

A Dual-LLM Supervised Workflow for Replicating Agent-based Models: From NetLogo to Mesa

Boyu Wang¹, JoAnn S. Lee¹, and Andrew Crooks¹

University at Buffalo, Buffalo, NY, USA
bwang44@buffalo.edu

Keywords: replication · NetLogo · Mesa · validation · large language models

1 Introduction

Replicating agent-based models (ABM) across platforms (e.g., NetLogo, Mesa) can support replication, reuse, and cumulative model development in social simulation [2]. It can be methodologically risky, because independent implementations can diverge across languages and toolkits, even though they are based on the same conceptual model [2], thus validation remains critical. Large language models (LLMs) add a further methodological challenge. While they can reduce programming effort, LLMs as code-generation tools still demand interpretation, repair, and user oversight rather than blind acceptance [5]. We argue that workflows that lower the effort of producing cross-platform replications can help social simulation accumulate through reuse, comparison, and extension.

This extended abstract studies a Netlogo-to-Mesa replication of the Ya-TASERPS model [3]. Ya-TASERPS is a youth-focused ABM grounded in interactional theory of delinquency [4]. In this study, its v1.0.0 release developed using NetLogo is treated as the canonical behavioral source [3], with the paper and ODD as supporting documents. The research question is whether a human-in-the-loop design with two distinct LLM roles can support a credible NetLogo-to-Mesa migration whose validity is assessed directly against the canonical NetLogo model. Our main contributions include: 1) Separating planning and evaluation from implementation which offers a practical design for LLM-assisted model migration under explicit human review; 2) Ya-TASERPS provides a demanding but tractable case because its spatial structure, staged daily schedule, and family influence logic are rich enough to expose subtle porting errors; 3) The resulting Mesa model reproduces major structural, directional, and pattern-level features of the NetLogo source, while still leaving some numeric differences visible as active validation findings rather than concealed discrepancies.

2 Replication Workflow

The workflow uses two GPT-5.4 models with distinct LLM roles under continuous human mediation. One LLM role handles planning and evaluation, from

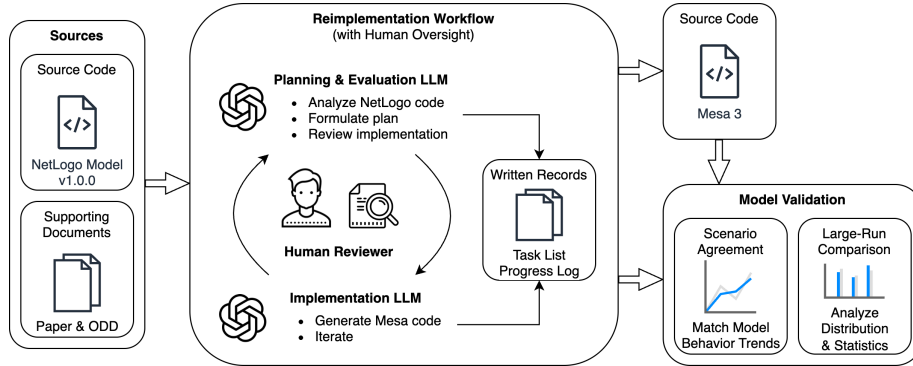


Fig. 1. The NetLogo-to-Mesa replication workflow utilizing two distinct LLM roles.

checking source code, decomposing tasks into phases, and reviewing whether implementations are consistent with the source NetLogo model. The other LLM role performs the actual implementations in phases, as outlined by the first LLM role. Each task phase defines acceptance criteria, proceeds through implementation and verification, and ends with human review before progression. The human researcher mediates both LLM roles, resolves ambiguities, and decides whether the resulting model state is acceptable. The aim is not towards full automation but to reduce prompt drift, limit uncontrolled rewrites, and keep generated code subordinate to explicit validation. Figure 1 summarizes the replication workflow while the full details are available at <https://github.com/wang-boyu/Ya-TASERPS>.

Traceability is supported by written records rather than chat history alone. Task plans and implementation logs preserve design choices, source-handling decisions, and validation results. Through such explicit artifacts, this workflow provides compact and inspectable record of how the Mesa model is implemented to align with the NetLogo source. The Ya-TASERPS model is used as a case study for this replication workflow as it combines a banded spatial layout (i.e., school, neighborhood, and home), staged daily dynamics, neighborhood movement, school and neighborhood opportunity processes, and family effects. These features expose the model to replication errors (e.g., when at-home logic activates), making it a practical stress test for our replication workflow.

3 Model Validation

Model validation is carried out sequentially at three levels, from structural parity checks, to directional validation of targeted scenarios, and then to large-run comparisons. Across these checks, agreement is the strongest between the structures of NetLogo and Mesa models. Specifically, the Mesa model reproduces the canonical world bounds $x \in [-16, 16]$ and $y \in [-20, 20]$, the same environment with 165 home, 1023 neighborhood, and 165 school patches, the default population of 150 youths, and the bounded risk and promotive setup ranges. It also preserves

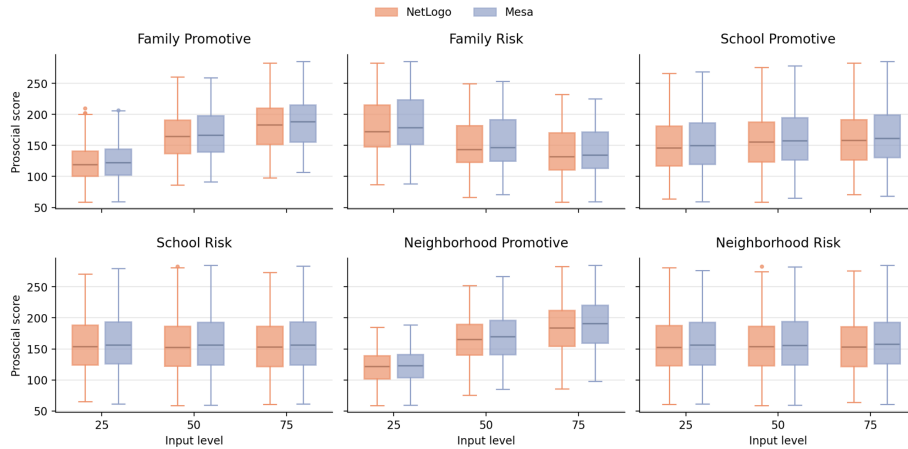


Fig. 2. Large-run prosocial outcomes across the six user-controlled inputs in NetLogo and Mesa with the same set of 729 parameter combinations, 10 random seeds, and 1800 days per run. The two implementations preserve the same broad significance and directional patterns.

the full daily sequence of `school` $\rightarrow$ `after_school` $\rightarrow$ `evening` $\rightarrow$ `step_day`. In terms of the `evening` procedure, it preserves the move-home-first logic, in which youths return home before the at-home update becomes active. These findings establish agreement between the source NetLogo model and the resulting Mesa model on the mechanisms generating downstream model behavior.

In directional scenarios, we perform model behavior checks on a small set of parameter bundles rather than a full parameter sweep. Using low, medium, and high input levels defined in the NetLogo experiment design, eight random seeds, and 60 simulated days per run, the Mesa model reproduces qualitative trends similar to those of the NetLogo model. Relative to a balanced baseline, risk-heavy settings generate more antisocial and less prosocial behavior, whereas promotive-heavy settings shift outcomes in the opposite direction. Targeted school-only and neighborhood-only shifts likewise change the mix of experiences in the expected directions. These short multi-seed checks indicate that the replication preserves the source model’s qualitative response to substantively important parameter changes, consistent with the overall sensitivity findings of the source model.

Large-run comparisons extend these results beyond a small parameter set. With the full set of 729 parameter combinations defined in the NetLogo model experiment, 10 random seeds, and 1800 simulated days per run, NetLogo and Mesa models agree on the significant pattern across all six user-controlled inputs in large-run ANOVA comparisons of prosocial outcomes. Results are shown in Fig. 2. The same agreement holds for antisocial outcomes under the same parameter sweep. This broader agreement suggests that the Mesa replication captures the same high-level sensitivities as the canonical NetLogo model, not only under a few hand-picked scenarios but across a wide experimental surface.

4 Discussion and Conclusion

Our study suggests that LLM-assisted model replication can be credible when it is treated as a methodological design problem rather than a coding shortcut. Using Ya-TASERPS as a case study, separated LLM roles help stabilize the tasks, written records make source-handling and validation decisions inspectable, and direct comparison to the NetLogo source prevents generated code from standing in for scientific judgment. This emphasis on human review is consistent with broader evidence that code-generation systems can be useful while still requiring substantial interpretation and repair by users [5].

The study also has clear limitations. It concerns a single model, it requires substantial human involvement, and it does not show that the same design will transfer successfully to other models or other LLM configurations. The evidence is strongest for structural, directional scenario, and pattern-level agreement across large experimental surfaces, and weaker for exact level matching and coefficient identity. Even so, this study supports a modest conclusion of methodological interest for social simulation: dual-LLM assistance can contribute to NetLogo-to-Mesa migration when it is constrained by explicit source priority, written decision records, human review, and direct NetLogo-versus-Mesa validation. More broadly, this study aligns with longstanding arguments that reimplementation, careful documentation, and explicit validation are central to credible agent-based modeling rather than optional afterthoughts [2,1]. Moreover, our methodology could be generalized and applied to numerous ABMs supporting replication and further model development on the platform of the users choice rather than being restricted to the initial implementation of the model.

References

1. Collins, A., Koehler, M., Lynch, C.: Methods that support the validation of agent-based models: An overview and discussion. *Journal of Artificial Societies and Social Simulation* **27**(1), 11 (2024). <https://doi.org/10.18564/jasss.5258>
2. Edmonds, B., Hales, D.: Replication, replication and replication: Some hard lessons from model alignment. *Journal of Artificial Societies and Social Simulation* **6**(4), 11 (2003), <https://www.jasss.org/6/4/11.html>
3. Lee, J., Crooks, A.T.: Youth and their artificial social environmental risk and promotive scores (ya-taserps): An agent-based model of interactional theory of delinquency. *Journal of Artificial Societies and Social Simulation* **24**(4), 2 (2021). <https://doi.org/10.18564/jasss.4660>
4. Thornberry, T.P.: Toward an interactional theory of delinquency. *Criminology* **25**(4), 863–892 (1987). <https://doi.org/10.1111/j.1745-9125.1987.tb00823.x>
5. Vaithilingam, P., Zhang, T., Glassman, E.L.: Expectation vs. experience: Evaluating the usability of code generation tools powered by large language models. In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. pp. 1–7 (2022). <https://doi.org/10.1145/3491101.3519665>