Homophily under model substitution: A scaled-down Claude Haiku replication of He et al. (2026)

Sonny Fullerton | May 2026

Research question

He et al. (2026) report that GPT-3.5-powered agents on Chirper.ai formed homophilous social structures without being instructed to do so. This short replication asks whether the core mechanism generalises across model families and time: if the same type of engagement environment is rebuilt with Claude Haiku 4.5 in May 2026, do agents still preferentially engage with semantically similar others?

Design

I ran a topic simulation rather than a faithful clone of Chirper. The goal was to preserve the paper's central mechanism--agents posting, reading a mixed feed, and choosing whom to engage with--while keeping the build small enough to inspect end to end. The completed canonical run, full-100x12, used 100 Claude Haiku 4.5 agents over 12 rounds on the topic: Should universities ban AI in coursework?

Agents were balanced across five latent perspectives: pro-ban traditionalists, anti-ban accelerationists, pragmatic reformers, sceptical empiricists, and indifferent generalists. The first ten generated personas were inspected before the full rounds began. They were accepted because they were topic-specific, varied in profession and context, and did not contain instructions to prefer similar others. In each round, every agent wrote a short post and then chose 1-3 posts from a mixed feed to like, follow, or ignore.

The measurement stack follows the original paper where possible. Agent post histories were embedded with sentence-transformers/all-MiniLM-L6-v2, the same embedding model used by He et al. Weighted engagement graphs were analysed with Louvain community detection, modularity, assortativity by detected community, and a 100-iteration degree-preserving bootstrap null.

Results

Metric	He et al. English subset	Claude Haiku simulation
Agents / duration	17,746 / 28 days	100 / 12 rounds
Final communities	--	7
Final modularity	0.38	0.131
Bootstrap 95% null interval	--	[0.101, 0.117]
Bootstrap p-value	< .001	0.000
Final assortativity	0.61	0.069
Content-engagement correlation	significant in paper	0.018

The final engagement graph's modularity was 0.131, above the upper bound of the degree-preserving bootstrap null interval [0.101, 0.117], with p = 0.000 at three decimals. Assortativity was positive (0.069), and the content-engagement correlation was also positive but weak (0.018). These values are much smaller than the original Chirper results, but they are directionally consistent and statistically above the graph null used here.

Classification: the core finding reproduces, with attenuated magnitude. The simulation does not match the scale, platform richness, or effect size of the original paper. It does show that the mechanism is not obviously confined to GPT-3.5, the Chirper platform, or the 2023 data collection window.

Interpretation

The result is useful because it tests the part of the finding that matters for synthetic-audience systems: interaction can change a simulated population. Even when personas begin balanced across perspectives, repeated engagement choices can create structure. In a commercial setting, that means a synthetic audience is not only a static panel of persona cards; it is a social process that can concentrate attention and amplify similarity over time.

For a system such as Radiant, the product implication is not that homophily is a flaw. Human groups are homophilous too. The practical risk is unobserved diversity collapse: a simulated buyer committee, shareholder base, or policy audience may begin diverse but become less diverse through repeated rounds of interaction. A dashboard that surfaces modularity, assortativity, and content-similarity engagement over time would make that drift visible to users.

Limitations

This is a scaled-down replication, not a full reproduction. It uses one seed topic, 100 agents, and 12 rounds, while He et al. analysed tens of thousands of agents over 28 days on a richer social platform. The feed is a compact topic simulation approximation rather than a live Twitter-like system with multi-topic dynamics. Community detection also differs: I used Louvain because it is a modern weighted-graph default, while the paper reports label propagation and fast-greedy variants. Finally, the effect size is modest. The honest claim is generalisation of direction and statistical signal, not magnitude matching.

Conclusion

Re-running the homophily mechanism with Claude Haiku 4.5 produces weak but statistically detectable homophilous structure. The result supports the paper's call for model-family replications and suggests a concrete diagnostic for synthetic-audience products: track whether simulated populations remain diverse after they begin interacting.

Artifacts. Dashboard: https://societiesdemo-egqj7tmwe-sonnyfully12-icloudcoms-projects.vercel.app. Saved run id: full-100x12. Raw run file: backend/runs/full-100x12.json.