LUSIM: AN LLM-DRIVEN SPATIAL-EQUILIBRIUM SIMULATOR FOR URBAN PLANNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantitative ex-ante urban-policy evaluation has historically relied on closed-form spatial-equilibrium and low-dimensional discrete-choice models that constrain analysis to homogeneous agents and fixed-form parametric utilities, and thereby to smooth proportional responses around the calibrated baseline. Integrating large language models into multi-agent simulation alleviates these constraints, but existing LLM-driven platforms lack the market-clearing mechanism and externally calibrated parameters that quantitative policy evaluation requires. We present **LUSim** (LLM-driven Urban Simulator), a multi-agent simulator in which heterogeneous households choose residence and workplace through a large language model and clear housing and labour markets inside an externally calibrated general-equilibrium city, with structural decision rules as baselines under identical conditions. We simulate a new rapid-transit line on a 96-zone Berlin instance calibrated from Ahlfeldt et al. (2015). The structural engines, which can only reweight the calibrated city in proportion, concentrate their response at the new stations, whereas with LLM-driven households approximately eight times as many relocate and the gains shift to peripheral districts. To test which engine predicts real urban change, we initialise every engine on the divided Berlin of 1986, let it predict the reunified city, and score, by criteria fixed in advance, its prediction of which districts gain or lose residents, jobs, prices, and wages against the changes observed by 2006. The simulation with LLM-driven households predicts these changes better than every structural baseline, also with district names withheld so that it cannot rely on memorised history, and the contrast persists across a second rail corridor, a zoning policy that expands floor space, and a second model family. We interpret these results as evidence that LLM-driven decision rules, disciplined by market clearing in a calibrated environment, predict urban change beyond the reach of structural models.

1 INTRODUCTION

Evaluating transport, housing, and place-based policies before they are enacted requires simulating how heterogeneous populations respond to environmental change. Recent advances in large language models open a substantial new design space at the decision-rule layer, lifting the homogeneous-agent and closed-form-utility constraints of existing urban-planning simulators. LLM agents reproduce many human economic-reasoning patterns (Horton et al., 2023; Hewitt et al., 2024; Manning & Horton, 2026), GPU-accelerated platforms run heterogeneous LLM-conditioned populations at scales of 10^4 to 10^6 (Chopra et al., 2024; Yang et al., 2024; Piao et al., 2025; Bougie & Watanabe, 2025), and a behavioural economics of AI is starting to characterise the systematic biases of LLM agents (Bini et al., 2026).

Yet none of these advances has been integrated into a calibrated spatial-equilibrium environment for urban-planning evaluation. Standard tools, including quantitative spatial-equilibrium models (Allen & Arkolakis, 2014; Ahlfeldt et al., 2015; Redding & Rossi-Hansberg, 2017), discrete-choice estimation (McFadden, 1978; Bayer et al., 2007), and agent-based simulators (Horni et al., 2016; Waddell, 2002), restrict counterfactual policy predictions to smooth proportional responses around the calibrated baseline. Existing LLM multi-agent platforms (Chopra et al., 2024; Yang et al., 2024; Piao et al., 2025; Bougie & Watanabe, 2025; Cong et al., 2024) simulate rich individual behaviour within a fixed city, but lack the market clearing and external calibration through which a policy

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

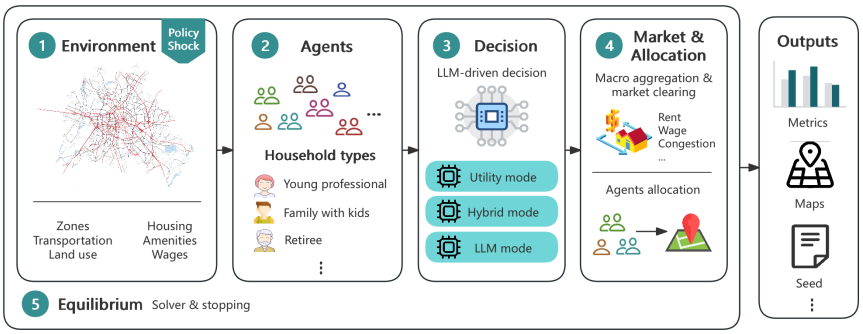

Figure 1: LUSim's five-stage workflow. The environment (zones, transport, and an optional policy intervention) feeds the agent population, which the decision engine maps to discrete location decisions. The market mechanism aggregates demand into clearing variables, and the equilibrium solver iterates the four micro-stages until prices and choices stabilise. The specific clearing variables and convergence criteria depend on the instance and are described in Sections 3 and 4.

changes the city itself, so their outputs cannot be compared with the prices, wages, and employment the empirical record contains.

To fill the gap, we present **LUSim** (LLM-driven Urban Simulator), a multi-agent spatial-equilibrium simulator in which heterogeneous households choose residence and workplace through a large language model and clear housing and labour markets inside an externally calibrated general-equilibrium city. LUSim consists of five components with defined interfaces (Figure 1), of which the decision rule is the one the language model occupies. Each household is instantiated from a census-derived demographic profile, scores candidate residences and selects a workplace through a two-stage prompt, and the market mechanism aggregates these choices, clears housing prices and wages, and returns the updated prices to the agents until a fixed point is reached. Holding the environment, population, and market mechanism fixed, the same simulation can be run with the closed-form, agent-based, and hybrid decision rules of the structural tradition, so the LLM-driven engine can be compared with four structural baselines under identical conditions and any difference is attributable to the decision rule alone.

We instantiate LUSim on Berlin, partitioned into its 96 Ortsteile (administrative districts), with the externally calibrated structural parameters of Ahlfeldt et al. (2015), and compare the LLM-driven engine with the structural baselines in two experiments. The first experiment introduces a hypothetical four-station rapid-transit line. Households driven by the structural engines, which can only reweight the calibrated cross-section in proportion, concentrate about half of all relocation in the four station zones, and central wages fall because a shorter commute lets firms pay less (the compensating differential). With LLM-driven households, about eight times as much residential mass relocates, the gains spread to peripheral zones, and central wages rise because more workers can now reach central jobs (labour-market pooling). To test whether the LLM-driven engine outperforms the structural baselines in predicting the effect of a real urban change, we initialise every engine on the divided Berlin of 1986, let it predict the new equilibrium once the transport network is reunified, and score its prediction of which districts gained or lost residents, workplaces, housing prices, and wages relative to the city against the changes observed between 1986 and 2006. The simulation with LLM-driven households predicts this pattern better than any structural baseline, and does so even when district names are withheld from the model, whereas the closed-form engine, calibrated to the observed city, gets the direction of change right less often than chance on three of the four variables. The contrast is not an artefact of the rail corridor, the policy, or the model family. It reappears on a second rail corridor, a zoning policy that expands floor space, and a second model family, and the pipeline runs unchanged on a Singapore instance with a different market mechanism and at the 12,309-block resolution of the replication data.

Our contributions are threefold. The first is LUSim, to our knowledge the first simulator to couple an LLM decision rule with externally calibrated spatial-equilibrium parameters and market clearing and to run structural decision rules alongside it under a shared mechanism (Table 1). The second is

Table 1: Capabilities matrix. To our knowledge, LUSim is the first simulator to combine an LLM-driven decision rule with externally calibrated spatial-equilibrium parameters and market clearing, and the first to run structural decision rules alongside it under a shared market mechanism.

Capability	CitySim	AgentSociety	UrbanSim	Ahlfeldt 2015	**LUSim**
LLM decision rule	Full LLM	Full LLM	No	No	Full LLM, hybrid
Market clearing (prices, wages)	No	No	No	Closed-form	Any rule
Externally calibrated parameters	No	No	Iterative tuning	Yes	Yes
Policy counterfactuals	No	No	Yes	Yes	Yes
Validation against observed outcomes	Behavioural	Social experiments	Case study	Reunification	Reunification
Structural baselines alongside	No	No	No	n/a	Yes
Domain	Urban behaviour	Society	Land use	Urban economics	Urban economics

a controlled comparison and mechanism decomposition establishing that LLM-driven households reach a density-relocation regime outside the proportional-shift range of the Cobb-Douglas and Fréchet family, and identifying the channel that separates the two. The third is a historical validation against the reunification of Berlin, with criteria fixed in advance and district names withheld, that ranks the LLM-driven engine ahead of the structural family in predicting where observed change concentrated.

Section 2 reviews related work, Section 3 formalises LUSim and its LLM-driven engine, Section 4 specifies the Berlin instance and the structural baselines, Section 5 reports the results, and Section 6 concludes. LUSim's codebase[1] and the Berlin data with the LLM response caches are released under non-restrictive licences (Appendix E).

2 Background and prior work

Three method families dominate quantitative urban planning, each with a distinctive commitment at the decision-rule layer. Quantitative spatial economics integrates closed-form discrete-choice systems with Fréchet shocks into gravity equations and equilibrium-clearing counterfactuals (Allen & Arkolakis, 2014; Redding & Rossi-Hansberg, 2017), and Ahlfeldt et al. (2015) develop the within-city analogue with residence-workplace pair Fréchet shocks, calibrated against block-level Berlin data and identified using the city's division and reunification. Solved analytically, such a model can be calibrated to a real city, but a policy can then only scale the calibrated pattern up or down. Discrete-choice estimation in the McFadden (1978) random-utility tradition recovers preference parameters from observed micro-data by maximum likelihood, extended by Bayer et al. (2007) with explicit sorting on unobservables. Agent-based microsimulation executes per-agent procedural rules over millions of synthetic agents, as in UrbanSim (Waddell, 2002) and MATSim (Horni et al., 2016). It represents heterogeneity directly, but its behavioural rules are hand-written and its parameters are typically tuned rather than estimated from a structural moment condition. LUSim treats all three as instances of the same `DecisionEngine` interface clearing the same market as the LLM-driven households.

Recent LLM multi-agent platforms simulate heterogeneous populations whose decision rules are expressed in natural language. AgentTorch (Chopra et al., 2024) compiles LLM-elicited behavioural archetypes into differentiable population-scale kernels. AgentSociety (Piao et al., 2025) places LLM agents in a synthetic society spanning mobility, employment, consumption, and social interaction, validated against four published social-science experiments. OASIS (Yang et al., 2024) scales social-media simulation to one million agents. CitySim (Bougie & Watanabe, 2025), the closest to an urban simulator, endows tens of thousands of LLM agents with beliefs, goals, and spatial memory to generate daily activity schedules, validating crowd densities and place popularity against observed data. LLM-in-the-loop simulators also exist for driving (Fu et al., 2024), where the language model controls a vehicle rather than a household. Cong (2025) combines LLM-based borrower simulations with data-driven lending models to study counterfactual equilibrium in online lending. None of these targets a spatial-equilibrium task with an externally calibrated parameter set, and in none of them do

[1] Anonymised mirror for review at `https://anonymous.4open.science/r/agent-urban-planning-iclr2027-45BC/` (MIT licence). The mirror bundles the source, documentation, and the PyPI-installable package. Public-repository, ReadTheDocs, and PyPI URLs will replace this footnote in the camera-ready version. See Appendix J.

the agents' choices feed back into prices and wages through market clearing, so a policy can change the agents' behaviour but not the city's equilibrium (Table 1).

3 THE LUSIM SIMULATOR

3.1 THE URBAN PLANNING PROBLEM

An urban area is populated with N heterogeneous agents indexed by i, each characterised by a type $\theta_i \in \Theta$ encoding the demographic, employment, and budget attributes relevant for residence and workplace decisions, and partitioned into Z zones $\mathcal{Z} = \{1, \dots, Z\}$. The baseline environment state is E_0, and a candidate policy s (an exogenous change to the environment, which economics calls a shock) maps it to a counterfactual environment $E^{(s)} = \text{apply_policy}(E_0, s)$. At iteration k of a within-period market loop with endogenous state S_k, agent i forms her information set through a type-dependent perception function $I_{i,k} = \text{obs}(E^{(s)}, S_k; \theta_i)$ and selects an action $a_{i,k} = \pi_i(I_{i,k}) \in \mathcal{A}$ according to her decision rule π_i. The market mechanism aggregates actions via $\mathcal{M}(\{a_{i,k}, w_i\}, E^{(s)}, S_k) \to S_{k+1}$ and is invoked iteratively until the stationarity criterion of Component 5 is met. The post-policy outcome is $R^{(s)}$ and aggregate welfare is $W^{(s)} = \sum_i w_i V_i(R^{(s)}; \theta_i)$, where w_i are distributional weights and V_i is agent i's value function. The data-flow sequence is drawn in Appendix F, Figure 8.

Component 1 (Environment). The environment carries the spatial, transport, and fundamentals primitives, $E_0 = (\mathcal{Z}, \tau, A, B, H, \psi)$, where $\tau \in \mathbb{R}_{\geq 0}^{Z \times Z}$ is the inter-zone travel-time matrix, $A, B \in \mathbb{R}_{>0}^{Z}$ are baseline productivity and amenity vectors, $H \in \mathbb{R}_{>0}^{Z}$ is per-zone floor-space supply, and ψ collects the structural parameters governing preferences, technology, and market clearing. Policies act through apply_policy. A new transit line shrinks selected entries of τ, while a zoning change perturbs H.

Component 2 (Agents). The agent population is a finite set of heterogeneous types θ_i drawn from a distribution F over Θ, with weights w_i encoding the population mass each agent represents, which feed both the market aggregation and the welfare aggregation.

Component 3 (Decision rule). Each agent has a decision rule $\pi_i \colon I_{i,k} \mapsto a_{i,k}$, on whose functional form LUSim places no commitment. The simulator's own rule is the LLM-driven engine of Section 3.2, a language-model call against a persona constructed from θ_i, and four structural reference engines (Section 4) serve as baselines.

Component 4 (Market mechanism). The mechanism $\mathcal{M}$ aggregates per-agent actions into per-zone aggregates and produces the state S_{k+1} that clears the markets at the next iteration, where S_k comprises the price vector and, optionally, density-dependent fundamentals such as agglomeration-recomputed productivities and amenities. $\mathcal{M}$ can be a single- or multi-market tâtonnement (iterative price adjustment until every market clears), an auction, or any other map from actions and state to next-period state.

Component 5 (Equilibrium). The simulator iterates the (decision, mechanism) pair until $\|S_{k+1} - S_k\| < \epsilon_S$, supplemented by $\max_i \|V_i^{k+1} - V_i^k\| < \epsilon_V$ for stochastic decision rules, and the fixed point $S^\star$ is the predicted post-policy equilibrium. Outputs collect into the per-zone tuple $R^{(s)}$ and the aggregate welfare $W^{(s)}$, and counterfactual comparison reads off $W^{(s)} - W^{(0)}$ and $R^{(s)} - R^{(0)}$.

3.2 THE LLM-DRIVEN DECISION ENGINE

The LLM-driven engine, denoted V5 in the tables below, dispenses with a parametric utility altogether and samples each agent's residence-workplace pair from a language model conditioned on a persona and on the current market state,

$$\pi_i^{V5}(I) \sim \text{LLM_decide}(\text{persona}(\theta_i), \text{context}(I)). \tag{1}$$

The persona renders the attributes of θ_i (Section 4) as a short natural-language description, and the context renders the per-zone housing prices and wages of the current state S_k together with the travel times τ from each candidate zone. The engine is nonparametric at the decision-rule layer and carries no structural parameter to calibrate. The decision runs in two stages per market iteration. In the first

stage the model scores every zone as a residence on a 0–10 scale given the persona and current prices, and the ten highest-scoring zones are shortlisted. In the second stage it selects a workplace for each shortlisted residence given wages and commute times, returning the best pair with a 0–100 self-rated satisfaction that serves as the engine's own welfare ruler V_i. The market mechanism then clears housing prices and wages exactly as for the structural engines, so the language model never sets a price and every equilibrium quantity it influences passes through market clearing.

Two devices keep the call volume tractable, following the archetype pattern of existing LLM platforms (Chopra et al., 2024; Piao et al., 2025; Yang et al., 2024). The N agent types are clustered into K archetypes by k-means over the attribute vector, the model is queried once per archetype per iteration with the decision broadcast to the cluster members under their population weights, and a response cache keyed by cluster, stage, and price bucket raises the late-iteration hit rate as prices converge. Prompt templates, clustering details, and the released cache are in Appendix C.

Pipeline. Each LUSim run is a five-step Python pipeline (load a calibrated scenario, sample the agent population, construct the decision engine, assemble the `SimulationEngine`, and call `sim.run(policy=...)` for the per-zone equilibrium tuple $R^{(s)}$ and welfare $W^{(s)}$), and the structural baselines of Section 4 are obtained by replacing only the `engine` constructor (Appendix F).

4 Berlin instance and structural baselines

We evaluate the simulator on a calibrated Berlin instance, holding the environment, agent population, market mechanism, and equilibrium criterion fixed while varying the decision rule across the LLM-driven engine and four structural baselines (V1–V4 share a structural inner utility while V5 alone dispenses with it).

Controlled Berlin instance. The instance reproduces the calibrated 2006 equilibrium of the Ahlfeldt et al. (2015) replication pack, aggregated to the 96-zone Berlin Ortsteile partition by a bundled crosswalk that maps the pack's 12,309 statistical blocks to Ortsteile with conserving sums. A second instance, the divided city of 1986 with the walled travel-time network of the same replication data, is used for the historical validation of Section 5.2 and is described in Appendix K.

The baseline state $E_0 = (\mathcal{Z}, \tau, A, B, H, \psi)$ collects the 96-zone partition, the public-transit travel-time matrix, the calibrated baseline productivity and amenity vectors A, B, the per-zone floor-space supply H, and the structural-parameter vector ψ (full enumeration in Appendix A). The household population $\mathcal{H}$ comprises $N = 10{,}000$ heterogeneous types θ_i encoding ten demographic, employment, and budget attributes (head age, household size, monthly net income, presence of children, presence of elderly, car ownership, education, migration background, employment status, tenure) with a home Ortsteil. Types are sampled from a joint distribution fitted to public Berlin marginals (Berlin Zensus 2011, Senate Department for Urban Development, Berlin Statistics Office) by iterative proportional fitting. Population weights w_i are proportional to the replication pack's residence mass (provenance in Appendix B, Table 4).

The structural inner utility U_i shared by the four baselines specialises to the Ahlfeldt et al. (2015) log-form

$$\ln u_{ij}(\theta) = \ln B_i + \ln w_j - (1-\beta)\ln Q_i - \kappa\,\tau_{ij} + \xi_{ij}, \tag{2}$$

where β and $(1-\beta)$ are Cobb-Douglas consumption and housing shares, κ is the per-minute commute decay, and ξ_{ij} is an i.i.d. idiosyncratic shock whose engine-specific distribution is Fréchet(ε) for V1, V2, and V4, Gaussian for V3, and nonparametric for V5. The first four terms are denoted $U_i(a; I, \psi)$ below. The Fréchet-integrated gravity probability and calibrated values of $\kappa, \varepsilon, \beta, \alpha$ are in Appendix A. The market mechanism is a two-market tâtonnement clearing housing prices Q (the price per unit of floor space) and wages w simultaneously. The household side aggregates residence demand from the per-agent choices, the firm side sets wages from a Cobb-Douglas production function with labour share $\alpha = 0.8$ (Appendix A), updates apply the replication-pack damping, and productivities and amenities recompute at every iteration from the current density (endogenous agglomeration, with formulas and parameters in Appendix A). For the Berlin instance $S = (Q, w)$, convergence is declared at $\|S_{k+1} - S_k\|_\infty < 10^{-3}$, and each engine outputs $R^{(s)} = (Q^\star, w^\star, HR^\star, HM^\star)$, i.e., housing prices, wages, residential employment (HR_i, workers

by residence) and workplace employment (HM_j, workers by workplace), as a 96-row table with the observed 2006 targets attached.

The counterfactual policy s is the East-West Express, a hypothetical four-station rapid-transit line that shortens travel times τ, linking Marzahn (eastern outer suburb), Mitte (central business district), Lichtenberg, and Charlottenburg (inner-west terminus) with five minutes of nominal in-vehicle time between adjacent stations. The travel-time update is the standard shortest-path relaxation $\tau_{ij}^{(s)} = \min(\tau_{ij}, \tau_{ij}^{\text{via-rail}})$ through the four station nodes (Allen & Arkolakis, 2014; Heblich et al., 2020) (Appendix B), reducing 164 of the 9,216 (residence, workplace) pairs by 5.93 minutes commuting time on average, so the policy is local to the four stations rather than city-wide. A second rail corridor and a floor-space supply policy are specified in Section 5.3 and Appendix M.

Structural baselines. The four structural engines plug the inner utility, Eq. (2), into different decision rules and, with V5, isolate closed-form against sampled choice, the noise family, and the source of preference primitives.

$$\pi_i^{V1}(I) = \text{softmax}_a[\,\varepsilon\, U_i(a; I, \psi)\,], \tag{3}$$

$$\pi_i^{V2}(I) = \arg\max_a \left[\, e^{U_i(a;I,\psi)} \cdot \xi_a \,\right], \quad \xi_a \sim \text{Fréchet}(\varepsilon), \tag{4}$$

$$\pi_i^{V3}(I) = \arg\max_a [\, U_i(a; I, \psi) + \xi_a \,], \quad \xi_a \sim \mathcal{N}(0, \pi^2/6), \tag{5}$$

$$\pi_i^{V4}(I) = \arg\max_a U_i(a; I, \hat{\phi}_i), \quad \hat{\phi}_i = \text{LLM_elicit}(\text{persona}(\theta_i)). \tag{6}$$

V1 is the closed-form softmax over the structural utility (deterministic). V2 is its per-agent agent-based realisation with Fréchet shocks, recovering V1 in expectation. V3 substitutes Gaussian shocks at matched variance. V4 is the hybrid engine, which elicits per-agent preference weights from the LLM once per demographic profile and applies them in a closed-form argmax (Appendix C).

V1 is deterministic and V2–V5 are stochastic with seed pinning. V1–V2 vary analytic against sampled choice, V2–V3 swap the noise family, V4 scales the inner-utility coefficients by an LLM-elicited weight vector $\hat{\phi}_i \in \mathbb{R}^4$ over (rent, commute, amenity, wage), and V5 dispenses with the structural utility entirely, using $K = 50$ archetypes in the controlled comparison. Implementation details and prompt templates are in Appendix C.

5 Results

Section 5.1 exercises all five engines on the East-West Express and decomposes the mechanism behind the divergence, Section 5.2 evaluates every engine against the reunification of Berlin, and Section 5.3 reports robustness and generality. Baselines use the observed warm start from the replication pack, converging to the observed equilibrium for V1 and to close neighbours for V2–V5, and every shock is warm-started from, and measured against, the engine's own baseline.

5.1 Controlled comparison on a hypothetical transit line

Table 2: Cross-engine response to the East-West Express across the 96 Ortsteile. $\Delta\bar{Q}$ is the level mean of post-policy housing-price changes, $\Delta Y\%$ the percentage change in aggregate output, $\Delta\langle U\rangle_{V1}$ the welfare change under the V1 formula applied uniformly, and $\Delta\langle U\rangle_{\text{own}}$ under each engine's own formula (for V5, the percentage drop in self-rated satisfaction). Full table in Appendix C, Table 6.

Engine	$\mu_{\Delta \log Q}$	$\sigma_{\Delta \log Q}$	$p_{95}\lvert\Delta \log Q\rvert$	$\Delta\bar{Q}$	$\Delta Y\%$	$\mu_{\Delta \log w}$	$\Delta\langle U\rangle_{V1}$	$\Delta\langle U\rangle_{\text{own}}$
V1 softmax	0.0004	0.0070	0.0083	0.0008	0.0007	−0.0001	−0.0032	−0.0032
V2 ABM	0.0004	0.0071	0.0080	0.0007	−0.0000	−0.0001	−0.0031	−0.0031
V3 Normal	0.0004	0.0042	0.0048	0.0001	0.0132	−0.0001	−0.0027	−0.0029
V4 Hybrid	0.0004	0.0068	0.0077	0.0007	0.0002	−0.0001	−0.0031	−0.0030
V5 LLM	**0.0016**	**0.0083**	**0.0172**	**0.0111**	**0.0299**	**−0.0003**	**−0.0056**	**−0.0051**

Only the decision rule π_i varies across the five engines. We report the per-zone $\Delta \log Q$ surface (Figure 2), distribution-shape moments across the 96 Ortsteile (Table 2), welfare under each engine's own formula and a common V1 ruler, and a decomposition of the response into how much mass moves, where, and through which wage channel (Appendix L, Table 18). The structural family is bounded by the Cobb-Douglas and Fréchet inner utility, which restricts the density response to proportional shifts of the calibrated cross-section, and V5 carries no such constraint.

Size and spread of the response. The structural engines move housing prices by less than 0.1% on average and confine the response to the station zones, whereas V5 moves them fourteen times as much and spreads the response across the city (Table 2). V1, V2, and V4 land on a tight consensus, with $\mu_{\Delta \log Q} = 0.0004$, $\sigma_{\Delta \log Q} \approx 0.0070$, and a welfare drop under the V1 formula of -0.0031 to -0.0032. V5 separates from all four. Its level mean rent change $\Delta \bar{Q} = 0.0111$ is approximately fourteen times the V1/V2/V4 pool of 0.0007 to 0.0008, and its $\Delta Y\% = 0.030$ is the only sizeable positive aggregate-output value.

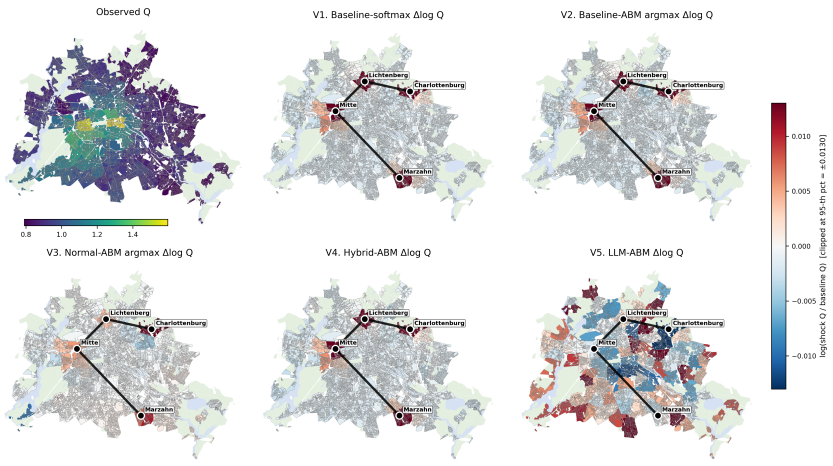

Figure 2: Per-zone $\Delta \log Q$ choropleths for V1 through V5 under the East-West Express. The structural engines V1, V2, V3, and V4 concentrate the response at the four station zones. V5 produces a city-wide pattern, with gains in peripheral zones and relief in central ones. The corresponding $\Delta \log w$, $\Delta \log HR$, $\Delta \log HM$, and $\Delta \log Y$ surfaces are in Appendix D.

Housing prices and wages. V5 produces a qualitatively different $\Delta \log Q$ surface (Figure 2). The structural engines concentrate $\Delta \log Q$ at the four station zones, reproducing the textbook accessibility-capitalisation effect in which the price gain lands on the zones whose access improved, whereas V5's response at the same zones is approximately zero or slightly negative, with the largest gainers in eastern peripheral zones ($\Delta \log Q \approx 0.024$ to 0.032). The resulting rent-gradient flattening is of the kind documented after large transit extensions in Paris (Mayer & Trevien, 2017), Beijing (Zheng & Kahn, 2008), and the Berlin S-Bahn re-integration after reunification (Ahlfeldt et al., 2015), and Section 5.2 evaluates the response against the historical record directly. The wage surface separates the engines even more cleanly (Figure 3, Appendix D). V1–V4 all predict that wages fall in central Berlin after the new line opens, because in their decision rule a shorter commute lets firms pay less, the Rosen-Roback compensating differential (Rosen, 1979; Roback, 1982) restored at a lower wage. V5 predicts that wages rise in the same central zones, because more workers can now reach central jobs.

Mechanism. To locate the source of the divergence we decompose the equilibrium response of each engine along three dimensions, namely how much residential mass relocates, where the response is located, and which economic channel carries the wage effect (Table 18). Three findings follow. First, V5 relocates approximately eight times as much residential mass as the structural family. The smooth decision rule shared by V1–V4 can only reweight the existing cross-section in proportion, whereas selecting the best option from language-model scores permits genuine movement of mass between zones. Second, the structural family places about half of its response in the four station zones, while V5 places 3% there and distributes the rest across peripheral zones. Third, and most

diagnostically, the wage responses of the structural engines are almost entirely explained by improvements in access, with a correlation of about -0.88 between $\Delta \log w$ and commuting minutes saved, the compensating differential operating as the theory prescribes. For V5 that correlation is zero, so the compensating-differential channel is absent and the wage response instead tracks changes in workplace density, consistent with labour-market pooling, in which a larger reachable pool of workers raises wages (Marshall, 1890; Duranton & Puga, 2004). The most informative row is V4: on every dimension the hybrid engine sits with the structural family, so eliciting preference weights from a language model heterogenises the structural architecture without changing its economic channel. Three further checks, a monotonicity test under a uniform commute reduction, a run of V5 with zone names replaced by anonymous identifiers, and a sweep over the number of archetypes $K \in \{25, 50, 75\}$, are in Appendix H.

5.2 Validation against the division and reunification of Berlin

Whether any engine's response tracks observed changes is a question for held-out validation, which the division and reunification of Berlin makes possible. Each engine is initialised on the divided city of 1986, with the walled travel-time network and the fundamentals of that year, and predicts the new equilibrium once the network is reunified. Predicted per-zone changes are compared with those recorded between 1986 and 2006 on the 55 West Berlin zones for which the replication data records 1986 values. Each engine's prediction is measured as a change against its own 1986 baseline, so that an engine whose baseline levels fit less well is not penalised twice. Twenty years contain far more than reunification, so the citywide mean is removed from both the predicted and the observed changes, leaving only the spatial pattern to be scored. Because the reunification outcome appears in the training data of any language model, we also run V5 with every district name replaced by an anonymous identifier, so the model cannot tell which real district it is looking at. The structural engines use the parameters calibrated on the replication data's three-year panel with the agglomeration terms held fixed, and V5 runs with $K = 25$ archetypes and a twenty-iteration cap. Instance and configuration choices are documented in Appendix K.

Each engine is scored on four variables, the per-zone changes in housing prices, residents, workplaces, and wages, by four metrics each, the Pearson and Spearman correlations between the predicted and observed patterns, the share of zones whose direction relative to the city is predicted correctly, and the relative RMSE. This gives sixteen cells. In every cell the six engines are ranked, and an engine's overall score is its average rank across the cells, lower being better, with ties broken by mean Pearson correlation. The criteria and the ranking rule were fixed before any engine was scored, and Appendix K lists every cell.

Table 3: Historical validation on the 55 West Berlin zones. Pearson correlations between centred predicted and observed changes, the share of zones whose direction relative to the city was predicted correctly (four variables in turn, 0.5 is chance), and the average rank under the scoring rule defined in the text, lower being better. V2, V3, and the remaining cells are in Appendix K.

Engine	$\Delta \log Q$	$\Delta \log HR$	$\Delta \log HM$	$\Delta \log w$	Correct direction	Average rank
V1 softmax	-0.32	-0.09	-0.03	-0.14	0.34 / 0.46 / 0.51 / 0.36	4.19
V4 Hybrid	-0.24	0.00	-0.01	-0.10	0.44 / 0.56 / 0.46 / 0.47	3.25
V5 LLM	$\mathbf{0.25}$	-0.13	$\mathbf{0.05}$	$\mathbf{0.00}$	0.66 / 0.38 / 0.46 / 0.58	$\mathbf{2.25}$
V5, names withheld	-0.16	$\mathbf{0.23}$	$\mathbf{0.05}$	-0.07	0.49 / 0.51 / 0.56 / 0.51	2.25

Table 3 reports the result. V5 ranks first with an average rank of 2.25, the run with names withheld ties with it and is placed second by the tie-break, and every structural engine follows, V4 at 3.25 and V1 at 4.19, with V2 and V3, which populate only part of the grid, ranked between them (Appendix K). Behind the ranking, both structural engines in the table are negatively correlated with the observed record on every variable, and the closed-form engine gets the direction right less often than guessing would on three of the four. The two V5 runs hold the only positive correlations, and the run with names withheld, which is given no information identifying the districts, is at or above the guessing level on all four. All correlations are modest. Withholding the district names removed the named run's advantage on housing prices, which we attribute to recognition of the zones rather than to the engine's reasoning. With names withheld, V5 still leads the structural engines on residents and workplaces and keeps its rank ahead of them, so the ranking does not depend on place recogni-

tion. Names made no difference to the East-West Express experiment of Section 5.1 (Appendix H), whose outcome is not in the training data.

The validation also supplies the criterion that separates two virtues an engine can have. On reproducing the observed 1986 levels of the city, the closed-form engine is strongest, with correlations of 0.93 for both residents and workplaces, where V5 reaches 0.80 and 0.77 (Appendix K). An engine calibrated to reproduce observed levels is not the engine that best predicts how those levels move under a policy, so the two criteria are reported separately, and a policy simulator is asked the second question. The general-equilibrium layer is what makes this experiment possible at all, since without market clearing a population of language-model agents produces choices but not the prices and wages that the historical record contains.

5.3 Robustness and generality

A second rail corridor and a different policy instrument. The contrast does not depend on the rail corridor chosen or on the type of policy. Appendix M (Tables 19 and 20) repeats the comparison on a different geography and on a different policy class. The second corridor is a hypothetical north-south line through Pankow, Prenzlauer Berg, Mitte, and Neukölln that reduces travel time on 287 of the 9,216 origin-destination pairs by 7.9 minutes on average, and on it the contrast of Section 5.1 reappears in full. The structural engines raise prices on average and lower wages while V5 lowers prices on average and raises wages, so the divergence is a property of the engine class rather than of the corridor chosen. The second policy is a zoning change that expands floor-space supply by 20% in eight eastern-periphery Ortsteile. It exercises the H channel rather than travel time and reaches V5 only through the equilibrium price vector, since floor-space supply is not in the prompts. Here all engines agree on the direction, with prices falling and residents rising where supply expands, and V5 responds with roughly twice the dispersion of the structural consensus.

Model family and prompt wording. In LUSim an engine is its decision rule together with the model and prompt that implement it, so we vary each in turn with everything else fixed, at $K = 25$ archetypes (Appendix N, Table 21). Replacing GPT 5.3 with DeepSeek V4 Flash changes the model family, and rewriting the scoring instruction with identical content and output schema varies the prompt on the original model. The equilibrium each configuration converges to is stable, with baseline household placement correlating above 0.99 across both changes, and the peripheral concentration of gains survives both. The response to a small localised policy is more configuration-dependent. The correlation between $\Delta \log Q$ and baseline $\log Q$, -0.17 in the original configuration, strengthens to -0.75 under the model change and reverses to 0.18 under the rewording.

A second city and a finer resolution. A Singapore instance (27 zones, Census 2020 household types, three policy packages, and a two-segment public and private housing market under mortgage-servicing and income-ceiling constraints) runs unchanged, with the structural and hybrid engines both ranking the transit-heavy package first on average utility and on inequality (Appendix O). Berlin at the full 12,309-block resolution runs the structural family to convergence.

6 Conclusion

We present LUSim, a multi-agent spatial-equilibrium simulator in which heterogeneous households choose residence and workplace through a language model and clear housing and labour markets inside an externally calibrated general-equilibrium city, with structural decision rules as baselines under identical conditions. On the Berlin instance LLM-driven households reach a density-relocation regime outside the proportional-shift range of the structural family, and on the reunification of Berlin it ranks first in predicting where observed change concentrated, with district names withheld as well as present. The contrast holds across rail corridors, policy classes, and model families, and the pipeline runs unchanged on a second city and at block resolution. The validation correlations are modest and rest on 55 zones and one event, the marginal response of the LLM-driven engine to a small policy change depends on prompt wording, so its results are reproducible only with the model and prompt pinned, and the agent population is held fixed across years. LUSim supports scenario-based counterfactual prediction under a specified policy, environment, and engine rather than operational forecasting, and is a step toward economic world models (Cong, 2025) coupling infrastructure, heterogeneous households, and equilibrium feedback.

AI USE STATEMENT

Large language models are the object of study in this work and are used to run its experiments. The V5 decision engine makes every household's residence and workplace choice through an LLM call inside the market loop, and the V4 engine elicits per-agent preference weights from an LLM (Sections 3.2 and 4 and Appendix C). The models are GPT 5.3 at `reasoning_effort=low` and, in the sensitivity analysis, DeepSeek V4 Flash. The prompt templates, parsing rules, model identifiers, and per-run configurations are documented in Appendices C and N, and every LLM response used in the paper is released as a cache so that all LLM-driven results reproduce bit-identically without further model calls (Appendices E and J). Among the tasks with required disclosure, generative AI tools were used to implement parts of the simulator and analysis code. They were not used to generate synthetic data sets, to develop the theoretical model or conceptual framework, to formulate or prove mathematical claims, to propose hypotheses, to design the experiments, to clean or reformat the data, or to interpret the results, and translation and qualitative data analysis do not arise in this work. Among the tasks with recommended disclosure, generative AI tools were used to edit software code and to draft and edit parts of the text. All AI-assisted code was verified and tested by the authors, and all AI-assisted text was checked by the authors against the underlying results. We take responsibility for the final content of this work, including text, claims, and artefacts produced with the aid of generative AI.

ETHICS STATEMENT

The simulator uses no human subjects and no individual-level data. The synthetic agent population is sampled from public census-scale aggregates, and the Berlin fundamentals are a derived aggregation of the publicly hosted Ahlfeldt et al. (2015) replication data. Foreseeable risks are policy misuse, for instance conflating rent-gradient flattening with displacement-protective gentrification, demographic and place-prior biases inherited from the foundation model, and the environmental footprint of LLM inference. LUSim is a scenario-based counterfactual-prediction tool rather than an operational forecasting system, and the package documentation recommends against commercial deployment as an investment-decision tool without re-validation on the target city. Further discussion is in Appendix I.

REPRODUCIBILITY STATEMENT

Every result in the paper reproduces without LLM credits or API keys from the anonymised mirror and the bundled response caches. Section 4 and Appendices A and B specify the instance, the structural parameters, the agent population, and the policy constructions. Appendix C records seeds, prompt templates, model identifiers, token counts, and wall-clock. Appendix K documents the 1986 instance, the scoring rule, and every configuration choice of the historical validation. Appendices M, N, and O document the additional shocks, the sensitivity configurations, and the Singapore and block-level runs, and Appendix E lists the cache hosting and the SHA-256 manifest. The programmatic entry point and the one-line engine swap are in Appendix F, and the cache-replay reproduction of the headline V5 numbers completes in five to ten minutes on commodity hardware (Appendix J).

REFERENCES

Gabriel M. Ahlfeldt, Stephen J. Redding, Daniel M. Sturm, and Nikolaus Wolf. The economics of density: Evidence from the Berlin Wall. *Econometrica*, 83(6):2127–2189, 2015.

Treb Allen and Costas Arkolakis. Trade and the topography of the spatial economy. *Quarterly Journal of Economics*, 129(3):1085–1140, 2014.

Patrick Bayer, Fernando Ferreira, and Robert McMillan. A unified framework for measuring preferences for schools and neighborhoods. *Journal of Political Economy*, 115(4):588–638, 2007.

Pietro Bini, Lin William Cong, Xing Huang, and Lawrence J. Jin. Behavioral economics of AI: LLM biases and corrections. Working Paper 34745, National Bureau of Economic Research, 2026.

Yvonne M. M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge, MA, 1975.

Nicolas Bougie and Narimasa Watanabe. CitySim: Modeling urban behaviors and city dynamics with large-scale LLM-driven agent simulation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2025.

Ayush Chopra, Arnau Quera-Bofarull, Aditya Subramanian, et al. AgentTorch: Differentiable agent-based modeling for population-scale simulation, 2024. AAMAS 2024 and follow-up arXiv preprint.

Lin William Cong. Economic world models and data-driven generative equilibria. *Cornell Legal Studies Research Paper (forthcoming)*, 2025.

Lin William Cong, Yanhong Guo, Xin Zhao, and Wenjun Zhou. Writing quality and soft information in the genai age: Evidence from online credit markets. *Available at SSRN 4959535*, 2024.

Gilles Duranton and Diego Puga. Micro-foundations of urban agglomeration economies. In J. Vernon Henderson and Jacques-François Thisse (eds.), *Handbook of Regional and Urban Economics*, volume 4, pp. 2063–2117. Elsevier, Amsterdam, 2004.

Daocheng Fu, Wenjie Lei, Licheng Wen, Pinlong Cai, Song Mao, Min Dou, Botian Shi, and Yu Qiao. LimSim++: A closed-loop platform for deploying multimodal LLMs in autonomous driving. In *IEEE Intelligent Vehicles Symposium (IV)*, 2024.

Stephan Heblich, Stephen J. Redding, and Daniel M. Sturm. The making of the modern metropolis: Evidence from London. *Quarterly Journal of Economics*, 135(4):2059–2133, 2020.

Luke Hewitt, Adarsh Ashokkumar, Isaias Ghezae, and Robb Willer. Predicting results of social science experiments using large language models. Working paper, Stanford University, 2024.

Andreas Horni, Kai Nagel, and Kay W. Axhausen (eds.). *The Multi-Agent Transport Simulation MATSim*. Ubiquity Press, London, 2016.

John J. Horton, Apostolos Filippas, and Brennan S. Manning. Large language models as simulated economic agents: What can we learn from Homo Silicus? Working Paper 31122, National Bureau of Economic Research, 2023.

Brennan S. Manning and John J. Horton. General social agents. Working Paper 34937, National Bureau of Economic Research, 2026.

Alfred Marshall. *Principles of Economics*. Macmillan, London, 1890.

Thierry Mayer and Corentin Trevien. The impact of urban public transportation: Evidence from the Paris region. *Journal of Urban Economics*, 102:1–21, 2017.

Daniel McFadden. Modelling the choice of residential location. In Anders Karlqvist, Lars Lundqvist, Folke Snickars, and Jörgen Weibull (eds.), *Spatial Interaction Theory and Planning Models*, pp. 75–96. North Holland, Amsterdam, 1978.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jingyi Yuan Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jianxi Su, and Yong Li. AgentSociety: Large-scale simulation of LLM-driven generative agents advances understanding of human behaviors and society. arXiv preprint arXiv:2502.08691, 2025.

Stephen J. Redding and Esteban Rossi-Hansberg. Quantitative spatial economics. *Annual Review of Economics*, 9:21–58, 2017.

Jennifer Roback. Wages, rents, and the quality of life. *Journal of Political Economy*, 90(6):1257–1278, 1982.

Sherwin Rosen. Wage-based indexes of urban quality of life. In Peter Mieszkowski and Mahlon Straszheim (eds.), *Current Issues in Urban Economics*, pp. 74–104. Johns Hopkins University Press, 1979.

Paul Waddell. UrbanSim: Modeling urban development for land use, transportation, and environmental planning. *Journal of the American Planning Association*, 68(3):297–314, 2002.

Ziyi Yang, Zhi Zhang, Zheng Zheng, et al. OASIS: Open agent social interaction simulations with one million agents. arXiv preprint arXiv:2411.11581, 2024.

Siqi Zheng and Matthew E. Kahn. Land and residential property markets in a booming economy: New evidence from Beijing. *Journal of Urban Economics*, 63(2):743–757, 2008.

A AHLFELDT ET AL. (2015) MODEL DERIVATION

The replication-pack calibration values used verbatim in the Berlin instance are $\beta = 0.75$, $\alpha = 0.8$, $\varepsilon = 6.694$, $\kappa\varepsilon = 0.0987$ (the gravity-equation coefficient on τ_{ij}, equivalently $\kappa = \kappa\varepsilon/\varepsilon \approx 0.0147$ as the per-minute commute-cost decay), $\lambda = 0.071$, $\delta = 0.362$, $\eta = 0.155$, and $\rho = 0.759$.

The Ahlfeldt et al. (2015) within-city framework comprises a household side and a firm side that close in joint residence-workplace pair Fréchet integration. Households of type θ have Cobb-Douglas preferences $U(c, h) = c^{\beta}h^{1-\beta}$ over a tradable consumption good c and residential floor area h, where $\beta \in (0, 1)$ is the consumption-good share and $(1 - \beta)$ is the housing share. In the Berlin replication-pack calibration $\beta = 0.75$. The consumption good is the numéraire, so wages and rents are denominated in units of the consumption good. Conditional on a (residence i, workplace j) pair, the household earns wage w_j at the workplace and pays housing price Q_i per unit of housing at the residence. Indirect utility from optimal expenditure shares is

$$V_{ij}(\theta) = B_i \cdot \beta^{\beta}(1 - \beta)^{1-\beta} \cdot \frac{w_j}{Q_i^{1-\beta}} \cdot e^{-\kappa\tau_{ij}} \cdot \xi_{ij}, \tag{7}$$

where B_i is the residential amenity, $e^{-\kappa\tau_{ij}}$ is the iceberg-style commute-cost discount with per-minute decay κ, and $\xi_{ij} \sim \text{Fréchet}(\varepsilon)$ is an idiosyncratic preference shock over (i, j) pairs. The constant $\beta^{\beta}(1 - \beta)^{1-\beta}$ is absorbed into the units of B_i since it does not vary across (i, j) and is therefore not a coefficient on any choice-relevant variable. Taking logs of indirect utility (Eq. (7)) yields the form in Section 4:

$$\ln u_{ij} = \ln B_i + \ln w_j - (1 - \beta) \ln Q_i - \kappa\tau_{ij} + \xi_{ij}, \tag{8}$$

where ξ_{ij} in this additive log-form is the log of the multiplicative Fréchet shock from Eq. (7), hence Gumbel-distributed. In this form the wage carries a unit coefficient because the consumption good is the numéraire (a one-percent rise in w_j raises real income by one percent), and the rent carries the housing-share coefficient $-(1 - \beta)$ because the household spends share $(1 - \beta)$ of income on floor area (a one-percent rise in Q_i reduces real income by $(1 - \beta)$ percent). The asymmetry between the wage and rent coefficients is therefore a substantive feature of the Cobb-Douglas indirect utility under a numéraire choice, not a normalisation artefact. The Fréchet integration over ξ_{ij} produces the closed-form gravity probability

$$P_{ij} \propto B_i^{\varepsilon} \cdot w_j^{\varepsilon} \cdot Q_i^{-\varepsilon(1-\beta)} \cdot e^{-\varepsilon\kappa\tau_{ij}}, \tag{9}$$

which V1's softmax engine evaluates analytically. In the engine equations (3)–(6) and (1) of Sections 4 and 3.2, i indexes the agent type (rather than the residence zone) and $a = (r, w) \in \mathcal{Z} \times \mathcal{Z}$ denotes the joint residence-workplace action. The per-action shock ξ_a is i.i.d. across actions, equivalent to ξ_{ij} in Eq. (2) under $a = (i, j)$. The deterministic log-utility $\ln u_{ij}$ in Eq. (2) carries $\kappa \cdot \tau_{ij}$, where κ is the per-minute commute-cost decay. Multiplication by the Fréchet shape ε inside the V1 softmax (Eq. (3)) yields the standard Ahlfeldt commuter-flow elasticity $\kappa\varepsilon = \varepsilon\kappa$ as the gravity-equation coefficient on τ_{ij}. Firms operate Cobb-Douglas production $Y_j = A_j HM_j^{\alpha} M_j^{1-\alpha}$ with labour share $\alpha = 0.8$, where HM_j is workplace employment (the labour input) and M_j is commercial floor at j (the capital input). This matches Ahlfeldt et al. (2015) Eq. (10) under their notation $H_{Mj} = HM_j$ and $L_{Mj} = M_j$. The wage first-order condition derived from $\partial Y_j / \partial HM_j$ is $w_j = \alpha A_j HM_j^{\alpha-1} M_j^{1-\alpha}$. Floor supply is generated by the construction-cost rule $\text{total_floor_area} = V \times K^{0.75}$ with $\rho = 0.759$ documenting the elasticity of floor-space supply to land rent. The two-market tatonnement updates Q and w simultaneously with the damping rule $Q^{(k+1)} = 0.25 \cdot Q^{e} + 0.75 \cdot Q^{(k)}$ and $w^{(k+1)} = 0.25 \cdot w^{e} + 0.75 \cdot w^{(k)}$, and convergence on $\|(Q, w)^{(k+1)} - (Q, w)^{(k)}\|_{\infty} < 10^{-3}$

is enforced. The agglomeration externalities $A_j = a_j \Upsilon_j^{\lambda}$ and $B_i = b_i \Omega_i^{\eta}$ recompute every iteration in all five engines (`endogenous_agglomeration: true` in the bundled YAML), where $a_j, b_i > 0$ are the location's exogenous productivity and amenity fundamentals (Ahlfeldt 2015 primitives) and the spillover kernels are $\Upsilon_j = \sum_k e^{-\delta \tau_{jk}} HM_k / \text{area}_k$ (workplace-density) and $\Omega_i = \sum_k e^{-\rho \tau_{ik}} HR_k / \text{area}_k$ (residential-density). The full derivation is in Ahlfeldt et al. (2015) Section 3.

B BERLIN DATA PREPARATION

The 12,309-block-to-96-Ortsteile crosswalk is constructed by spatial join of the Ahlfeldt block centroids against the Berlin Geoportal Ortsteile boundary polygons. The conserving-sum aggregation rules declare floor area, residence count, and workplace count as additive, productivity A and amenity B as residence-weighted means, and travel-time τ as a population-weighted mean for the off-diagonal entries plus a self-time of two minutes for the diagonal. The East-West Express shock specification declares the four stations (Marzahn, Lichtenberg, Mitte, Charlottenburg) and the inter-station nominal travel time of five minutes, with the shortest-path update applied to the baseline τ matrix detailed below. The crosswalk, aggregation rules, and shock specification ship as YAML and CSV files under `data/berlin/` in the codebase. The agent-population provenance is summarised in Table 4.

Table 4: Provenance of the ten attributes encoded in `data/berlin/agents/berlin_ortsteile_richer_10k.yaml`. The three attributes flagged Synthetic-fallback are assigned uniform-by-Bezirk because the public Berlin Statistics Office release used at the data-fetch date in April 2026 did not include an Ortsteil-resolution table for them. Richer-resolution sources (notably Berlin's Sozialatlas and the SOEP regional aggregates) were not yet integrated into the data pipeline and constitute a known direction for a future revision rather than a hard data-availability limit. Income is raw monthly net household income at 2011 Berlin Zensus brackets, not equivalised. The home Ortsteil is sampled with weight proportional to the residence mass `emprsd_06` from the Ahlfeldt replication pack. The joint distribution is constructed via iterative proportional fitting against the available marginals, with cells of zero or near-zero marginal support Laplace-smoothed at $\alpha = 0.5$ before iteration following the standard small-area-estimation prior choice (Bishop et al., 1975).

Attribute	Source	Resolution	Synthetic-fallback
Head age and sex	Berlin Zensus 2011	Ortsteil	No
Household size	Berlin Zensus 2011	Ortsteil	No
Income	Berlin Statistics Office (Bezirk income distribution)	Bezirk	Yes (uniform)
Has children	Berlin Zensus 2011	Ortsteil	No
Has elderly	Berlin Zensus 2011	Ortsteil	No
Car ownership	Berlin Senate Department for Urban Development	Bezirk	Yes (uniform)
Education	Berlin Zensus 2011	Ortsteil	No
Migration background	Berlin Zensus 2011	Ortsteil	No
Employment status	Berlin Statistics Office	Bezirk	Yes (uniform)
Tenure (own/rent)	Berlin Zensus 2011 GWZ	Ortsteil	No

East-West Express shock: shortest-path update. The shock recomputes the bilateral travel-time matrix τ on the network augmented with the four-station rail subgraph, the standard within-city counterfactual-network construction in quantitative spatial equilibrium (Allen & Arkolakis, 2014; Heblich et al., 2020). Let $\{a_1, \ldots, a_K\} \subset \{1, \ldots, N\}$ denote the $K = 4$ station Ortsteile zone indices, ordered along the rail line as Marzahn, Lichtenberg, Mitte, Charlottenburg. Define the cumulative in-vehicle rail time between stations as $T_{\text{rail}}(p, q) = |p - q| \cdot \delta$ minutes, with $\delta = 5$ the inter-station nominal time and $p, q \in \{1, \ldots, K\}$ station indices along the line. For each origin-destination pair $(i, j) \in \{1, \ldots, N\}^2$,

$$\tau_{ij}^{(s)} = \min\Big(\tau_{ij},\ \min_{p,q \in \{1,\ldots,K\}} \big[\, \tau_{i,a_p} + T_{\text{rail}}(p, q) + \tau_{a_q,j} \,\big]\Big), \tag{10}$$

Table 5: Reproducibility tiers for the Berlin Ortsteile instance. The cache-replay path is the recommended option for reviewers and replicators without LLM credits. The bundled Hugging-Face-hosted cache contains approximately 82,000 prompt-response pairs and decompresses to roughly 320 MB. The Provider column records the LLM provider and reasoning-effort setting used to generate the bundled cache.

Tier	Variant	Wall-clock	LLM credits	Provider	Determinism
1	Install only	<30 s	None	N/A	N/A
2	Quickstart V1	<1 min	None	N/A	Bit-identical
3a	V1 baseline + shock	∼3 hr	None	N/A	Bit-identical
3b	V2, V3 each	∼3 hr	None	N/A	Seeded (seed=42)
3c	V4 cache-replay	∼3 hr	None (cache)	GPT 5.3, reasoning=low	Bit-identical
3d	V4 from scratch	∼3 hr	∼$5 (GPT 5.3)	GPT 5.3, reasoning=low	Provider
4	V5 cache-replay	∼5–10 min	None (cache)	GPT 5.3, reasoning=low	Bit-identical
4	V5 from scratch	∼10 hr	$30–50	GPT 5.3, reasoning=low	Provider

i.e., the original travel time, or the minimum over boarding station p and alighting station q of (origin-to-a_p) + (rail a_p-to-a_q) + (a_q-to-destination), whichever is shorter. The procedure is a single-step shortest-path relaxation through the K station nodes, computed in $O(K^2N^2)$ time as two $(\min, +)$ matrix multiplications: $b_{i,q} = \min_p (\tau_{i,a_p} + T_{\text{rail}}(p, q))$ followed by $\tau_{ij}^{(s)} = \min(\tau_{ij}, \min_q (b_{i,q} + \tau_{a_q,j}))$. Symmetry of τ is preserved if τ is initially symmetric, and the operation is idempotent. The reference implementation is `src/agent_urban_planning/research/berlin/railway_shock.py`.

C Per-engine implementation details and prompt templates

Compute, seeds, and provider pinning. V1, V2, V3 run on a single Apple M-series workstation (10 cores, 64 GB RAM). V4 and V5 use the same workstation with LLM calls dispatched to the GPT 5.3 API. V1 is deterministic. V2, V3, V4 are seeded-stochastic at the single pinned `seed=42`, matching the Ahlfeldt et al. (2015) replication-pack convention. V5 is cache-deterministic against the bundled GPT 5.3 snapshot (bit-identical replay against the cache). The bundled V5 cache was generated against GPT 5.3 with `reasoning_effort=low`. V5 issued 81,844 prompts ($\sim 1.45 \times 10^8$ input tokens, $\sim 1.87 \times 10^7$ output tokens) at an estimated billed cost of forty-two US dollars and 9.7 hours wall-clock at concurrency 15. V4 issued 402 elicitation prompts at approximately five US dollars; the resulting per-demographic-profile preference cache (402 JSON entries, $\sim$ 24 KB total) is bundled at `data/berlin/llm_cache_v4/` for bit-identical Tier 3c reproduction without LLM credits.

Reproducibility tiers. The simulator ships in four tiers (Table 5) ranging from a thirty-second installation check through full V5 cache-replay. The cache-replay tier is the recommended option for reviewers and replicators without LLM credits.

Full cross-variant moments table. Table 6 extends the abridged main-text Table 2 with five additional columns ($\sigma_{\Delta \log w}$, $\Delta\bar{W}$, $\langle U \rangle_0$, $\Delta\langle U \rangle_{\text{own}}$). All entries are single-seed point estimates: V1 is deterministic, V2, V3, V4 are at `seed=42` matching the Ahlfeldt et al. (2015) replication-pack convention, and V5 is at the bundled GPT 5.3 cache (bit-identical replay). V5's $\Delta\langle U \rangle_{\text{own}}$ entry is the percentage drop in the LLM's self-rated 0–100 satisfaction score, qualitatively comparable to the surrounding rows though the unit is architecture-specific.

Baseline equilibrium fit. Table 7 reports the per-zone Pearson correlation between simulated and observed 2006 quantities (Q, HR, HM, w) across the 96 Ortsteile, before the East-West Express shock. V1 is calibrated against the Ahlfeldt 2015 replication pack and serves as the structural reference. V2, V3, V4 share V1's Cobb-Douglas and Fréchet inner utility and converge to close neighbours of V1's baseline equilibrium. V5 has no structural calibration constraint, so its baseline differs notably from the 2006 target ($r_Q = -0.24$ rather than 0.45, $r_{HR} = 0.72$ rather than 0.93); the paper's headline cross-variant comparison is therefore on the shock-response moments in

Table 6: Full cross-variant Berlin shock moments. Distribution-shape moments are across the 96 Ortsteile. $\Delta\bar{Q}$ and $\Delta\bar{W}$ are the level means of post-shock housing-price and employment-weighted-wage changes across the 96 Ortsteile. $\langle U\rangle_0$ is the population-weighted average per-agent utility at the baseline (pre-shock) equilibrium under each engine's own welfare formula, and the Δ rows below are computed against this reference. $\Delta\langle U\rangle_{V1}$ applies the V1 (Baseline-softmax) welfare formula uniformly to each variant's output state. $\Delta\langle U\rangle_{\text{own}}$ applies each variant's own welfare formula, with V5's entry the percentage drop in the LLM's self-rated 0–100 satisfaction score. All entries are single-seed point estimates: V1 is deterministic, V2, V3, V4 are at `seed=42`, and V5 is at the bundled GPT 5.3 cache (bit-identical replay).

Moment	V1 softmax	V2 ABM	V3 Normal	V4 Hybrid	**V5 LLM**
$\mu_{\Delta \log Q}$	0.0004	0.0004	0.0004	0.0004	**0.0016**
$\sigma_{\Delta \log Q}$	0.0070	0.0071	0.0042	0.0068	**0.0083**
$\mu_{\Delta \log w}$	−0.0001	−0.0001	−0.0001	−0.0001	**−0.0003**
$\sigma_{\Delta \log w}$	0.0018	0.0018	0.0010	0.0017	**0.0021**
$p_{95}\lvert\Delta \log Q\rvert$	0.0083	0.0080	0.0048	0.0077	**0.0172**
$\Delta Y\%$	0.0007	−0.0000	0.0132	0.0002	**0.0299**
$\Delta\bar{W}$	0.0003	0.0003	0.0005	0.0003	**0.0004**
$\Delta\bar{Q}$	0.0008	0.0007	0.0001	0.0007	**0.0111**
$\langle U\rangle_0$	−2.5894	−2.5895	−2.5498	−2.5905	**−2.7448**
$\Delta\langle U\rangle_{V1}$	−0.0032	−0.0031	−0.0027	−0.0031	**−0.0056**
$\Delta\langle U\rangle_{\text{own}}$	−0.0032	−0.0031	−0.0029	−0.0030	**−0.0051**

Table 2, which are deltas against each engine's own baseline. The companion supplementary file `comparison_baseline_fit.csv` reports the matching relative-RMSE values.

Table 7: Baseline equilibrium fit to observed 2006 targets across the 96 Ortsteile (Pearson correlation r between simulated and observed quantities). V1, V2, V4 cluster tightly because they share the calibrated Cobb-Douglas and Fréchet inner utility. V3 is a mild Gaussian-tail outlier. V5 is the most-departing case, consistent with its non-structural decision rule.

Variant	r_Q	r_{HR}	r_{HM}	r_w
V1 softmax	0.45	0.93	0.88	0.95
V2 ABM	0.45	0.93	0.89	0.95
V3 Normal	0.44	0.92	0.81	0.92
V4 Hybrid	0.44	0.93	0.89	0.95
V5 LLM	**−0.24**	**0.72**	**0.89**	**0.88**

No-op sanity checks. The framework satisfies the elementary "no-policy yields no change" invariant by construction. (i) V1 (Baseline-softmax) is closed-form deterministic, so an identity shock $\tau^{(s)} = \tau$ produces zero per-zone deltas in (Q, w, HR, HM) across all 96 Ortsteile. (ii) V5 in cache-replay mode is bit-deterministic against the bundled GPT 5.3 cache; an identity shock against the same cache subdirectory produces zero deltas. (iii) V2, V3, V4 are seeded-stochastic at `seed=42`; running baseline and an identity shock from the same seed shares the RNG state and produces zero deltas across all zones.

Per-engine implementation. V1 (Baseline-softmax) evaluates P_{ij} analytically for each agent type and aggregates to per-zone residence and workplace masses by type-weighted sums. V2 (Baseline-ABM) draws $\xi_{ij} \sim \text{Fréchet}(\varepsilon)$ per agent and returns the argmax pair, and V3 substitutes $\xi_{ij} \sim \mathcal{N}(0, \pi^2/6)$ to match the Fréchet (Gumbel-base) variance. V4 (Hybrid-elicited) queries the LLM once per unique demographic profile (the 10,000 agent types resolve to 402 unique profiles) with a preference-elicitation prompt that asks for four 0–100 integers (rent, commute, amenity, wage). After population-weighted per-axis renormalisation across the unique profiles, the elicited integers become per-agent scaled multipliers $\hat{\phi}_i = (\hat{\phi}_i^Q, \hat{\phi}_i^\tau, \hat{\phi}_i^B, \hat{\phi}_i^w)$ that modulate the inner-utility coefficients on Q, τ, B, and w. V4's elicitation prompt is in Listing 1 below. V5 (LLM-decide) executes a two-stage decision loop per cluster per market iteration: stage 1 scores all 96 zones

for residence given a persona and the current price vector, and stage 2 conditions on a top-K=10 residence shortlist and selects a workplace. Both V5 stages are in Listings 2 and 3.

Listing 1: V4 preference-elicitation prompt template, abridged.

```
You are a Berlin household with the following demographics: <persona>.
Rate the importance of each factor on a 0-100 scale: monthly rent
budget, commute time, neighbourhood amenity, workplace wage. Return
exactly four integers, comma-separated, in the order above.
```

Listing 2: V5 stage-1 residence-scoring prompt template, abridged.

```
You are a Berlin household with the following demographics: <persona>.
The current housing prices and travel times to a workplace cluster
are: <96 zone rows: zone name, rent index, commute minutes>. Score
each of the 96 zones on a 0-10 attractiveness scale. Return one
integer per zone, comma-separated, in the same order.
```

Listing 3: V5 stage-2 workplace-given-residence prompt template, abridged.

```
You are a Berlin household with the following demographics: <persona>,
choosing where to work given residence in <selected zone>. Wages and
commute times to the top-10 workplace candidates are: <10 rows>.
Pick one workplace and return its zone identifier.
```

V5 decision protocol. For each cluster $c \in \{1, \ldots, K{=}50\}$ at each market iteration k, V5 runs:

- **Inputs.** Persona θ_c, current housing prices $Q^{(k)}$ and wages $w^{(k)}$, travel-time matrix τ, and the 96 candidate Ortsteile.
- **Stage 1.** The LLM scores all 96 zones as residences (Listing 2); the top K_R=10 residences are passed forward.
- **Stage 2.** For each shortlisted residence, the LLM selects a workplace (Listing 3).
- **Output.** Residence-workplace pair $(i^\star, j^\star)$ plus the LLM's 0–100 self-rated satisfaction score s_c that V5 uses as its own welfare ruler.
- **Aggregation.** Broadcast $(i^\star, j^\star)$ to every type in cluster c under population weights w_i, then clear Q and w via the same two-market tâtonnement as V1–V4.
- **Cache key.** $(c, \text{stage}, \text{residence_name}, \text{prompt_version}, \text{bucket}(Q), \text{bucket}(w))$, partitioned into `baseline/` and `shock/` subdirectories.

The full prompt templates, including the system prompts, the rebalance instruction, and the parsing rules, are in `src/agent_urban_planning/llm/prompts/`.

D ADDITIONAL PER-ZONE SHOCK-EFFECT FIGURES

This appendix complements the per-zone $\Delta \log Q$ choropleth in the Results (Figure 2) with the $\Delta \log w$ surface (Figure 3) and the corresponding $\Delta \log HR$, $\Delta \log HM$, and $\Delta \log Y$ surfaces (Figures 4, 5, 6) and a cross-measure $\Delta \log$ vs log-baseline scatter (Figure 7). Together they characterise the full per-zone equilibrium tuple $R^{(s)} = (Q^\star, w^\star, HR^\star, HM^\star)$ released for each engine V1–V5, with per-zone output Y_j measured as $A_j \cdot HM_j^\alpha$ at $\alpha = 0.8$ (Section 4). For the per-zone surface and the Table 2 aggregate $\Delta Y\%$, both A_j and the commercial floor input M_j are held at their baseline-calibrated values, with $M_j^{1-\alpha}$ absorbed into the per-zone normalisation, so $\Delta \log Y_j = \alpha \cdot \Delta \log HM_j$. Each map uses the same six-panel layout as Figures 2–3: the top-left panel shows the observed 2006 baseline value (sequential viridis colormap), and the other five panels show $\log(\text{shock}/\text{baseline})$ for each engine variant on a shared diverging RdBu_r scale clipped at the 95th percentile of $|\Delta \log|$.

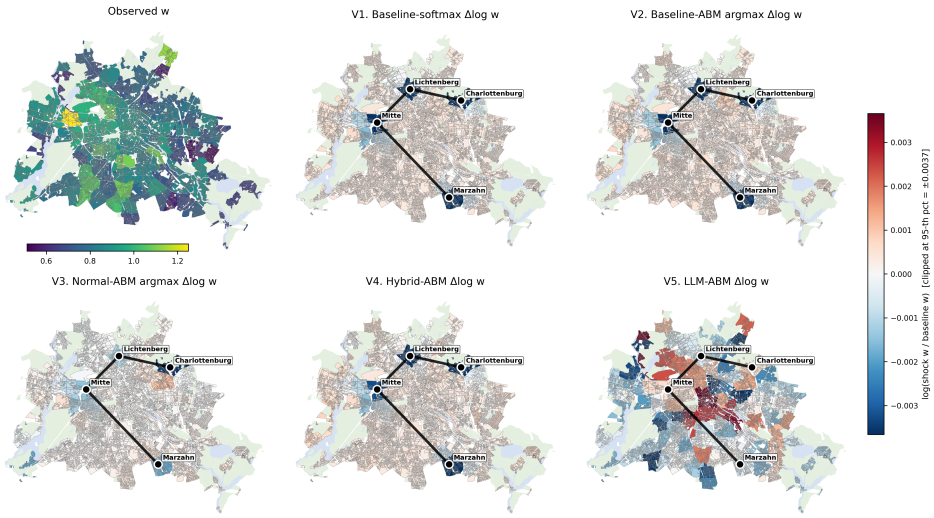

Figure 3: Per-zone $\Delta \log w$ choropleths for V1 through V5 under the East-West Express shock. The structural engines drop wages in central Berlin (compensating differential), while V5 raises them (labour-market pooling); the mechanism decomposition in Section 5.1 and Appendix L separates the two channels.

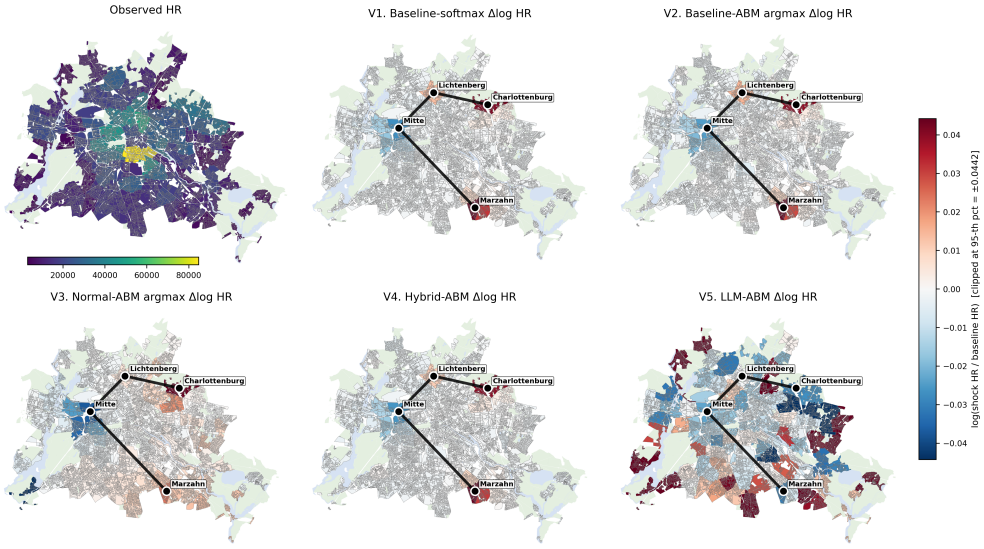

Figure 4: Per-zone $\Delta \log HR$ choropleths (residential employment density — workers by residence) for V1 through V5 under the East-West Express shock. Top-left: observed 2006 baseline. V5 (LLM-ABM) produces the largest peripheral residential gains, consistent with the non-smooth density relocation discussed in Section 5. The structural variants V1–V4 produce smooth proportional shifts in the same neighbourhood as their $\Delta \log Q$ surfaces (Figure 2).

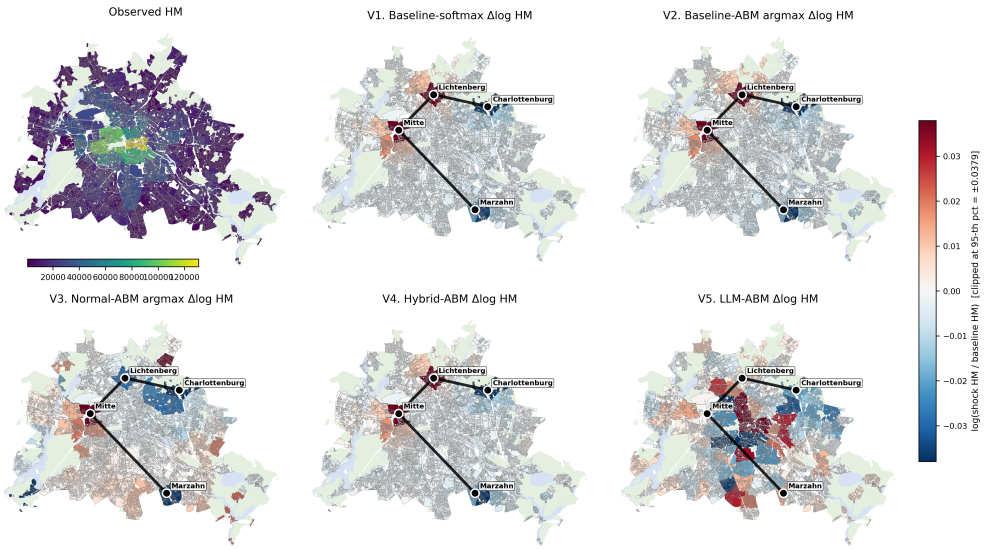

Figure 5: Per-zone $\Delta \log HM$ choropleths (workplace employment density — workers by workplace) for V1 through V5. The structural variants concentrate the workplace response in or near the four station zones, while V5 reallocates workplace mass more broadly across central-eastern corridors with a centre-versus-periphery sign pattern.

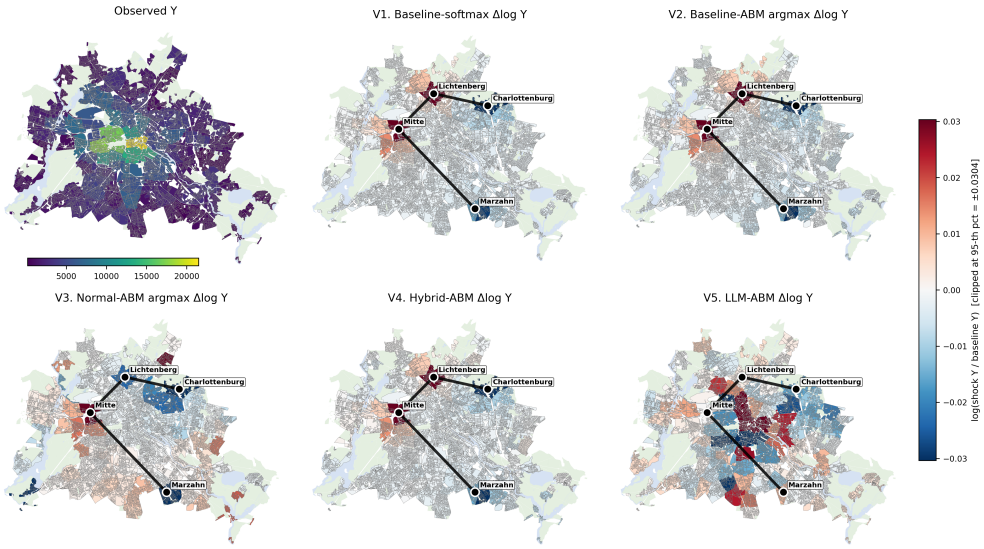

Figure 6: Per-zone $\Delta \log Y$ choropleths under the firm Cobb-Douglas $Y_j = A_j \cdot HM_j^{\alpha} \cdot M_j^{1-\alpha}$ with $\alpha = 0.8$ and both A_j and the commercial floor M_j held at their baseline-calibrated levels (matching the Table 2 $\Delta Y\%$ measurement, which absorbs $M_j^{1-\alpha}$ into the per-zone normalisation). At fixed A_j and M_j, the per-zone surface equals $\alpha \cdot \Delta \log HM_j$. The aggregate $\Delta Y\%$ in Table 2 weights these per-zone log-changes by zone-level Y_j magnitudes, which is why V5's aggregate $\Delta Y\% = 0.0299$ is approximately fifteen times the V1, V2, V4 structural pool despite per-zone responses of similar order across most non-station zones.

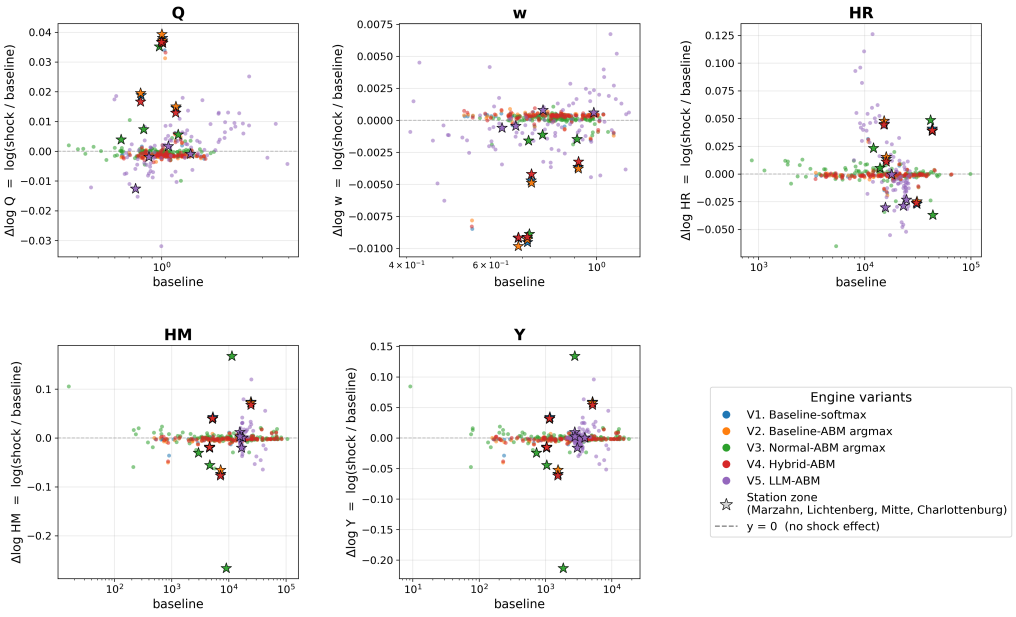

Figure 7: Per-zone $\Delta \log$ vs log baseline for the five measures (Q, w, HR, HM, Y), with 96 zones $\times$ 5 engine variants per panel (one colour per variant). The dashed $y = 0$ horizontal line is the no-shock-effect reference; vertical deviation from this line is the per-zone shock impact. Star markers identify the four East-West Express station zones (Marzahn, Lichtenberg, Mitte, Charlottenburg).

E V5 LLM CACHE HOSTING AND SHA-256 MANIFEST

The V5 LLM cache is hosted on Hugging Face Datasets at `https://huggingface.co/datasets/aup-anon-2026/LUSim-V5-LLM-cache-Berlin-ICLR2027` (anonymised for double-blind review). The public-account URL will be substituted in the camera-ready version. The cache is distributed as a single gzipped tar archive `llm_cache_v5.tar.gz` of size approximately 15.4 MB compressed (320 MB uncompressed). The archive contains two top-level directories `baseline/` and `shock/`, holding approximately 54,784 and 27,060 JSON files respectively (81,844 total) keyed by `<persona-cluster-id>_<iteration>_<stage>.json`. The Tier 4 reproduction script invokes V5 with `--no-llm`, which fails fast on a cache miss rather than falling through to a live LLM call, ensuring no LLM credits are accidentally consumed during a routine reproduction run.

The SHA-256 manifest for the bundled artefacts is in Table 8.

Table 8: SHA-256 manifest of the two bundled artefacts (full hashes follow the table).

File	SHA-256
`llm_cache_v5.tar.gz`	`7f94ea6c...64cb868e`
`data/berlin/scenarios/berlin_2006_ortsteile.yaml`	`c8b9f0ee...3dfc8fae`

Full SHA-256 hashes (truncated above for layout):

- `llm_cache_v5.tar.gz`: `7f94ea6c56c4d2894f9b63aa8a61e91780d14fddedf24d3f4300a55264cb868e`
- `berlin_2006_ortsteile.yaml`: `c8b9f0ee57d68b633e72a4770dda2c04a7e700b11fccdbb09ffa20f03dfc8fae`

The nine additional V5 caches of Appendices K, M, N, and H.3 (the 1986 baseline and reunified runs, the names-withheld pair, the north-south corridor, the floor-space expansion, the DeepSeek and reworded-prompt runs, and the $K = 25$ and $K = 75$ sweeps) are hosted in the same dataset, one tarball per run, and extract under `data/berlin/llm_cache_v5/`. Their truncated SHA-256 hashes are in Table 9, and the full hashes and extraction paths are in `data/berlin/llm_cache_v5/MANIFEST.md` in the mirror.

Table 9: The nine additional V5 caches hosted in the same dataset, with truncated SHA-256 hashes and the number of files in each tarball.

File	SHA-256	Files
`llm_cache_v5_1986_baseline.tar.gz`	`ea45b8f8...8d79dabb`	5,500
`llm_cache_v5_1986_reunified.tar.gz`	`3edb211a...3a1d8fbd`	5,500
`llm_cache_v5_1986_names_withheld.tar.gz`	`0556108f...066c3a30`	11,000
`llm_cache_v5_north_south.tar.gz`	`06325f82...a951d550`	7,150
`llm_cache_v5_floor_space.tar.gz`	`299c15f2...8e3d70a9`	5,500
`llm_cache_v5_deepseek.tar.gz`	`e776a14c...0b0031cf`	9,075
`llm_cache_v5_paraphrase.tar.gz`	`4d90f02f...42e250c5`	11,000
`llm_cache_v5_k25.tar.gz`	`dc7ab3fc...f84f909b`	27,500
`llm_cache_v5_k75.tar.gz`	`70c14998...e979d9fb`	71,775

F PROGRAMMATIC API WALKTHROUGH

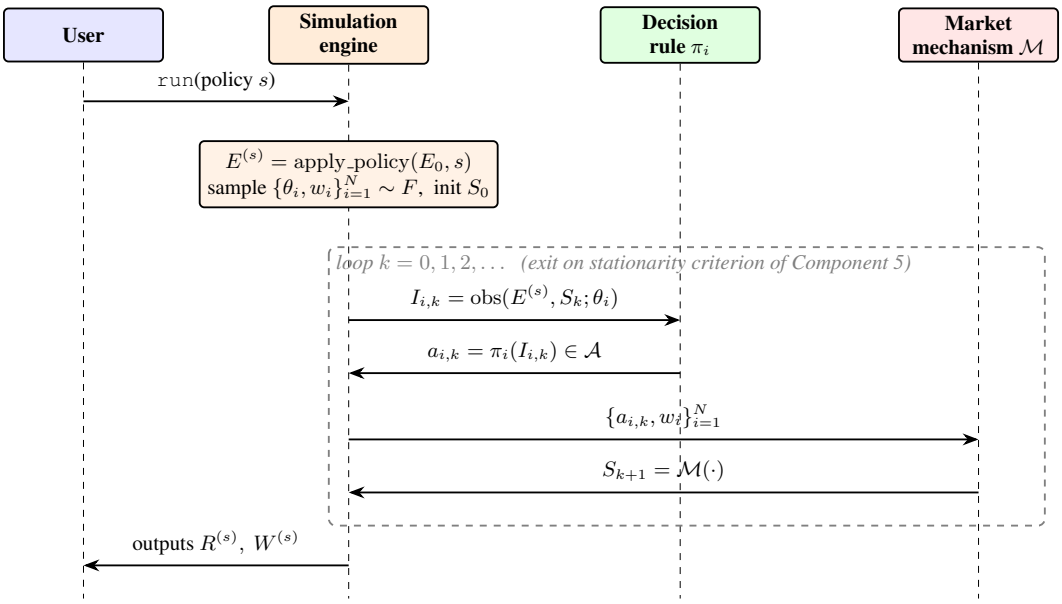

Figure 8: Data-flow sequence diagram for LUSim. The user invokes `SimulationEngine.run` with policy s. The engine applies s to the baseline environment E_0, samples the agent population from the type distribution F, and initialises the endogenous state S_0. It then iterates between the decision rule π_i, which receives the per-agent observation $I_{i,k}$ and returns an action $a_{i,k} \in \mathcal{A}$, and the market mechanism $\mathcal{M}$, which aggregates the actions, clears the markets, and returns the updated state S_{k+1}. The loop runs until the stationarity criterion of Component 5 is met, at which point the fixed point yields the per-zone equilibrium tuple $R^{(s)}$ and the aggregate welfare $W^{(s)}$, returned to the caller.

The five-step LUSim pipeline (Section 3.2) is realised as a Python script of fewer than ten lines. We illustrate with the V1 (closed-form softmax) engine on the calibrated 96-zone Berlin Ortsteile baseline, and the four other engines are obtained by replacing only the constructor on the highlighted line.

```
import agent_urban_planning as lusim
from agent_urban_planning.data.loaders import load_scenario, load_agents

# (i)  Calibrated environment: zones, tau, A, B, H, psi.
scenario = load_scenario("data/berlin/scenarios/berlin_2006_ortsteile.
    yaml")

# (ii) Heterogeneous agent population: N = 10,000 types from public
    Berlin sources.
agents = load_agents("data/berlin/agents/berlin_ortsteile_richer_10k.yaml
    ")

# (iii) Decision engine. SWAP HERE for V2/V3/V4/V5; the rest is invariant
    .
engine = lusim.UtilityEngine(scenario.ahlfeldt_params, mode="softmax")

# (iv) Assemble. Seed pinning for reproducibility.
sim = lusim.SimulationEngine(scenario=scenario, agent_config=agents,
                             engine=engine, seed=42)

# (v)  Run. policy=None gives the baseline equilibrium; pass a Policy
#      object (transit shock, zoning change) for counterfactuals.
result = sim.run(policy=None)
```

The four other engines are constructed with identical surrounding code:

```
engine = lusim.UtilityEngine(scenario.ahlfeldt_params, mode="argmax",
    noise="frechet")  # V2
engine = lusim.UtilityEngine(scenario.ahlfeldt_params, mode="argmax",
    noise="normal")   # V3
engine = lusim.HybridDecisionEngine(elicitor=...)
                  # V4
engine = lusim.LLMDecisionEngine(response_format="score_all",
                                 rebalance_instruction=True,
                                 stage2_top_k_residences=10)
                # V5
```

`result` carries the per-zone equilibrium tuple $R^{(s)} = (Q^\star, w^\star, HR^\star, HM^\star)$, the aggregate welfare $W^{(s)}$, the per-iteration price-and-action trajectory, and a `RunMetadata` payload recording the seed, engine class, LLM provider (where applicable), wall-clock, and cache hit rate. Counterfactual comparison reads off `result.metrics` differences between baseline and shock runs.

G CODE LISTINGS FOR THE FIVE REFERENCE ENGINES

The full code listings for the five reference engines live in the released codebase under `src/agent_urban_planning/decisions/` and exceed 100 lines each, so we describe their dispatch structure here rather than reproduce them in the paper. All engines implement a common `decide_batch` interface that returns a list of residence-workplace assignments per agent type. V1 evaluates the closed-form softmax over the structural log-utility analytically. V2 and V3 add a Fréchet or Gaussian shock to the level utility and return the argmax. V4 elicits per-agent preference weights from an LLM once at the start of clearing (cached by demographic-profile hash, 402 unique queries for the 10,000 agent types) and feeds them into the same closed-form argmax used by V2 and V3. V5 runs a two-stage score-all loop per cluster per market iteration, scoring all 96 zones for residence in stage one and the top-K=10 residences for workplace in stage two. Concrete file paths, class names, and helper-method signatures are documented in the README of the anonymised mirror (Appendix J).

H ROBUSTNESS AND SENSITIVITY STUDIES

This appendix collects three robustness checks complementing the headline East-West Express shock results in Section 5: (i) a uniform-commute monotonicity sanity check across V1–V4, (ii) a leakage ablation that re-runs V5 with anonymised zone names, and (iii) a K-cluster sensitivity sweep at $K \in \{25, 50, 75\}$. All three reuse the paper's seed (42), agent population (1,000,000 Monte Carlo replicates), iteration cap (50), and convergence criteria, while only the perturbed quantity differs.

H.1 MONOTONICITY UNDER A UNIFORM COMMUTE SHOCK

The headline East-West Express shock is corridor-localized (it touches 164 of the 9,216 Origin-Destination pairs). To exercise the pipeline in a regime where every Origin-Destination pair is perturbed identically, we apply a uniform 5-minute reduction $\tau_{ij}^{(s)} = \max(0, \tau_{ij}^{(0)} - 5)$ across all 9,216 pairs and re-clear the market for V1–V4. Engines V2/V3/V4 are based on argmax over $U_i(a; I, \hat{\phi}_i)$ and are therefore mathematically translation-invariant under a uniform additive shift of τ, so the argmax does not change. V1's closed-form softmax inherits the same property at the level of choice probabilities (the uniform $\exp(5\varepsilon\kappa)$ factor cancels in the softmax ratio), but V1's response can be non-zero through endogenous-agglomeration channels (recomputation of $A_j = a_j \Upsilon_j^\lambda$ and $B_i = b_i \Omega_i^\eta$ under the warm-start re-clear).

The check is a non-trivial pipeline-integrity test in which a buggy implementation that broke argmax-translation-invariance (e.g., by inserting an absolute-value commute term inside the decision rule, or a magnitude-dependent affordability filter that did not cancel) would fail this table. It does not. We do not include V5 in this table. V5's prompt explicitly contains τ values, so it carries no structural translation-invariance guarantee and a non-trivial response would not be diagnostic. The classical

Table 10: Shock-response moments under a 5-minute uniform commute reduction. As predicted, the argmax-based engines (V2/V3/V4) honour translation-invariance to $\sim 10^{-4}$ precision, and V1's small response reflects the endogenous-agglomeration channel.

Variant	$\mu_{\Delta \log Q}$	$\sigma_{\Delta \log Q}$	$p_{95}\lvert\Delta \log Q\rvert$	$\mu_{\Delta \log w}$	$\sigma_{\Delta \log w}$	$\mathrm{frac}(\Delta \log Q > 0)$
V1 softmax	-0.0044	0.0318	0.0674	0.0011	0.0080	0.46
V2 ABM	0.0000	0.0000	0.0001	0.0000	0.0000	0.47
V3 Normal	-0.0001	0.0010	0.0007	-0.0000	0.0002	0.64
V4 Hybrid	-0.0000	0.0000	0.0001	-0.0000	0.0000	0.46

monocentric "uniform commute reduction to uniform rent increase" prediction is also not the relevant target here because the Ahlfeldt et al. (2015) framework is polycentric with endogenous joint (residence, workplace) choice, and the monocentric sign prediction is not implied by this model class.

H.2 Leakage ablation: anonymised zone names

One possible concern is that the foundation model's pretraining priors over Berlin geography (e.g., that Marzahn is a peripheral panel-housing district, that Mitte is the historic centre) may be doing load-bearing work in V5's headline output. To test this, we re-run V5 (paper config, $K{=}50$, score-all-96 prompt) with synthetic identifiers `z96_000` through `z96_095` replacing the real Ortsteil names. All numerical inputs (price, wage, amenity, productivity, commute time) are unchanged. If V5's qualitative shock response persists under anonymisation, the reasoning is being driven by the numerical inputs rather than the place names, while if it collapses, the result is partially attributable to place priors.

Table 11: V5 shock-response moments under the East-West Express, with and without zone-name anonymisation. Anonymised V5 produces moments within $\sim 25\%$ of the named V5 baseline along every dimension. Mean values are within sampling noise of zero (consistent with V5's argmax-translation-invariance under polycentric inner utility).

Variant	$\mu_{\Delta \log Q}$	$\sigma_{\Delta \log Q}$	$p_{95}\lvert\Delta \log Q\rvert$	$\mu_{\Delta \log w}$	$\sigma_{\Delta \log w}$
V5 named (paper Table 2)	0.0016	0.0083	0.0172	-0.0003	0.0021
V5 anonymised	-0.0010	0.0079	0.0156	-0.0001	0.0017

Table 12: Per-zone Pearson correlation of V5 equilibrium variables between the named and the anonymised runs. Both the baseline equilibrium and the post-shock equilibrium are essentially identical across the two runs: every $r \geq 0.95$ across all four equilibrium variables and both market states. The four equilibrium variables together summarise the model's full per-zone output, so the column-wise $r \geq 0.95$ pattern indicates that V5 places agents in the same spatial equilibrium with or without real Berlin Ortsteil names in the prompt.

Variable	Baseline (named $\leftrightarrow$ anon)	Shock (named $\leftrightarrow$ anon)
Q (residential housing price)	0.9942	0.9945
HR (residential employment)	0.9782	0.9729
HM (workplace employment)	0.9603	0.9545
w (wage)	0.9980	0.9982

The per-zone correlations in Table 12 are the most direct evidence, where across the four equilibrium variables and both market states, $r \geq 0.95$ between the named and the anonymised runs. The two V5 instantiations therefore land at essentially the same spatial equilibrium regardless of whether real Ortsteil names or synthetic identifiers are presented in the prompt. The shock-response moments in Table 11 are consistent with this picture. Dispersion is preserved ($\sigma_{\Delta \log Q}$ drops by only 5%, $p_{95}\lvert\Delta \log Q\rvert$ by 9%, $\sigma_{\Delta \log w}$ by 19%) and both means are small in absolute terms ($\lvert\mu\rvert < 0.002$, within the LLM's stochastic-sampling band so the mean's sign is not diagnostic, and V5 also inherits its argmax-translation-invariance at the level of the decision rule's mean). The headline V5 result

is therefore driven by V5's reasoning over the numerical inputs, not by the foundation model's pretraining priors over Berlin geography. Two caveats remain. First, the synthetic IDs `z96_XXX` preserve the integer ordering of zone indices, and a stronger anonymisation would shuffle IDs randomly per prompt and is left for future work. Second, this test addresses one specific intervention (replacing names with synthetic IDs), while alternative interventions (shuffled names, names from a different city, anonymised but stylistically rich text) are out of scope here. Section 5.2 applies the same intervention to the historical validation, where the outcome is in the model's training data, and there the names do matter for the housing-price pattern; together the two runs locate the effect of place recognition where the outcome is known to the model and not where it is not.

H.3 K-CLUSTER SENSITIVITY

The paper's headline V5 result uses $K{=}50$ k-means personas summarising the $N{=}10{,}000$ agent types. To test whether the shock-response pattern is sensitive to this choice, we re-run V5 baseline and shock at $K{=}25$ and $K{=}75$. Each alternative run rebuilds its own LLM cache (cluster IDs feed the cache key), and the paper's primary cache at `data/berlin/llm_cache_v5/` is untouched.

Table 13: V5 shock-response moments at three values of K. The mean response is near zero across all three K (consistent with V5's argmax-translation-invariance), and the dispersion $\sigma_{\Delta \log Q}$ decreases monotonically with K, matching the predicted ensemble-averaging effect (more clusters leads to tighter LLM-voice averaging which results in smaller σ).

K	source	$\mu_{\Delta \log Q}$	$\sigma_{\Delta \log Q}$	$p_{95}\lvert\Delta \log Q\rvert$	$\mu_{\Delta \log w}$	$\Delta\bar{Q}$
25	this work	-0.0011	0.0109	0.0217	0.0003	-0.0012
50	**paper primary**	$\mathbf{0.0016}$	$\mathbf{0.0083}$	$\mathbf{0.0172}$	$\mathbf{-0.0003}$	$\mathbf{0.0019}$
75	this work	-0.0005	0.0069	0.0145	0.0002	-0.0011

Table 14: Cross-K Pearson correlation of per-zone V5 baseline equilibrium variables (no shock). Baseline equilibrium is essentially K-invariant: every $r \geq 0.989$ across all four equilibrium variables and all three pairwise comparisons. Mean per-zone ratios are within 0.7% of unity in every cell.

Variable	$K{=}25$ vs $K{=}50$	$K{=}25$ vs $K{=}75$	$K{=}50$ vs $K{=}75$
Q (residential housing price)	0.9996	0.9997	0.9999
HR (residential employment)	0.9892	0.9939	0.9944
HM (workplace employment)	0.9912	0.9918	0.9954
w (wage)	0.9999	0.9999	1.0000

The $K{=}25$ and $K{=}75$ runs sandwich the $K{=}50$ paper headline smoothly. Shock-response means are all near zero (consistent with V5's translation-invariance), and dispersion decreases monotonically with K (Table 13). The σ-with-K trend is the predicted direction that more clusters generate more LLM "voices" contributing to the per-cluster decision, and the resulting ensemble averaging tightens the cross-zone dispersion. Most importantly, the three baseline equilibria are essentially identical across K at the per-zone level (Table 14, every $r \geq 0.989$), so the choice of K does not change *where* V5 places agents in the spatial equilibrium but only the granularity of the cluster-prompt grouping.

I LIMITATIONS AND BROADER IMPACT

The historical validation rests on one event and 55 zones, and all correlations are modest, so the ranking is evidence of relative rather than absolute predictive power. V2–V4 estimates are reported at one seed without seed-replicate error bars. V4 and V5 results are pinned to specific LLM provider snapshots, so behavioural reproduction depends on provider-snapshot stability while bit-identical replay is supported by the bundled caches, and the marginal response of V5 to a small localised shock depends on prompt wording and on the archetype count (Appendix N). The agent-type distribution is held fixed across years in the historical validation, and the structural family and V5 run under different parameterisations on the 1986 instance, as documented in Appendix K. The Singapore

instance demonstrates that the pipeline runs on a second city with a different market mechanism but cannot play the validating role, since its parameters are not externally estimated, and the block-level run supports a scale claim only (Appendix O). Scaling the LLM-driven engine to substantially larger instances is gated by LLM compute budgets and by prompt length.

LUSim is a scenario-based counterfactual-prediction tool, not an operational urban-policy forecasting system, and should not be deployed for live decisions without held-out historical validation on the target city and shock class. Foreseeable risks include policy misuse (e.g., conflating rent-gradient flattening with displacement-protective gentrification), demographic and place-prior biases inherited from the foundation model, and the environmental footprint of LLM inference. The package licence (MIT for code, CC BY 4.0 for data and the V5 cache) is paired with a non-binding author-recommended use-policy norm in the README against commercial deployment as an investment-decision tool without re-validation on the target city.

J SUBMISSION ARTEFACT AND CACHE-REPLAY REPRODUCTION

Anonymised artifact access. For the double-blind review period the LUSim codebase, documentation, the bundled Berlin Ortsteile instance, and the V5 LLM cache are mirrored at `https://anonymous.4open.science/r/agent-urban-planning-iclr2027-45BC/`. The mirror reflects the submission-time commit, ships under the MIT licence (code) and CC BY 4.0 (data and cache), and contains the README instructions reviewers need to install the package and run a Tier 4 cache-replay reproduction without LLM credits. Public-repository, ReadTheDocs, and PyPI URLs (the package is published as `agent-urban-planning`) will replace the anonymised URL in the camera-ready version.

Tier 4 cache-replay reproduction. The headline V5 Berlin numbers in Table 2 reproduce from a clean download with the cache-replay invocation in Listing 4, requiring no LLM credits, no API keys, and approximately five to ten minutes of wall-clock on commodity hardware. Anonymous.4open.science serves only ZIP archives (no git protocol), so the mirror is fetched via `curl`; the camera-ready version will use the public GitHub URL with `git clone`:

Listing 4: Tier 4 cache-replay reproduction of the V5 Berlin Ortsteile result.

```
# 1. Download the anonymised mirror, create a Python venv (3.9+ required)
    , and install:
curl -L -o lusim.zip 'https://anonymous.4open.science/api/repo/agent-
    urban-planning-iclr2027-45BC/zip'
unzip lusim.zip -d agent-urban-planning
cd agent-urban-planning
python3 -m venv .venv
source .venv/bin/activate
pip install --upgrade pip
pip install -e ".[llm,plot,berlin]"

# 2. Download and extract the V5 LLM cache (~15 MB compressed):
curl -L -o llm_cache_v5.tar.gz https://huggingface.co/datasets/aup-anon
    -2026/LUSim-V5-LLM-cache-Berlin-ICLR2027/resolve/main/llm_cache_v5.
    tar.gz
tar -xzf llm_cache_v5.tar.gz -C data/berlin/

# 3. Replay V5 against the cache (no live LLM calls):
python examples/02_berlin_replication/run_v5_score_all.py --no-llm
# Outputs per-zone CSVs to output/v5_score_all/ (baseline) and
# output/v5_score_all_shock_east_west/ (shock).

# 4. (Optional) V4 cache replay for bit-identical Tier 3c reproduction:
python examples/02_berlin_replication/run_v4_hybrid.py --no-llm
# The V4 preference cache (~24 KB, 400 entries) ships in the ZIP at
# data/berlin/llm_cache_v4/, so no separate download is required.
```

The `--no-llm` mode runs at the paper configuration (`cluster_k=50`, `num_agents=1_000_000`, `iters=50`, `seed=42`) and replays cached responses for

every prompt. Here `num_agents` is the count of individual Monte-Carlo agent draws sampled from the $N = 10{,}000$ type distribution under the population weights w_i (Section 4); the LLM is only queried at the cluster level via the $K = 50$ archetypes, so `num_agents` controls market aggregation granularity rather than LLM call volume. The cache is partitioned by phase into `data/berlin/llm_cache_v5/baseline/` and `data/berlin/llm_cache_v5/shock/`. The `--no-llm` flag hard-fails on the first cache miss with a pointer to this appendix, so reviewers cannot accidentally consume LLM credits during a routine reproduction run. The structural variants V1, V2, V3 reproduce without the cache via `run_v1_softmax.py`, `run_v2_argmax_frechet.py`, and `run_v3_argmax_normal.py` in the same directory.

K HISTORICAL VALIDATION: INSTANCE, SCORING RULE, AND FULL RESULTS

The divided city of 1986. The Ahlfeldt et al. (2015) replication data record block-level fundamentals for 1936, 1986, and 2006. The 1986 instance fuses the East and West halves of the city into the same 96 Ortsteile as the 2006 instance and uses the walled travel-time matrix, in which 6,150 of the 9,216 origin-destination pairs cross the Wall and carry a prohibitive travel time, so that no commute crosses it. The replication data observe 1986 fundamentals for West Berlin only, so the 41 East Ortsteile carry zero housing prices, residents, workplaces, and wages in 1986, and observed 1986–2006 changes exist for the 55 West Ortsteile alone, which are therefore the scored zones. The counterfactual run restores the travel-time matrix of the calibrated 2006 instance, so that each engine's reunified equilibrium lives in the network geometry of its known 2006 behaviour, and holds every other fundamental at its 1986 value. All engines receive the 1986 commercial-floor input and the 1986 worker mass of 1,778,958. The agent-type distribution is held fixed across years, since the public sources from which it is constructed are 2011 marginals, so the test scores the response to the network change rather than demographic drift.

Engine configurations. The structural family uses the parameters calibrated on the three-year panel of the replication data, with the agglomeration terms held at their calibrated values, because the 2006 parameters do not transfer to the walled city. On the 1986 instance the 2006 parameters populate only 25 of the 55 West zones, with a baseline fit of $r = 0.33$ for log residents, whereas the panel parameters populate all 55 zones with $r = 0.93$ for log residents and $r = 0.90$ for log wages and reproduce the calibration's own 1986 run exactly. V5 has no structural parameter to calibrate and runs in its published configuration with the same 1986 inputs, $K = 25$ archetypes, and a cap of twenty market iterations. The archetype count matches the sweep configuration of Appendix H.3, and the cap follows from the sweep trajectories, in which the equilibrium state reaches its stochastic plateau by iteration eleven to fifteen (ΔQ falls from 0.035 to about 0.002 by iteration eleven and then fluctuates on the plateau through iteration fifty without approaching the 10^{-3} tolerance). Both runs of every pair share the same cap and all scored quantities are within-pair differences, so the cap cannot differentially affect any scored cell. The run with names withheld uses the anonymised-identifier machinery of Appendix H.2. The configuration asymmetry, the structural family at panel parameters and V5 as published, is inherent to a test that scores each engine against observed data rather than engines against each other under a shared mechanism, and it is disclosed wherever the results appear.

Scoring rule. The rule was fixed before any engine was scored. The prediction for each engine is the per-zone change in $\log Q$, $\log HR$, $\log HM$, and $\log w$ of the reunified run against that engine's own 1986 baseline run. The target, identical for all engines, is the per-zone observed log change from 1986 to 2006. Table 15 reports the raw means of the target, which embed twenty years of population change, deflators, and unit conventions that no network counterfactual should be asked to reproduce, so both predicted and observed changes are centred across the 55 zones before scoring. Per variable and per engine, four pattern metrics are computed on the centred changes: the Pearson correlation r, the Spearman correlation ρ, the sign agreement, namely the share of zones for which the centred prediction and the centred observation share a sign, and the relative RMSE, namely the RMSE of the centred prediction divided by the standard deviation of the centred observation. Table 16 lists every cell for the six engines and Table 17 the resulting ranking. The aggregate rank of an engine is its average rank across the sixteen variable-by-metric cells, computed over the cells

the engine has, with rank one the best per cell, higher r, ρ, and sign agreement better, lower relative RMSE better, and ties broken by mean Pearson r across variables. The level fit of the 1986 baseline, namely the Pearson correlation between the baseline run and the observed 1986 West-zone values, is reported alongside and never enters the ranking.

Table 15: Observed 1986–2006 target changes on the 55 West Berlin zones, raw log changes before centring.

Variable	Mean $\Delta \log$	Std. dev.
Q (housing price)	0.091	0.071
HR (residence employment)	−0.522	0.113
HM (workplace employment)	−0.536	0.270
w (wage)	0.235	0.098

Table 16: Full historical-validation cells for the six engines. n is the number of scoreable zones, r and ρ the Pearson and Spearman correlations between centred predicted and observed changes, sign the share of zones with agreeing sign, rel. RMSE the relative RMSE, and fit r the Pearson correlation of the engine's 1986 baseline run with the observed 1986 levels. The Monte Carlo argmax engines produce sparse residence masses and off-scale response variance on the walled instance, so V2's HM cell and V3's HR and HM cells cannot be computed and V2's HR cell rests on 14 zones; the aggregate rank averages the cells each engine has. V4 is the hybrid engine of the main text.

Engine	Var.	n	r	ρ	sign	rel. RMSE	fit r
V1 softmax	Q	55	−0.32	−0.32	0.34	1.18	0.21
	HR	55	−0.09	−0.04	0.46	1.22	0.93
	HM	55	−0.03	0.03	0.51	1.08	0.93
	w	55	−0.14	−0.20	0.36	1.01	0.90
V2 ABM	Q	55	0.26	0.27	0.58	14.28	0.21
	HR	14	−0.36	−0.50	0.43	18.19	0.93
	w	55	−0.15	−0.16	0.47	2.11	0.90
V3 Normal	Q	55	0.06	0.07	0.60	19.08	0.11
	w	55	−0.05	−0.07	0.49	2.77	0.64
V4 Hybrid	Q	55	−0.24	−0.17	0.44	1.11	0.20
	HR	55	0.00	0.09	0.56	1.15	0.93
	HM	55	−0.01	0.04	0.46	1.08	0.94
	w	55	−0.10	−0.08	0.47	1.01	0.91
V5 LLM	Q	55	0.25	0.26	0.66	0.99	−0.35
	HR	55	−0.13	−0.21	0.38	1.25	0.80
	HM	55	0.05	0.05	0.46	1.07	0.77
	w	55	0.00	0.05	0.58	1.00	0.88
V5, names withheld	Q	55	−0.16	−0.10	0.49	1.10	−0.29
	HR	55	0.23	0.11	0.51	0.97	0.87
	HM	55	0.05	0.03	0.56	1.05	0.74
	w	55	−0.07	−0.06	0.51	1.01	0.90

Table 17: Aggregate rank over the sixteen variable-by-metric cells (average rank, lower is better) and the tie-break statistic.

Engine	Cells	Average rank	Mean r
V5 LLM	16	2.25	0.043
V5, names withheld	16	2.25	0.013
V4 Hybrid	16	3.25	−0.087
V3 Normal	8	3.50	0.007
V2 ABM	12	4.08	−0.083
V1 softmax	16	4.19	−0.143

L MECHANISM DECOMPOSITION

Table 18 reports the decomposition discussed in Section 5.1. The relocation share is the L_1 distance between the post-shock and baseline residence-mass vectors divided by twice the total mass, i.e., the share of residents whose zone changes. $p_{95}|\Delta \log HR|$ is the 95th percentile of the absolute per-zone log change in residence employment, and the next column is the share of zones whose residence employment changes by more than 5%. The station share is the share of the total absolute change in residence mass occurring in the four zones containing new stations. The final two columns are correlations across zones between the wage change and, respectively, the change in workplace employment and the improvement in access measured as commuting minutes saved. The structural family's wage response is explained almost entirely by access, whereas V5's wage response is uncorrelated with access and tracks workplace density, which is the labour-market-pooling reading given in Section 5.1.

Table 18: Full mechanism decomposition of the East-West Express response.

Engine	Relocation share	$p_{95}\lvert\Delta \log HR\rvert$	Zones $> 5\%$	Station share	$\mathrm{corr}(\Delta \log w, \Delta \log HM)$	$\mathrm{corr}(\Delta \log w, \mathrm{access})$
V1 softmax	0.40%	0.0154	0.00%	49.64%	0.27	−0.88
V2 ABM	0.41%	0.0163	0.00%	47.97%	0.26	−0.89
V3 Normal	0.59%	0.0241	1.04%	38.57%	0.68	−0.49
V4 Hybrid	0.39%	0.0141	0.00%	48.85%	0.36	−0.87
V5 LLM	3.05%	0.0938	22.92%	2.99%	0.23	0.00

M ADDITIONAL POLICY SHOCKS

North-south corridor. The second rail corridor is a hypothetical four-station rapid-transit line running from Pankow (northern terminus) through Prenzlauer Berg and Mitte to Neukölln (southern terminus), with five minutes of nominal in-vehicle time between adjacent stations, constructed by the same shortest-path relaxation as the East-West Express (Appendix B, Eq. (10)). It reduces travel time on 287 of the 9,216 origin-destination pairs by 7.92 minutes on average and by 30.6 minutes at most, against 164 pairs, 5.93 minutes, and 24.8 minutes for the East-West Express. The specification ships as a forty-line configuration file, and the construction code is unchanged.

Floor-space expansion. The floor-space supply policy raises total, residential, and commercial floor area by 20% in eight eastern-periphery Ortsteile, namely the three Marzahn Ortsteile, Hellersdorf, Biesdorf, Kaulsdorf, Mahlsdorf, and Falkenberg, the large-panel-housing districts where the V5 response to the East-West Express concentrated its peripheral gains. The commercial share of each zone is preserved, so the policy is a pure supply expansion acting on H in $\mathrm{apply_policy}$. Floor supply is not an input to the V5 prompts, so the shock reaches the LLM-driven engine only through the equilibrium price vector, which is the economically correct information channel for a supply shock. A no-policy run through the supply path reproduces the baseline to machine precision.

Table 19: Full shock moments for the north-south corridor across the 96 Ortsteile, measured against each engine's own baseline.

Engine	$\mu_{\Delta \log Q}$	$\sigma_{\Delta \log Q}$	$p_{95}\lvert\Delta \log Q\rvert$	$\Delta \bar{Q}$	$\mu_{\Delta \log w}$
V1 softmax	0.0043	0.0267	0.0167	0.0034	−0.0011
V2 ABM	0.0044	0.0270	0.0177	0.0034	−0.0011
V3 Normal	0.0038	0.0270	0.0037	0.0017	−0.0009
V4 Hybrid	0.0039	0.0245	0.0167	0.0031	−0.0010
V5 LLM	−0.0035	0.0154	0.0296	−0.0104	0.0009

Table 20: Full shock moments for the floor-space expansion across the 96 Ortsteile, measured against each engine's own baseline.

Engine	$\mu_{\Delta \log Q}$	$\sigma_{\Delta \log Q}$	$p_{95}\lvert\Delta \log Q\rvert$	$\Delta\bar{Q}$	$\mu_{\Delta \log w}$
V1 softmax	−0.0121	0.0164	0.0648	−0.0122	0.0030
V2 ABM	−0.0121	0.0162	0.0647	−0.0122	0.0030
V3 Normal	−0.0112	0.0069	0.0299	−0.0105	0.0028
V4 Hybrid	−0.0121	0.0161	0.0630	−0.0122	0.0030
V5 LLM	−0.0146	0.0361	0.1157	−0.0215	0.0037

N SENSITIVITY OF THE LLM-DRIVEN ENGINE TO MODEL FAMILY AND PROMPT WORDING

The two sensitivity runs of Section 5.3 hold the environment, the agent population, the market clearing, the archetype grouping, the random seed, and the response validators fixed, at $K = 25$ archetypes and a cap of twenty market iterations, so that one factor varies at a time. The first replaces GPT 5.3 with DeepSeek V4 Flash, changing the model family together with its decoding configuration. Twenty live prompts were checked before the run and the second model followed the same decision procedure with no formatting failures, so no prompt adaptation was needed and the comparison carries no wording confound. The second rewrites the stage-one scoring instruction with identical content and an identical output schema on the original model, holding the stage-two instruction fixed. Table 21 reports the full shock moments of the three configurations. The baseline equilibrium is stable across all three: baseline household placement under the reworded prompt correlates with the original at 0.999 for housing prices and 0.993 for residence employment across the 96 zones, and the model change gives correlations above 0.99 on the same quantities. The peripheral concentration of gains, measured as the share of top-decile $\Delta \log Q$ zones lying off the corridor, is 1.00, 1.00, and 0.80 across the three configurations. The relationship between $\Delta \log Q$ and baseline $\log Q$, which is -0.17 in the original configuration at $K = 25$, strengthens to -0.75 under the model change and reverses to 0.18 under the rewording. The headline gradient flattening of Section 5.1 is at $K = 50$, where the dispersion is tighter ($\sigma_{\Delta \log Q} = 0.0083$ against 0.0109 at $K = 25$, Appendix H.3), and the signal at $K = 25$ is weak in the original configuration to begin with; the sensitivity runs were made at $K = 25$ for cost and are reported as they stand. The mean wage change in the central zones is slightly negative in all three configurations at $K = 25$, so the positive central wage response of Section 5.1 is specific to the $K = 50$ configuration and is sensitive to the archetype count rather than to the model family. The finding and the historical validation answer different questions, since the validation concerns how well an engine reproduces a large historical reorganisation while this experiment concerns the fine structure of the response to a small hypothetical shock.

Table 21: Full shock moments for the three V5 configurations on the East-West Express at $K = 25$.

Configuration	$\mu_{\Delta \log Q}$	$\sigma_{\Delta \log Q}$	$p_{95}\lvert\Delta \log Q\rvert$	$\Delta\bar{Q}$	$\mu_{\Delta \log w}$
GPT 5.3, original prompt	−0.0011	0.0109	0.0217	−0.0018	0.0003
DeepSeek V4 Flash, original prompt	−0.0030	0.0192	0.0395	−0.0106	0.0003
GPT 5.3, reworded prompt	0.0001	0.0096	0.0205	0.0016	−0.0007

O GENERALITY: A SECOND CITY AND BLOCK-LEVEL RESOLUTION

Singapore. The Singapore instance partitions the city into 27 planning zones linked by 729 transport connections, with a total housing supply of 1,096,380 units, base rents from S$1,733 to S$3,208, and 482 facility records across five types. The household population comprises 1,000 representative types drawn from Census 2020 marginals, with a weighted mean monthly income of S$10,924, 20.3% of households with children, 13.3% with elderly members, and 28.6% owning a car. Three candidate public-investment packages for the Punggol new town share a common budget of S$500 in budget units: a transit-heavy package (two rapid-transit connections and one facility), a facilities-heavy package (six facilities and no transit upgrade), and a balanced package (one connection and

three facilities). The market mechanism differs from Berlin's: a two-segment public and private housing market cleared by elasticity-based tâtonnement (price elasticity 0.5, convergence at a maximum excess demand of 0.05), with a mortgage-servicing ratio ceiling of 30% and the public-housing income ceiling applied as per-agent budget constraints at every iteration. Two engines are run: the structural engine with a literature-calibrated linear utility (coefficients -2.0 on price relative to income, -0.015 on commute minutes, 0.5 on facilities, and 0.8 on amenity) and the hybrid engine, which scales those coefficients by LLM-elicited per-agent weights. Table 22 reports the measured outcomes. Both engines rank the transit-heavy package first on average utility and lowest on the Gini coefficient; they differ on the order of the remaining two packages and on the Rawlsian criterion, which the hybrid engine's heterogeneous weights shift toward the facilities-heavy package. The instance demonstrates that the simulator runs on a second city with a different market mechanism; it cannot play the validating role, since its parameters are not externally estimated, which is why Berlin is where the simulator is validated.

Table 22: Singapore instance: measured outcomes of the three policy packages under the structural engine (literature-calibrated linear utility) and the hybrid engine (LLM-elicited per-agent weights). Utilities are population-weighted; the long-commute share is the share of households commuting more than 45 minutes; the unaffordable share is the share whose housing cost exceeds the mortgage-servicing ceiling. The last two rows report market convergence (tolerance 0.05 on maximum excess demand).

	Structural engine			Hybrid engine		
Metric	Transit	Facilities	Balanced	Transit	Facilities	Balanced
Average utility	-0.0311	-0.0322	-0.0525	-0.0623	-0.0777	-0.0728
Gini coefficient	0.1345	0.1431	0.1454	0.1131	0.1367	0.1332
Minimum utility (Rawlsian)	-1.6770	-1.6764	-1.6534	-2.0685	-1.7243	-1.9702
Average commute (minutes)	21.34	21.88	21.91	19.83	19.63	19.98
Long-commute share	0.060	0.060	0.057	0.055	0.053	0.064
Housing-unaffordable share	0.161	0.120	0.121	0.119	0.158	0.085
Market iterations to convergence	8	7	11	6	7	9
Final maximum excess demand	0.0496	0.0414	0.0449	0.0475	0.0483	0.0497

Block-level Berlin. At the replication data's full resolution of 12,309 statistical blocks, the structural engine runs five seeded Monte Carlo replicates (seed 42) of the baseline and of the replication pack's counterfactual; all five converge in both runs and the agglomeration recomputation converges in every replicate. Table 23 compares the simulated moments with the five targets the replication pack reports for that counterfactual (its Table A18). The two wage moments fall within the pack's 15% tolerance band, at 8.1 and 8.3% from target, whereas the three housing-price moments are off by a factor of about thirty in this configuration, so the block-level run supports a claim about scale rather than about replication fidelity at block resolution. The LLM-driven engine is not run at block level, since its stage-one prompt would carry 12,309 candidate rows; the Ortsteile aggregation is therefore a prompt-length choice for the LLM-driven engine rather than a limit of the simulator.

Table 23: Block-level Berlin (12,309 blocks): replication-pack moments for the pack's counterfactual, targets against the mean and standard deviation over five seeded replicates of the structural engine, with the deviation from target and whether it lies within the pack's 15% tolerance band. The pack's variable names are, in order, `wttshock_ratio`, `wttadjust_ratio`, `glfloor_mean`, `glfloor_treat_mean`, and `glfloor_control_mean`.

Moment	Target	Simulated mean	Std. dev.	Deviation	Within 15%
Wage ratio under the travel-time shock	1.1806	1.2765	0.0098	8.1%	Yes
Wage ratio after adjustment	0.9218	0.8454	0.0243	−8.3%	Yes
Mean log housing-price change, all blocks	-0.2004	-6.7239	0.0000	×33.6	No
Mean log housing-price change, treated blocks	-0.2921	-6.7097	0.0000	×23.0	No
Mean log housing-price change, control blocks	-0.1087	-6.7381	0.0000	×62.0	No