South American Publishing

SAP Transport, Mobility & Society. 2026;5:307
doi: 10.62486/tms2026307

REVIEW

Generative artificial intelligence and large language models in transport planning: a narrative review

Inteligencia artificial generativa y modelos de lenguaje de gran tamaño en la planificación del transporte: una revisión narrativa

Tammy Natalia Arellano Morales[1]
Lusby Depablos-Machado[2]
Juan Ernesto Pérez-Pérez[3]

[1] Universidad Autónoma de Sinaloa, Vicerrectoría. Mazatlán, México.
[2] Investigador independiente, Venezuela
[3] Universidad Simón Bolívar. Cúcuta, Colombia

Submitted: 17-03-2026 | Revised: 12-05-2026 | Accepted: 22-07-2026 | Published: 27-07-2026

Corresponding Author: Tammy Natalia Arellano Morales

ABSTRACT

Generative artificial intelligence and large language models are moving from general-purpose tools into the technical core of transport planning. This narrative review synthesizes the evidence published on their applications, contributions, and limitations. Across the axes the literature converges on the capacity of these models to reproduce activity and travel patterns, code open-ended mobility text, and support simulation and routing at a fraction of conventional cost, while it diverges on their numerical reliability, demographic representativeness, and readiness for decisions with public consequences. The review concludes that generative models are best understood as components of hybrid, human-supervised planning workflows rather than as replacements for behavioral models or deliberative processes, and it proposes a research agenda centered on transport-specific benchmarks, bias auditing, and reproducible evaluation.

Keywords: generative artificial intelligence; large language models; transport planning; travel demand; urban mobility; synthetic data.

RESUMEN

La inteligencia artificial generativa y los modelos de lenguaje de gran tamaño están pasando de ser herramientas de propósito general a formar parte del núcleo técnico de la planificación del transporte. Esta revisión narrativa sintetiza la evidencia publicada sobre sus aplicaciones, contribuciones y limitaciones. Entre ejes, la literatura converge en la capacidad de estos modelos para reproducir patrones de actividad y viaje, codificar textos abiertos de movilidad y apoyar la simulación y el enrutamiento con un costo menor que los métodos convencionales, mientras diverge en su fiabilidad numérica, representatividad demográfica y preparación para decisiones con consecuencias públicas. La revisión concluye que estos modelos se entienden mejor como componentes de flujos de trabajo híbridos supervisados por personas que como sustitutos de los modelos conductuales o de los procesos deliberativos, y propone una agenda centrada en benchmarks específicos del transporte, auditoría de sesgos y evaluación reproducible.

Palabras clave: inteligencia artificial generativa; modelos de lenguaje de gran tamaño; planificación del transporte; demanda de viajes; movilidad urbana; datos sintéticos.

© 2026; The Authors. This Open access article is distributed under CC BY 4.0 licence.

INTRODUCTION

Transport planning depends on travel demand models that translate socioeconomic structure into forecasts of trips, destinations, modes, and schedules. Activity-based systems of the kind formalized two decades ago remain the analytical backbone of regional planning, and they require detailed household travel surveys that are expensive, slow, and increasingly difficult to complete at acceptable response rates.(1)

This data burden has pushed agencies toward mobile traces, administrative records, and machine learning, yet the core modeling task still demands synthetic populations and activity diaries that preserve demographic and behavioral realism. The consequence is a modeling pipeline in which the most consequential step, the synthetic representation of the population, is also the least transparent to the agencies that depend on it.

Generative artificial intelligence changed what such models can attempt. Adversarial networks introduced in the mid-2010s(2)and the transformer architecture that underlies contemporary language models(3)now support the synthesis of tabular, sequential, and textual data from the same underlying distributions as observed mobility.

Two related lineages matter here: deep generative models that learn to sample from a data distribution, and large language models that reason over text and code. The second lineage extends to tasks that transport planners previously coded by hand, including open-ended survey responses, policy documents, and public comments.(4)

The last of these extends a participation tradition that planning formalized decades ago and has defended ever since.(5)The volume of applications published since 2023 has nonetheless grown faster than the evidence about their reliability.

The synthesis reported here covers what the published record between 2013 and 2026 demonstrates about these tools in transport planning. The objective is to characterize the main application domains, to summarize the reported performance and the recurring limitations, and to delineate a research agenda for practice. The synthesis is organized into five thematic axes: synthetic demand and population generation; travel behavior analysis and unstructured data; communication, public participation, and policy support; operational transport systems and mobility services; and evaluation, reliability, and governance.

DEVELOPMENT

Synthetic demand and population generation

The available evidence converges on deep generative models as the dominant route to synthetic activity and travel data. Early work reconstructed individual travel demand from aggregated mobile phone records with adversarial training and produced digital twin travelers whose trip distributions matched observed mobility.(7)

Subsequent studies extended the same logic to recurrent networks and to conditional generative adversarial networks that impute missing qualitative trip attributes and generate realistic population mobility with privacy constraints.(8,9)

Variational autoencoders and diffusion models widened the generative family. A systematic investigation of variational autoencoders for population synthesis documented their suitability for reproducing household distributions, while diffusion-based trajectory generators captured co-movement and sequence structure that single-step models miss.(10,11)

Apart from these architectural advances, the field also released reusable synthetic populations with documented activity and travel patterns, which allow independent replication rather than another one-off demonstration.(12)Across studies, the reported gains concentrate on fidelity and speed rather than on behavioral interpretability.

A further pattern concerns the target of generation. Some works synthesize a full synthetic population with household attributes, while others generate only the travel layer, conditional on a population that already exists, and the two problems differ in both difficulty and use. Population synthesis must respect joint demographic constraints, whereas activity generation must respect the temporal logic of a day, and the reviewed studies do not compare the two on the same input.

The motivation for much of this work is as much about data access as about accuracy. Synthetic populations and diaries let agencies work with realistic disaggregate inputs without redistributing sensitive records, and several generators are framed as privacy-preserving alternatives to the release of survey microdata.(8)

This explains the recurring evaluation style of the axis: generated distributions are compared against census margins, trip length statistics, and activity sequence frequencies, while the downstream effect on a planning decision is left unmeasured. The comparison studies confirm that the predictive fit is real but silent about behavioral mechanisms.(13)

The modeling choices beneath these results are not neutral. Trip-based generators condition on household attributes and reproduce marginal trip rates, whereas activity-based generators must also reproduce the sequence and duration of activities, a substantially harder target that several of the language-model and agent frameworks set for themselves.(14,15)

The choice of conditioning variables therefore determines which fidelity statistics can be reported, and it explains part of the divergence between studies that claim success and studies that report modest gains.

Large language models entered this axis as semantic constraints on generation. Population synthesis built on language models reproduced feasible and diverse households, and retrieval-augmented approaches generated activity schedules with household coordination under limited information.(16,17)

Case evidence from a Korean city synthesized human mobility patterns with a language model, which illustrates the transfer of the method from research settings to a concrete municipal context.(18)

Generative agents that embed cognitive and social rules produced urban mobility simulations whose daily patterns resembled observed behavior.(19,20)Irrespective of the specific architecture, these studies agree on the value of language-derived priors for plausibility, and they diverge regarding the numerical precision of the generated counts.

The comparative record is more cautious. Benchmarks that pitted hundreds of machine-learning models against discrete-choice models found that the generative and deep models matched or exceeded predictive fit while offering no behavioral parameters, and a parallel comparison of mobility tool ownership reached a similar verdict.(21)

Activity schedule modeling with machine learning showed the same trade-off in a narrower setting, where predictive accuracy came at the expense of the behavioral parameters that planners use to test policies.(22)

Direct demand modeling with Wasserstein adversarial networks reproduced ride-hailing trips, and graph-guided generative models imputed multimodal demand, yet both depended on the training geography.(23,24)

The fidelity statistics themselves deserve scrutiny. Marginal distributions can match while joint distributions do not, and a generator that reproduces average trip rates may still produce households whose members travel together unrealistically or not at all. Household coordination is therefore the dimension on which the newest work concentrates, and it is also the dimension for which observed references are scarcest.(17,25)

The general transportation agent that produces synthetic travel survey data points to the next step, in which generation and survey design merge into a single tool.(26)Such a tool would let an agency pilot a questionnaire on a synthetic population before fielding it, but its value depends on whether the synthetic respondents behave like the real ones, which returns the axis to its central unsolved question.

The clearest gap in the axis is the absence of shared benchmarks. Each study reports its own fidelity statistics on its own city, and no common corpus lets an agency compare a variational autoencoder against a language-model agent on identical inputs.(10,16)

The newest evidence integrates the two families. Frameworks for household activity coordination, household relationship generation, and electric vehicle trip synthesis combine generative sampling with structural constraints.(15,25,27)

Large language models appear in this line as generators of believable behavior, as estimators of regional travel structure, and as cognitive engines for personalized travel, while a general transportation agent produced both mobility forecasts and synthetic travel survey data.(14,26,28,29)

A hierarchical cognitive framework extended the same principle to interpretable human mobility behavior.(30)The axis therefore documents a coherent shift from purely statistical generators toward hybrid systems that embed semantic knowledge, at the price of a wider gap between demonstration and validation. In total, 24 core studies support this axis.

Travel behavior analysis and unstructured data

The evidence on travel behavior converges on two complementary uses. The first is the interpretation of unstructured text, in which language models code open-ended survey responses,

citizen complaints, social media messages, and charging infrastructure reviews at scale.(31-33)

The second is the reproduction of behavior itself, in which models infer trip purpose, estimate bicycle volumes, and predict daily travel patterns from traces or auxiliary text.(34-36)

The coding applications matter because open-ended survey and complaint text is abundant but underused in practice. A language model can assign trip purposes, classify dissatisfaction with a bus network, or extract preferences about charging infrastructure from thousands of comments without a manual codebook, which is precisely the labor that small planning agencies cannot afford.(31,32,34) The same capacity supports the automated reading of social media as a proxy for urban traffic conditions, with the caveat that the signal reflects who posts rather than who travels.(33)

The results agree on competence in bounded tasks. Language models reproduced stated and revealed preferences in experiments on the value of travel time, on route choice, and on the decoy effect, and persona-based embedding aligned model outputs with observed travel choices.(37-40)

Studies that fused structural knowledge with language-inferred context reported better route choice models than either source alone.(41)Across studies, the models performed well when the task was textually specified and poorly when it required local knowledge or precise arithmetic.

Divergence concentrates on generalizability. A real-world experiment on mobility data and transport planning documented both the reasoning capacity and the knowledge-conditioning failures of current models, and an agenda paper for agentic language models in travel behavior research makes the same point from the design side.(42,43)

Event-aware analyses of visitor flows and street-level safety inference extended the method to multimodal inputs, with accuracy that depended on the alignment between the corpus and the target city.(44,45)

These patterns repeat across the coding applications, in which a model trained or prompted for one city does not transfer cleanly to another.

A related distinction separates the model as coder from the model as respondent. As a coder, the language model reads text that real people produced, which is a task with a verifiable reference; as a respondent, it answers survey-like questions in place of travelers, which is synthetic preference data whose reference is unclear.(37,40)

The two uses are conflated across the reviewed work, and their validity conditions are not the same.

The evaluation gap is visible here as well. Behavioral reproduction is judged against aggregate shares or correlation with survey answers, while the decision process that produces the behavior remains unobserved, so a model can match the outcome and miss the mechanism. This matters for planning because behavioral models are used to simulate counterfactuals, and a model that mimics without mechanism may extrapolate incorrectly when a policy changes the context in which choices are made.(38,43)

The value-of-time experiments illustrate the promise and the limit at once. A language model can return plausible stated-preference answers because the underlying trade-off is textual and familiar, yet the same model has no experience of the congestion that the parameter is meant to price.(37)Reported agreement with survey data therefore measures internal consistency with the training corpus, not external validity for a specific population.

A further validity question concerns the reference standard. When a language model codes open-ended text, the reference is a human codebook, and the model is then judged against the same inter-coder agreement that human coders achieve.(31,34)

When the reference is instead a stated-preference survey, the comparison confounds model error with survey error, and neither source of error is quantified in the reviewed studies.

The axis is nonetheless the most active of the five in behavioral terms. Prediction studies report that language models approach dedicated sequence models for travel behavior forecasting, and commuting pattern representations derived from language models reproduced intercity differences with open data.(46,47)

Agentic models for day-to-day route choices add a further layer, in which the model reasons over the same information a traveler would use instead of learning a single mapping.(38)

The boundary between statistical mimicry and behavioral realism remains the central unresolved question of the axis, because most studies validate against aggregate outcomes rather than against the decision process itself. In total, 16 core studies support this axis.

Communication, public participation, and policy support

This axis is smaller and newer than the previous two, and its studies converge on the use of language models as readers and drafters of planning text. Comparative analyses of transporta-

tion policy frameworks for equity and accessibility, and knowledge graph frameworks for electric vertical takeoff and landing policy, treated public documents and citizen concerns as analyzable corpora.(48,49)

Generative artificial intelligence was proposed as a pathway for complex urban planning and for net-zero urban planning, with case studies that emphasize potential over demonstrated impact.(50,51)

Participatory applications move from reading to engaging. Language models were used to assess and map contributions in participatory urban planning, and a policy alignment framework addressed the mismatch between model proposals and stated public objectives.(52,53)

The proposed workflows position the model between the public comment and the planner, which is efficient when the task is summarization and hazardous when the task is adjudication.

The equity risk is concrete. Automated coding and drafting can reproduce the priorities of the texts on which they were trained, and if those texts underrepresent low-income or non-driving residents, the summarized consultation will underrepresent them as well.(48)

None of the reviewed studies reports a test of whether its summarization preserved or amplified existing disparities.

Two design choices recur across the six works. The first is the coupling of a language model with a knowledge graph or a corpus of policy documents, which grounds the output in source text rather than in parametric memory.(49,53)

The second is the explicit positioning of the model as a support to human judgment rather than a substitute for it, although the reviewed studies do not report how that human step actually functioned.

These studies agree that the tools scale the processing of public text, and they diverge regarding whether the resulting engagement is more or less representative than conventional channels. A model can read every comment in a consultation, but it cannot restore the voices that never commented, and a generator that drafts policy language can embed assumptions that no participant endorsed. The divergence is unresolved because none of the six works measured participation against an outcome that residents experienced. Six core studies support this axis.

Operational transport systems and mobility services

The operational literature is the broadest of the five axes, and it splits between systems that replace a numerical module and systems that coordinate an existing one. Reviews of language models for intelligent transportation systems and for traffic forecasting map this territory and report that most applications remain proofs of concept.(54,55)

In parallel, a system-oriented review organized the contributions by subsystem and identified simulation and control as the fastest-growing uses.(56)

Hybrid architectures are the consistent pattern. Language models enhanced with reinforcement learning produced generic bus holding control strategies, and language-model agents steered conversational traffic simulation in SUMO.(57,58)

Multimodal routing, interlining bus scheduling, and preference-aware planning for mobility as a service combined language reasoning with optimization or neuro-symbolic components.(59-61)

Irrespective of the domain, these hybrids outperformed the language model used in isolation on numerical tasks, which suggests that the language component contributes control logic rather than arithmetic.

Simulation and prediction occupy the remainder of the axis. Generative diffusion models reconstructed traffic states, spatial-temporal language models with diffusion predicted multimodal flows, and open-source language models served as a foundation for individual mobility prediction.(62-64)

Digital twin architectures and path generation frameworks applied the same principle to urban mobility simulation and to dynamic route construction.(65,66)

Studies that predicted mobility under public events and that selected electric vehicle charging locations illustrate the operational maturity the field has reached.(67,68)

The two reviews also expose a classification problem that the field has not settled. Whether a bus holding controller, a route planner, and a traffic simulator belong to the same class of application is not obvious, and the resulting heterogeneity of metrics makes the reported gains difficult to compare across subsystems.(55,56)

A shared task taxonomy would do more for the field than another isolated demonstration.

What remains manual in these systems is as instructive as what is automated. The reviewed frameworks still require a human to define the scenario, supply the network, and judge the output,

so the language model accelerates rather than eliminates the work of the analyst.(58,60)

This division of labor is consistent with the hybrid pattern and with the governance caution of the fifth axis.

The prediction applications carry a different risk profile from the control applications. A forecasting error degrades an estimate, whereas a control error can propagate into schedules and headways that riders experience in real time, which is why the bus holding and traffic simulation studies emphasize safety and feasibility constraints that pure forecasting ignores.(57,58)

This distinction is a useful organizing principle for agencies deciding where to begin.

Language also changes the interface to these systems. Conversational agents allow an operator to describe a scenario in natural language and receive a simulated response, which lowers the barrier between a planner and a traffic model.(58,65)

The same property can conceal the assumptions inside the model, because a fluent answer reads as a confident answer regardless of its factual basis.

The maturity is nonetheless uneven. The control and routing applications are demonstrated on one network with one model in the reviewed studies, and the reviews note that standardized benchmarks are absent, so reported gains are hard to compare across systems.(55,56)

The dependence on the knowledge base that conditions the language model is the second recurrent weakness: when the base is incomplete, the hybrid inherits its gaps and can propose infeasible schedules or routes. These are engineering limits rather than conceptual ones, and they mark the axis as the most ready for incremental adoption. In total, 15 core studies support this axis.

Evaluation, reliability, and governance

The fifth axis is the least developed, and its studies converge on a shared diagnosis rather than on a shared solution. A socio-technical examination documented bias in generative artificial intelligence travel planners, and trust research identified privacy and data security as the moderators of acceptance in personalized travel planning.(69,70)

The real-world experiment cited in the second axis is again informative here, because it shows that output quality varies with the conditioning data provided to the model.(42)

Resilience work adds a planning-specific concern. A review of generative artificial intelligence and urban transport resilience argues that the same flexibility that helps planners explore scenarios can obscure the assumptions on which a plan rests.(71)

A generated forecast that looks like a modeled forecast can therefore travel through a planning process without the audit trail that a conventional model carries.

Bias extends beyond demographic categories. A socio-technical reading of travel planners shows that the model can also encode assumptions about what a desirable trip or destination is, and these assumptions then surface as recommendations that look neutral.(69)

Because planners do not see the alternatives the model discarded, this form of bias is harder to notice than a skewed summary statistic.

The trust findings point in the same direction. Acceptance of conversational travel planners depends less on the accuracy of the model than on how privacy and data security are handled, which implies that governance is a design parameter rather than an afterthought.(70)

A planner that cannot explain which data it used will face resistance even when its output is correct.

Auditability is the common thread. Bias, privacy, and trust all depend on whether an external party can inspect what the model was trained on and how it was prompted, and none of the four works provides such an inspection protocol.(69,70)

The axis therefore identifies the problem that the other four axes would need to solve before their outputs can carry weight in a contested decision.

What the axis lacks is as telling as what it contains. No study in this set defines a transport-specific benchmark, a standardized bias audit, or a protocol for verifying a generated quantity before it enters a statutory decision. The absence is not an accident: the earlier axes demonstrate capability, while this axis would have to demonstrate trustworthiness, and that question is harder to answer with a single architecture or corpus. Four core studies support this axis.

The five axes describe a coherent but uneven field. The synthesis table condenses the applications, the convergent findings, and the divergent findings of each axis (Table 1).

Table 1. Thematic synthesis of generative artificial intelligence and large language models in transport planning.

Axis	Core applications	Convergent findings	Divergent findings	Studies (n)
Synthetic demand and population generation	Synthetic populations, activity and travel diaries, trip and demand synthesis	Distributional fidelity, speed, privacy by design	Numerical precision, interpretability, dependence on training geography	24
Travel behavior analysis and unstructured data	Survey coding, trip purpose inference, value of time, route choice, sentiment	Accuracy comparable to human coders on bounded tasks	Local knowledge, arithmetic, heterogeneous preferences	16
Communication, public participation, and policy support	Policy analysis, equity assessment, public comments, plan drafting	Scalable text processing, wider engagement	Representativeness, automation of deliberative decisions	6
Operational transport systems and mobility services	Simulation, forecasting, routing, bus and fleet control	Hybrid models outperform isolated language models	Dependence on knowledge base, proof-of-concept maturity	15
Evaluation, reliability, and governance	Bias auditing, privacy and trust, benchmarks	Shared recognition of risk	Few transport-specific benchmarks and protocols	4

DISCUSSION

The immediate implication of this synthesis is that transport planning should treat generative models as components of hybrid workflows rather than as autonomous replacements for behavioral models or deliberative processes. The evidence is strong where the task is synthetic data generation and weak where the task carries public consequence, and the two observations are connected: models that reproduce distributions well are precisely the ones whose outputs planners might trust without inspection. This tension runs across all five axes.

For a journal concerned with social equity and sustainable mobility, the equity reading of this finding is direct. The synthetic populations and coded consultations that generative models produce will shape whose trips are counted and whose preferences are summarized, and an unexamined model can quietly reproduce the biases of its training data.(48,69)

The sustainability applications follow the same logic, because a net-zero scenario generated from biased demand inherits the bias before any mitigation is debated.(51)

The axes relate to one another through a shared dependence on data and evaluation. Synthetic demand generation and behavior analysis both inherit the limitations of the survey or trace data they are trained on, which explains why studies that train on one city fail when transferred to another.(23,44)

Irrespective of the application, the recurring failure is not architectural but evidential: the field reports fidelity against internal benchmarks far more often than it reports performance on planning decisions with measurable outcomes.(42,55)

In parallel, the participation and governance axes show that the barriers to adoption are organizational and normative as much as technical.(53,71)

Comparison with prior reviews sharpens the contribution of this work. Earlier surveys treated intelligent transportation systems or mobility forecasting as the unit of analysis and organized the literature by subsystem.(55,56)

Asurvey of language models in transportation planning organized contributions by modeling, design, and decision-making,(72)and a methodological review placed the transport uses within the wider picture of language model capabilities.(73)

The present review differs by separating the deep generative lineage from the language model lineage and by reporting the evidence at the level of planning tasks, which is where the divergent findings become visible. This organization also makes explicit that the two lineages now merge in hybrid frameworks.(26,30)

Two strengths distinguish the present work from that prior literature. It covers both the deep generative and the language model literatures within one analytic frame, which makes the convergence toward hybrid systems visible, and it reports evidence at the task level, where the practical decision about adoption is actually made. The five-axis organization is therefore a deliberate compromise between the subsystem taxonomy of the operational surveys and the capability

taxonomy of the methodological reviews.

Separating the two lineages also clarifies a common confusion. Deep generative models produce numbers and are evaluated against numbers, whereas language models produce text and are evaluated against both text and numbers, and the second evaluation is the less standardized of the two.(72,73)

Claims that a language model replaces a travel demand model therefore need a level of numerical evidence that the current literature does not supply.

The transferability findings also carry a practical warning. A model demonstrated in one city embeds that city's data and assumptions, and an agency that adopts it elsewhere inherits both.(23,47)

Adoption guidance should therefore treat every published result as a demonstration on a specific population rather than as a general certificate for the method.

The divergences have concrete explanations. Studies that report accurate behavioral reproduction used bounded, textually specified tasks with a single model and a single corpus,(37,39) whereas studies that report failures introduced local knowledge, arithmetic, or heterogeneous populations.(42,47)

Population, measurement, and context therefore explain most disagreements: a model that reproduces the value of travel time in a stated-preference experiment says little about its accuracy for a city it has never seen. The bias and trust literature supports this reading, because it locates the risk not in the model alone but in the fit between the model, the data, and the decision.(69,70)

The same caution appears in qualitative digital research, which warns that automated coding inherits the biases of its training text.(74)

The practical implications follow directly. Planning agencies can adopt generative models today for tasks with a closed evaluation loop, such as synthetic population generation checked against census margins, and for the coding of open-ended text under human review. They should be slower to adopt them for unilateral decisions on equity, network investment, or public communication, where the cost of an unnoticed error falls on residents rather than on the model developer.

The participation tradition that planning has maintained since the earliest formalized ladders of citizen involvement argues for the same caution: a tool that reads and drafts public text must not displace the deliberation itself.(5)

The SANRA framework used here also implies a reporting obligation, namely that methods sections should state the model, the conditioning data, and the verification step, which most of the reviewed studies do only partially.(6)

The deeper question is where professional judgment belongs. Transport planning has always treated its models as advisory, and the danger of a fluent generative system is that fluency invites deference rather than judgment.

The remedy is procedural: generated inputs should enter the same review chain as any other modeled input, with the same owners and the same sign-offs, so that the novelty of the tool does not bypass the accountability that the rest of the pipeline carries.(42,53)

A minimal evaluation protocol would resolve part of the tension. It would require, for any generated quantity, a statement of the conditioning data, a comparison against an observed reference, and a named human owner for the decision in which the quantity is used.(6,42)

Such a protocol is not a technical breakthrough; it is the transport analogue of the reporting standards that surveys and models already follow, extended to a class of tools that currently escapes them.

The research agenda follows from these gaps. Transport-specific benchmarks that measure generated quantities against observed counts, standardized bias audits across demographic and geographic strata, and reproducibility reports with shared synthetic populations would convert the current proof-of-concept literature into evidence that planners can act on.(42,69)

A second line should test generative models inside real participatory processes and policy workflows, where the participation axis is thinnest.(52,53)

A third line should formalize governance, including when a generated forecast may enter a statutory plan and when it may not.(71)

Progress on these three lines would not require new model architectures; it would require the field to treat evaluation as a first-class deliverable rather than an appendix.

A practical roadmap follows from this balance. In the near term, agencies can deploy the coding and simulation applications with human verification, adopt population synthesis against census margins, and hold the participation and decision applications in pilot. In the longer term, the research agenda above should supply the benchmarks and audit protocols that would let the more consequential uses graduate from demonstration to practice.

Accountability also requires a named owner for every generated input, because the model cannot be held responsible for a decision that a planner signs. The reviewed studies concentrate on technical performance and leave this institutional question open, which is another reason the governance axis remains the least developed.(71)

CONCLUSIONS

Generative artificial intelligence and large language models already contribute to transport planning across five axes, with the strongest evidence in synthetic demand generation, behavior analysis, and operational simulation, and the weakest in participation and governance. The literature converges on hybrid, human-supervised use and diverges on numerical reliability, representativeness, and readiness for public decisions. The field needs transport-specific benchmarks, bias auditing, and reproducible evaluation before these tools should carry consequences for residents, and until then they are best positioned as accelerators of the tasks planners already perform rather than as substitutes for behavioral models or deliberation.

REFERENCES

1. Bowman JL, Ben-Akiva ME. Activity-based disaggregate travel demand model system with activity schedules. Transportation Research Part A: Policy and Practice. 2001;35(1):1-28. doi: 10.1016/s0965-8564(99)00043-9
2. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. Commun ACM. 2020;63(11):139-144. doi: 10.1145/3422622
3. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need [Internet]. 2017. Available from: https://arxiv.org/abs/1706.03762
4. Fajardo Quesada AJ, Hernández González EA, Herrero Pacheco R, Álvarez Pérez RJ. Chapter 12. Generative AI and Large Language Models. In: Primum non nocere: Software engineering for medicine. South American Publishing; 2026. doi: 10.62486/978-9915-9876-8-2.ch12
5. Arnstein SR. A Ladder Of Citizen Participation. Journal of the American Institute of Planners. 1969;35(4):216-224. doi: 10.1080/01944366908977225
6. Baethge C, Goldbeck-Wood S, Mertens S. SANRA—a scale for the quality assessment of narrative review articles. Res Integr Peer Rev. 2019;4(1):5. doi: 10.1186/s41073-019-0064-8
7. Anda C, Ordonez Medina SA, Axhausen KW. Synthesising digital twin travellers: Individual travel demand from aggregated mobile phone data. Transportation Research Part C: Emerging Technologies. 2021;128:103118. doi: 10.1016/j.trc.2021.103118
8. Berke A, Doorley R, Larson K, Moro E. Generating synthetic mobility data for a realistic population with RNNs to improve utility and privacy. Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing. 2022:964-967. doi: 10.1145/3477314.3507230
9. Kim EJ, Kim DK, Sohn K. Imputing qualitative attributes for trip chains extracted from smart card data using a conditional generative adversarial network. Transportation Research Part C: Emerging Technologies. 2022;137:103616. doi: 10.1016/j.trc.2022.103616
10. Sané AR, Vandanjon PO, Belaroussi R, Hankach P. A comprehensive investigation of variational auto-encoders for population synthesis. J Comput Soc Sc. 2025;8(1):13. doi: 10.1007/s42001-024-00332-0
11. Abbar H, Kassan S, Bidet F, Cherigui A, Guépin F, Renaud D. TrajDD-GAN: A Synthetic Mobility Trajectory Generation Solution Based on Diffusion Models. IEEE Access. 2025;13:158018-158027. doi: 10.1109/access.2025.3602274
12. Tozluoğlu Ç, Dhamal S, Yeh S, Sprei F, Liao Y, Marathe M, et al. A synthetic population of Sweden: datasets of agents, households, and activity-travel patterns. Data in Brief. 2023;48:109209. doi: 10.1016/j.dib.2023.109209
13. Wang S, Mo B, Zheng Y, Hess S, Zhao J. Comparing hundreds of machine learning and discrete choice models for travel demand modeling: An empirical benchmark. Transportation Research Part B: Methodological. 2024;190:103061. doi: 10.1016/j.trb.2024.103061
14. Huang F, Zhang Z, Tang J. Believable urban planning via large language model-enabled behavior modeling: A conceptual framework. Cities. 2026;176:107243. doi: 10.1016/j.cities.2026.107243
15. Liao X, Ma H, Liu Y, Wei Y, He BY, Stanford C, et al. Next-Generation Travel Demand Modeling with a Generative Framework for Household Activity Coordination. 2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC). 2025:561-568. doi: 10.1109/itsc60802.2025.11423529
16. Lim SY, Yun H, Bansal P, Kim DK, Kim EJ. A large language model for feasible and diverse population synthesis. Transportation Research Part C: Emerging Technologies. 2026;185:105581. doi: 10.1016/j.trc.2026.105581
17. Liu Y, Liao X, Ma H, He BY, Stanford C, Ma J. Human Mobility Modeling with Household Coordination Activities Under Limited Information via Retrieval-Augmented LLMs. 2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC). 2025:951-958. doi: 10.1109/itsc60802.2025.11423539
18. ROH C, WON M. Synthesizing Human Mobility Patterns Using Large Language Models - A Sejong City Case Study. jkst. 2026;44(2):270-288. doi: 10.7470/jkst.2026.44.2.270
19. Liu Q, Li C, Ma W. GATSim: Urban mobility simulation with generative agents. Transportation Research Part C: Emerging Technologies. 2026;186:105576. doi: 10.1016/j.trc.2026.105576
20. Liu Q, Li C, Ma W. Generative agents for urban mobility: A cognitive framework for realistic travel behavior simulation. Simulation Modelling Practice and Theory. 2026;147:103234. doi: 10.1016/j.simpat.2025.103234
21. Püschel J, Barthelmes L, Kagerbauer M, Vortisch P. Comparison of Discrete Choice and Machine Learning Models for Simultaneous Modeling of Mobility Tool Ownership in Agent-Based Travel Demand Models. Transportation Research Record: Journal of the Transportation Research Board. 2024;2678(7):376-390. doi: 10.1177/03611981231206175
22. Koushik A, Manoj M, Nezamuddin N, Prathosh A. Activity Schedule Modeling Using Machine Learning. Transportation Research Record: Journal of the Transportation Research Board. 2023;2677(8):1-23. doi: 10.1177/03611981231155426
23. Rahman MH, Abrar M, Rifaat SM. Linear regression coupled Wasserstein generative adversarial network for direct demand modeling of ride-hailing trips in Chicago and Austin. Transportation Letters. 2025;17(4):653-665. doi: 10.1080/19427867.2024.2372944
24. Li C, Liu W, Ma W, Yang H. Simultaneous Multimodal Demand Imputation and Forecasting via Graph-Guided Generative and Adversarial Network. Transportation Science. 2025;59(4):763-781. doi: 10.1287/trsc.2023.0326
25. La DM, Vu HL. TreeCSP: An Efficient Way to Generate Household Relationships for Population Synthesis. 2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC). 2025:2148-2153. doi: 10.1109/itsc60802.2025.11423173

26. Salvador I, Furno A, Derrible S. Large Language Model-Enhanced General Transportation Agent Framework for Human Mobility Forecasting and Synthetic Travel Survey Data Generation. Transportation Research Record: Journal of the Transportation Research Board. 2026:03611981261456273. doi: 10.1177/03611981261456273
27. Chen X, Hamim OF, Ukkusuri SV. Electric vehicle trips detection and synthesis using Sequential Generative Adversarial Networks. Expert Systems with Applications. 2025;280:127410. doi: 10.1016/j.eswa.2025.127410
28. Shen H, Weng J, Lin P. An LLM-driven estimation framework for estimating regional travel structures based on spatiotemporal multi-modal travel data. Sustainable Cities and Society. 2026;138:107196. doi: 10.1016/j.scs.2026.107196
29. Ge S, Ye P, Zhang R, Zhou M, Dong H, Wang FY. LLM-Driven Cognitive Modeling for Personalized Travel Generation. IEEE Trans Comput Soc Syst. 2025;12(5):3455-3464. doi: 10.1109/tcss.2025.3556397
30. Li Q, Ji C, Liu X. From narrative to action: A hierarchical cognitive large language model agent framework for interpretable and adaptive human mobility behavior generation. Engineering Applications of Artificial Intelligence. 2026;184:116203. doi: 10.1016/j.engappai.2026.116203
31. Wang X, Zhang D, Li X, Peng J. Diagnosing public EV charging deficits via LLM-KG analysis of citizen complaints in Wuhan. Sustainable Cities and Society. 2026;151:107970. doi: 10.1016/j.scs.2026.107970
32. Huang Y, Ren H, Xie D, Cui Q, Jin Y, Chen D. Decoding EV user preferences with large language model and charging infrastructure reviews. Transportation Research Part D: Transport and Environment. 2026;156:105348. doi: 10.1016/j.trd.2026.105348
33. Li M, Fan C, Yan H, Wu P, Su Z, Ma W. Urban traffic evaluation with social media data: A consensus-based LLM negotiation paradigm. Transportation Research Part A: Policy and Practice. 2026;208:104980. doi: 10.1016/j.tra.2026.104980
34. Zhang C, Li S, Ma X, Chen X, Huang Z, Bai Z, et al. Adapting large language models for urban trip purpose inference: A domain-informed textual representation and fine-tuning framework. Cities. 2026;179:107680. doi: 10.1016/j.cities.2026.107680
35. Azizi Soldouz S, Hellinga B. Estimating Bicycle Volume Levels at Urban Intersections using Large Language Models. Transportation Research Record: Journal of the Transportation Research Board. 2026;2680(9):701-718. doi: 10.1177/03611981251407921
36. Sharifi D, Mirzahossein H, Kalantari N, Sedighi F. Integrating hybrid recurrent neural networks and large language models for daily travel behavior prediction. Transportation Research Interdisciplinary Perspectives. 2026;35:101793. doi: 10.1016/j.trip.2025.101793
37. Yan Y, Liu T, Yin Y. Valuing time in silicon: Can large language models replicate human value of travel time. Travel Behaviour and Society. 2026;44:101245. doi: 10.1016/j.tbs.2026.101245
38. Wang L, Duan P, He Z, Lyu C, Chen X, Zheng N, et al. Agentic Large Language Models for day-to-day route choices. Transportation Research Part C: Emerging Technologies. 2025;180:105307. doi: 10.1016/j.trc.2025.105307
39. Hassan M, Narupiti S, Bhuiyan T, Hassan MM. Leveraging Large Language Models to Investigate the Decoy Effect in Route Choice Behavior. 2025 International Conference on Sustainable Technology and Engineering (i-COSTE). 2025:1-6. doi: 10.1109/i-coste68047.2025.11467175
40. Liu T, Li M, Yin Y. Aligning LLM with Humans for Travel Choices: A Persona-Based Embedding Learning Approach. Transportation Science. 2026;60(5):806-827. doi: 10.1287/trsc.2025.0330
41. Qiu S, Huang X, Qin G, Chen X, Sun J. Generalized Route Choice Modeling via Fusion of Structural Knowledge and LLM-Inferred Context. 2025 IEEE 28th International Conference on Intelligent Transportation Systems (ITSC). 2025:1560-1567. doi: 10.1109/itsc60802.2025.11423276
42. Gioia S. Large Language Models for Mobility Data and Transport Planning: A Real-World Experiment on Output Data, Reasoning, and Knowledge-Conditioning. 2026 IEEE Swiss Conference on Data Science and AI (SDS). 2026;(2026):52-59. doi: 10.1109/sds70563.2026.00015
43. Salvador I, Furno A, Derrible S. Toward agentic large language models in travel behavior research. Transportation Research Interdisciplinary Perspectives. 2026;40:102220. doi: 10.1016/j.trip.2026.102220
44. Wang X, Zhao Z, Wang R, Xu Y. Event-aware analysis of cross-city visitor flows using large language models and social media data. Transportation Research Part A: Policy and Practice. 2026;209:105023. doi: 10.1016/j.tra.2026.105023
45. Tang Y, Qu A, Yu X, Deng W, Ma J, Zhao J, et al. From street views to urban science: Discovering road safety factors with multimodal large language models. Transportation Research Part C: Emerging Technologies. 2026;188:105692. doi: 10.1016/j.trc.2026.105692
46. Mo B, Xu H, Ma R, Cho JH, Zhuang D, Guo X, et al. Large language models for travel behavior prediction. Transportation Research Interdisciplinary Perspectives. 2026;38:102124. doi: 10.1016/j.trip.2026.102124
47. Cao E, Tang JHCG, Yang X, Liu J, Zhuge C. LLM-aided representation of human commuting patterns in 42 Chinese cities with open data. Journal of Transport Geography. 2026;135:104749. doi: 10.1016/j.jtrangeo.2026.104749
48. Aghaabbasi M, Anjomshoae A. Comparative analysis of transportation policy frameworks for equity and accessibility using LLM-assisted methods: Evidence from Northeast Florida. Case Studies on Transport Policy. 2026;26:101952. doi: 10.1016/j.cstp.2026.101952
49. Liu T, Yang C, Chung H, Zhu M, Han K, Ding H. How can public concerns shape eVTOL policy design? A knowledge graph-enhanced large language model (LLM) framework. Transportation Research Part A: Policy and Practice. 2026;212:105185. doi: 10.1016/j.tra.2026.105185
50. Wu T, Chen Z, Yang S, Zhao J, Wei R, Jiang QM, et al. Generative AI for complex urban planning: Pathways, potentials, and challenges. Journal of Urban Management. 2026;15(3):1071-1088. doi: 10.1016/j.jum.2025.12.006

51. Leong WY. Generative AI-driven urban planning for net-zero emissions cities. IET Conf Proc. 2026;2025(53):51-52. doi: 10.1049/icp.2026.1919
52. Zali MH, Ghanbarei Baba ahmadi M, Argany M, Mostafavi MA. Leveraging Large Language Models for Automated Assessment and Mapping in Participatory Urban Planning. Int Arch Photogramm Remote Sens Spatial Inf Sci. 2026;XLIX-B4-2026(B4-2026):671-678. doi: 10.5194/isprs-archives-xlix-b4-2026-671-2026
53. Yan X, Dai T, Yu (Marco) Nie. Addressing the alignment problem in transportation policy making: an LLM approach. Transportation Research Part C: Emerging Technologies. 2026;188:105686. doi: 10.1016/j.trc.2026.105686
54. Kaur S, Sehra SS, Ebrahimi D, Wang X, Singh J, Sehra SK. Harnessing Large Language Models for Intelligent Transportation Systems: A Systematic Review. Multimodal Transportation. 2026;5(3):100308. doi: 10.1016/j.multra.2026.100308
55. Zhang Z, Sun Y, Wang Z, Nie Y, Ma X, Li R, et al. Large Language Models for Mobility Analysis in Transportation Systems: A Survey on Forecasting Tasks. Transportation Research Record: Journal of the Transportation Research Board. 2026;2680(2):756-774. doi: 10.1177/03611981251367699
56. Zhang C, Wei B, Yang L. TrafficMind: A System-Oriented Review of Large Language Models for Intelligent Transportation Systems. J Transp Eng, Part A: Systems. 2026;152(9):03126005. doi: 10.1061/jtepbs.teeng-9794
57. Yu J, Wang Y, Ma W. Large language model-enhanced reinforcement learning for generic bus holding control strategies. Transportation Research Part E: Logistics and Transportation Review. 2025;200:104142. doi: 10.1016/j.tre.2025.104142
58. Li S, Ma M, Azfar T, Ke R. ChatSUMO Agent: An LLM-Based Agent for Conversational Traffic Simulation in SUMO. Transportation Research Part C: Emerging Technologies. 2026;190:105759. doi: 10.1016/j.trc.2026.105759
59. Liu Y, Chung H, Song Y, Liu D, Chen T. From fastest to safer and more experience-aware: an LLM-driven framework for personalized multimodal routing. Transportation Research Part A: Policy and Practice. 2026;214:105290. doi: 10.1016/j.tra.2026.105290
60. Li Y, Li X, Li H, Zhang Y, Yuan Y, Li M. Decision generation for interlining bus scheduling: a large language model-empowered approach. Transportation Research Part C: Emerging Technologies. 2026;188:105640. doi: 10.1016/j.trc.2026.105640
61. Tu W, Li J, Xiao F, Li L. PA-LLMNeSy: Preference-Aware Neuro-Symbolic LLMs for robust Multi-Modal travel planning in MaaS. Transportation Research Part C: Emerging Technologies. 2026;189:105754. doi: 10.1016/j.trc.2026.105754
62. Zhang H, Dong H, Yang Z. TSGDiff: Traffic state generative diffusion model using multi-source information fusion. Transportation Research Part C: Emerging Technologies. 2025;174:105081. doi: 10.1016/j.trc.2025.105081
63. Shao Z, Xi H, Lu H, Wang Z, Bell MGH, Gao J. A spatial–Temporal Large Language Model with Denoising Diffusion Implicit for predictions in centralized multimodal transport systems. Transportation Research Part C: Emerging Technologies. 2025;179:105249. doi: 10.1016/j.trc.2025.105249
64. Qin Z, Wang L, Ling Y, Pereira FC, Ma Z. A foundational individual mobility prediction model based on open-source large language models. Transportation Research Part C: Emerging Technologies. 2026;185:105562. doi: 10.1016/j.trc.2026.105562
65. Vicente-Martínez P, Soria-Olivas E, Chust-Ros A, García-Escrivà MÁ, William-Secin E, Sánchez-Montañés M. Generative AI-Driven Digital Twin Architecture for Urban Mobility Simulation and Decision Support. Smart Cities. 2026;9(7):109. doi: 10.3390/smartcities9070109
66. Li X, Xian K, Wen H, Bai S, Xu H, Yu Y. PathGen-LLM: A Large Language Model for Dynamic Path Generation in Complex Transportation Networks. Mathematics. 2025;13(19):3073. doi: 10.3390/math13193073
67. Liang Y, Liu Y, Wang X, Zhao Z. Exploring large language models for human mobility prediction under public events. Computers, Environment and Urban Systems. 2024;112:102153. doi: 10.1016/j.compenvurbsys.2024.102153
68. Lee EH, Lee E. Electric vehicle charging station location selection using generative artificial intelligence. Transportation Research Part E: Logistics and Transportation Review. 2026;213:104930. doi: 10.1016/j.tre.2026.104930
69. Jia S, Ma C, Chi OH, Fan A. A socio-technical exploration of bias in generative AI travel planners. Tourism Management. 2026;116:105442. doi: 10.1016/j.tourman.2026.105442
70. Abou-Shouk M, Elbaz AM, Al-Leheabi SMZM, AbuElEnain E, Shabana MM. Trusting ChatGPT Usage in Personalized Travel Planning: The Moderating Role of Privacy and Data Security. Tourism and Hospitality Research. 2026;26(4):753-765. doi: 10.1177/14673584241306346
71. Wu J, Tian X, Liu Q, Li T, Liu C, Xu J, et al. Generative artificial intelligence and urban transport resilience: Rethinking sustainable mobility. Safety Science. 2026;198:107148. doi: 10.1016/j.ssci.2026.107148
72. Jin Y, Ma J. A survey of large language models in transportation planning: modelling, design and decision-making. Transportmetrica A: Transport Science. 2026:1-46. doi: 10.1080/23249935.2026.2631152
73. Yan Y, Zou B, Liao Y, Xu G, Yao R, Fan H, et al. Large language models for transportation research: Methodologies, state of the art, and future opportunities. Information Fusion. 2026;136:104546. doi: 10.1016/j.inffus.2026.104546
74. Vanegas-Nieto MS, Castro Brenes SB. AI in Qualitative Digital Research: A Systematic Mapping Review of Methodological Rigor, Ethics, and Human-AI Collaboration with Implications for Netnography and Digital Ethnography. SAP Netnography. 2026;4:270. doi: 10.62486/net2026270

FUNDING

No financing

CONFLICT OF INTEREST

None

AUTHORSHIP CONTRIBUTIONS

Drafting – original draft: Tammy Natalia Arellano Morales, Lusby Depablos-Machado, Juan Ernesto Pérez-Pérez.
Writing–review and editing: Tammy Natalia Arellano Morales, Lusby Depablos-Machado, Juan Ernesto Pérez-Pérez.