnan, A.K., Solomon, S.S., Lucas, G.M., Mehta, S.H., 2026. Characterizing non-fatal overdose among people who inject drugs in India: prevalence and risk factor analysis, 2022-2024. *International Journal of Drug Policy* 147, 105073. <https://doi.org/10.1016/j.drugpo.2025.105073>

Hahn, J.A., Evans, J.L., Davidson, P.J., Lum, P.J., Page, K., 2010. Hepatitis C virus risk behaviors within the partnerships of young injecting drug users. *Addiction* 105, 1254–1264. <https://doi.org/10.1111/j.1360-0443.2010.02949.x>

Hahn, J.A., Page-Shafer, K., Lum, P.J., Bourgois, P., Stein, E., Evans, J.L., Busch, M.P., Tobler, L.H., Phelps, B., Moss, A.R., 2002. Hepatitis C virus seroconversion among young injection drug users: relationships and risks. *J Infect Dis* 186, 1558–1564. <https://doi.org/10.1086/345554>

Jenkins, L.M., Banta-Green, C.J., Maynard, C., Kingston, S., Hanrahan, M., Merrill, J.O., Coffin, P.O., 2011. Risk Factors for Nonfatal Overdose at Seattle-Area Syringe Exchanges. *J Urban Health* 88, 118–128. <https://doi.org/10.1007/s11524-010-9525-6>

Kan, M., Holmén, E., Brynte, C., Ekström, A.M., Kåberg, M., 2026. Needle and syringe coverage per injection and factors associated with needle sharing among clients of the Stockholm Needle and Syringe Program, 2013–2023. *International Journal of Drug Policy* 153, 105299. <https://doi.org/10.1016/j.drugpo.2026.105299>

Karlsson, N., Kåberg, M., Berglund, T., Hammarberg, A., Widman, L., Ekström, A.M., 2021. A prospective cohort study of risk behaviours, retention and loss to follow-up over 5 years among women and men in a needle exchange program in Stockholm, Sweden. *Int J Drug Policy* 90, 103059. <https://doi.org/10.1016/j.drugpo.2020.103059>

Khezri, M., Kimball, S., McKnight, C., Rouhani, S., Bunting, A.M., Karamouzian, M., Ompad, D.C., Des Jarlais, D., 2025. Harms associated with injecting in public spaces: a global systematic review and meta-analysis. *Int J Drug Policy* 140, 104819. <https://doi.org/10.1016/j.drugpo.2025.104819>

Kral, A.H., Lambdin, B.H., Browne, E.N., Wenger, L.D., Bluthenthal, R.N., Zibbell, J.E., Davidson, P.J., 2021. Transition from injecting opioids to smoking fentanyl in San Francisco, California. *Drug Alcohol Depend* 227, 109003. <https://doi.org/10.1016/j.drugalcdep.2021.109003>

Mackesy-Amiti, M.E., Boodram, B., Page, K., Latkin, C., 2023. Injection partnership characteristics and HCV status associations with syringe and equipment sharing among people who inject drugs. *BMC Public Health* 23, 1191. <https://doi.org/10.1186/s12889-023-16133-5>

Morris, M.D., Evans, J., Montgomery, M., Yu, M., Briceno, A., Page, K., Hahn, J.A., 2014. Intimate Injection Partnerships Are at Elevated Risk of High-Risk Injecting: A Multi-Level Longitudinal Study of HCV-Serodiscordant Injection Partnerships in San Francisco, CA. *PLOS ONE* 9, e109282. <https://doi.org/10.1371/journal.pone.0109282>

Pouget, E.R., Hagan, H., Des Jarlais, D.C., 2012. Meta-analysis of hepatitis C seroconversion in relation to shared syringes and drug preparation equipment. *Addiction* 107, 1057–1065. <https://doi.org/10.1111/j.1360-0443.2011.03765.x>

Rafful, C., Orozco, R., Rangel, G., Davidson, P., Werb, D., Beletsky, L., Strathdee, S.A., 2018. Increased non-fatal overdose risk associated with involuntary drug treatment in a longitudinal study with people who inject drugs. *Addiction* 113, 1056–1063. <https://doi.org/10.1111/add.14159>

Rahimi-Movaghar, A., Amin-Esmacili, M., Shadloo, B., Noroozi, A., Malekinejad, M., 2015. Transition to injecting drug use in Iran: A systematic review of qualitative and quantitative evidence. *International Journal of Drug Policy* 26, 808–819. <https://doi.org/10.1016/j.drugpo.2015.04.018>

Rudolph, A.E., Rhodes, S.E., 2023. Network Correlates of Using a Syringe After an Injection Partner Among Women Who Inject Drugs in Philadelphia Pennsylvania. *AIDS Behav* 27, 957–968. <https://doi.org/10.1007/s10461-022-03832-0>

Schiavon, S., Hodgin, K., Sellers, A., Word, M., Galbraith, J.W., Dantzler, J., Cropsey, K.L., 2018. Medical, psychosocial, and treatment predictors of opioid overdose among high risk opioid users. *Addict Behav* 86, 51–55. <https://doi.org/10.1016/j.addbeh.2018.05.029>

Spelman, T., Sacks-Davis, R., Dietze, P., Higgs, P., Hellard, M., 2019. Injection drug network characteristics as a predictor of injection behaviour. *Epidemiol Infect* 147, e173. <https://doi.org/10.1017/S095026881900061X>

Tran, L.T., Peacock, A., Colledge, S., Memedovic, S., Grebely, J., Leung, J., Larney, S., Trickey, A., Stone, J., Vickerman, P., Hickman, M., Degenhardt, L., 2020. Injecting risk behaviours amongst people who inject drugs: A global multi-stage systematic review and meta-analysis. *International Journal of Drug Policy* 84, 102866. <https://doi.org/10.1016/j.drugpo.2020.102866>

Walker, J.G., Akiyama, M.J., Artenie, A., Cleland, C.M., Lizcano, J.A., Musyoki, H., Nyakowa, M., Cherutich, P., Kurth, A.E., Vickerman, P., 2025. Impact of scaling up harm reduction interventions on injecting risk behaviours, ART outcomes and HIV incidence among people who inject drugs in Kenya. *International Journal of Drug Policy* 140, 104824. <https://doi.org/10.1016/j.drugpo.2025.104824>