

Exploring the Use of Local Large Language Models to Generate Guidance and Stylized Facts for Agent-Based Models

Luke J. Osterritter¹[0000-0002-4447-9928] and Kathleen M. Carley¹[0000-0002-6356-0238]

¹ Carnegie Mellon University, Pittsburgh PA 15213, USA
losterritter@cmu.edu, carley@cs.cmu.edu

Abstract. Where can modelers find value in using large language models (LLMs) in the creation of agent-based models (ABMs)? We seek to begin answering this question by evaluating the output of several small open-weight LLMs on a set of four elicitation prompts regarding the creation of a model of insider threat. These prompts have two purposes: guidance for the creation of ABMs, and generation of “stylized facts” – empirical regularities that are generally regarded as true even if they don’t hold for every specific case – for the purpose of model validation. For each case we employed two prompt approaches: zero-shot, wherein a directive and schema for results are provided and nothing else; and few-shot, wherein examples from other domains are provided. The models are capable of adhering to structure, with 93% returning the requested JSON schema. The few-shot approaches tended to worsen model guidance fidelity as they led models to return unrelated information. Finally, we find that asking these models to provide empirical sources results in fictitious citations 62% of the time, and only 3.6% of the sources could be verified. We conclude that using these relatively small local models can be useful for structured ideation but are less useful for grounding work in empirical evidence.

Keywords: Large Language Models, Agent-based Models, Validation, Stylized Facts, Insider Threat.

1 Introduction

The creation of an agent-based model is often no small feat. While there do exist platforms that allow creators to draw on vast libraries of existing models, creating a model in one’s field of interest still requires substantial domain knowledge and significant coding expertise. Modelers must identify appropriate micro-level behaviors, interaction rules, decision heuristics, and empirical regularities the model should adhere to. This knowledge collection is time-consuming, ad hoc, and often requires identification and participation of subject domain experts. Absent this work, the ability for the modeler to create something that usefully reflects the empirical phenomena (and their ability to assess the model’s validity) is weakened dramatically.

We have identified two key areas wherein the use of LLMs could assist with and improve the process of creating ABMs: model guidance and model validation. While there are several popular frontier models offered as commercial platforms that are likely up to this task, they are often cost-prohibitive, constantly evolving, sometimes ill-suited to security-conscious work, and have the possibility to retrain themselves on provided prompts, which could potentially poison the results of a study like this one. Further, it is an interesting research question to ask: what value can you derive for agent-based model creation and validation from large language models capable of running on personal hardware?

We ground this study in the domain of insider risk and threat; that is, the potential for an individual who has or had authorized access to an organization’s critical assets to use their access, either maliciously or unintentionally, to act in a way that could negatively affect the organization [1]. This area of study has many modeling efforts, both agent-based and system dynamics, applied to it over the years [2–6], and has an evidence base of academic literature spread across many disparate disciplines [7]. Further, empirical data surrounding insider incidents can be difficult to obtain, and organizations are often reluctant to share or otherwise publicize when such events occur [8]. Given these points, we believe this domain lends itself well to the testing of model guidance generation and of “stylized facts”, empirical regularities that hold broadly, though not necessarily in each individual case [9], that can be used to ensure model output aligns with empirical phenomena.

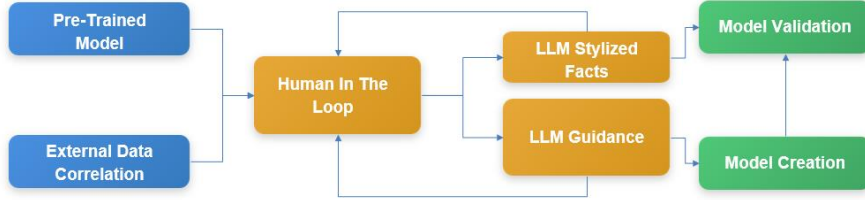


Fig. 1. Conceptual Model for using LLMs in the Modeling Process

The conceptual model in figure 1 describes the overall flow of using LLMs for ABMs. Modelers, in concert with domain experts, use their knowledge of modeling processes and their domains of interest to craft prompts appropriate to the goals of generating guidance and empirical facts for validation. In the case of using frontier models, models that can otherwise search outside their pre-trained knowledge through the use of Retrieval-Augmented Generation (RAG) to supplement with appropriate corpora, the modeler can further ensure the usefulness of output. Generated guidance is used for the creation of the model, which can then be validated based on the empirical truths of the generated stylized facts. Throughout this whole process, the modeler and any domain experts remain as the humans-in-the-loop; we do not suggest to simply allow the LLMs to generate agent-based models themselves but rather suggest that they are used

as another tool in the modeler’s arsenal for which they provide valuable context, subject matter expertise, and control.

2 Related Work

Exploratory scholarship regarding the appropriate use of LLMs for ABM creation has only recently begun to emerge. Vanh  e et al. provide the most complete look at the issue thus far [10], contributing a rapid literature review on how LLMs are being used for ABMs as of now, as well as providing some foundational guidance on how they could be used in the future. Other work has explored how LLMs can be used to generate the actual code for an ABM [11, 12] and how LLMs can create, evaluate, and validate ABMs [13]. A great many other works explore the use of LLMs as the decision making function of agents themselves within simulations [14–17]. Other work has explored using LLMs for studying social dynamics [18], creating synthetic social media data [19], and even an ABM that explores how autonomous agents operate [20].

That some amount of auditing is needed for model output is established for large models: fabrication has been measured in ChatGPT bibliographies [21]. Self-reported confidence, which our prompts ask for, is known to be systematically overconfident [22]. What is missing is measurement at the small open-weight end, on a task where a citation is not simply a decoration but the entire justification for a modeling choice. Our few-shot conditions rest on in-context learning, where demonstrations are known to convey label space and format more than task content [23]; our few-shot contamination result described below is likely a function of failed domain transference between the example and the prompted subject.

3 Methodology

3.1 Task Formulation

For this study, we focus on two tasks. The first is generating model guidance: that is, asking the LLM to work out which types of agents are needed, what attributes they should take, how they interact with one another, how they make their decisions, what environmental parameters should be modeled, which ranges values can take and how they are to be initialized, what assumptions are to be made, and which empirical sources these decisions are drawn from.

The second task involves the creation of stylized facts. Stylized facts are used by modelers as a way to validate their models, ensuring that their output adheres to these empirical regularities as a way of demonstrating their applicability to the empirical phenomena they represent [24–26].

3.2 Prompt Design

The set of prompts is broken into two sets of two. They each share a preamble that frames the work as being for academic research, calls for results to be returned strictly

in a defined JSON schema with no further context or commentary, instructs the model to refrain from inventing sources, and asks the model to self-identify which version it is and as well provide a timestamp of the activity.

The prompts are P1 (ABM, zero-shot), P2 (facts, zero-shot), P3 (ABM, few-shot) and P4 (facts, few-shot); the few-shot exemplars come from deliberately unrelated domains: an abbreviated residential burglary ABM for P3 [27] and abbreviated financial return stylized facts for P4 [28], each with an explicit instruction not to copy its content.

The model guidance tasks (P1 and P3) request nine items: agent types with attributes, behavioral rules, decision rules, interactions, environmental parameters, initialization values, variable ranges, key assumptions, and references.

The stylized-fact tasks (P2 and P4) request an array of facts, each typed against a five-category taxonomy (Temporal, Spatial, Distributional, Relational, Sequence) and carrying a statement, qualitative pattern, evidence basis, citation (in Chicago format), confidence level, modeling considerations, and caveats.

We used a fully crossed “7 models $\times$ 4 prompts $\times$ 3 replicates” design ($n = 84$) with settings held constant (temperature 0.7, top-p 1.0, max_tokens 4096, no fixed seed). We initially used a lower max_tokens threshold, but the chain-of-thought models required a higher value to reliably return results.

3.3 Model Selection

The models used in this study were selected to balance performance, size, hardware limitations, and comparability. The models are deepseek-r1-8b, gemma4-e4b, llama3.1-8b, llama3.2-3b, mistral-7b, phi4-mini-3.8b and qwen3.5-9b. Three of these have noticeably fewer parameters at 3-4 billion and are included both to capture the largest variety of models that can run on our available hardware, as well as to compare against the larger models to gauge efficacy in the tasks under test. We used Ollama [29] to procure the models and run our prompts against them locally.

3.4 Output Verification

Every string in a references array or a stylized fact’s source field is a claim about empirical sources. All 390 of these were normalized and triaged. Three classes need no external verification as they already violate the instruction to supply primary sources in Chicago format: placeholder (filler identities such as “Doe”), non-citation (a topic label rather than a work), and author-year stub. The 275 distinct, citable entries were checked against Crossref, DOI resolution, publisher and repository pages, and against whether the named venue exists at all, each receiving one final disposition: verified, partial (a real work with erroneous metadata), fabricated, or unverifiable. We used a frontier LLM, Anthropic’s Claude (Opus 5), to perform these verification tasks and to gather summary metrics, with human spot-check validation.

4 Results

4.1 Schema Conformity

Across 84 runs, 78 (92.9%) returned strictly schema-valid JSON (Fig. 2a). Five models were schema-valid on every run, while qwen3.5-9b returned valid structure on 11 of 12 (91.7%) runs and deepseek-r1-8b on 7 of 12 (58.3%) runs. All six failures are the same failure: a structurally complete JSON object containing a syntax error, five of them an unescaped quotation mark inside a string value. None is truncated, empty or off task, so these runs are still able to be parsed with corrections. Note that an initial test run with a lower max_tokens threshold caused both gemma4-e4b and deepseek-r1-8b, both “reasoning models” that return a chain-of-thought walkthrough of their logic, to fail to generate the appropriate JSON results on most runs.

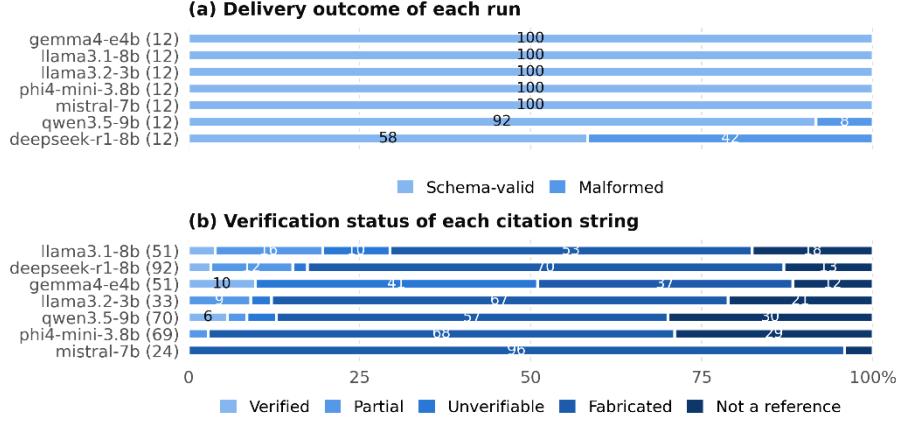


Fig. 2. (a) Percentage of schema-valid output; (b) Verification status of all citations.

4.2 Citation Accuracy

Of the 390 citations generated, 242 (62.1%) name works that do not exist; 76 (19.5%) are not references at all but rather author-year stubs, placeholder identities, or bare topic labels; 26 (6.7%) are partial, a real work with erroneous metadata; 32 (8.2%) are unverifiable; and 14 (3.6%) corresponding to twelve distinct works are verified. Together 81.5% of reference slots are unsubstantiated and 8.2% were unverifiable as written. Fabrication varies by model from 37.3% (gemma4-e4b) to 95.8% (mistral-7b; see Fig. 2b); the best performer still invented more than a third of its references.

The fabrications tend to seem very plausible: they carry correct Chicago punctuation, plausible surnames, credible journal names, and realistic volume and page ranges. Listed venues repeatedly did not exist, and of the seventeen citations carrying a DOI, none resolved successfully. A characteristic seems to be unresolved placeholders: 18 of 69 phi4-mini-3.8b citations are attributed to a “Smith” or a “Doe”, complete with

journals, volumes and pages: “Doe, A. (2019). Geography and Insider Threat: A Case Study Analysis”.

The twelve verified works are largely institutional documents that are indeed of importance to the insider threat community of practitioners and researchers: NIST special publications, ISO/IEC 27001, the Verizon Data Breach Investigations Report, ENISA’s Threat Landscape. Some partial successes pair a real institution and report series with the wrong year, title or author. Institutional reports and white papers tend to be retrievable while the peer-reviewed scholarship of the field tends to be fabrication. The stylized fact runs contain a mean of 7.3 citations versus 2.2 for ABM guidance runs, explainable by the schema requiring a source for every stylized fact. The few-shot condition reduces fabrication on the fact task by about 30% (77.1% for P2 against 47.9% for P4), making it the only prompt-level intervention that improves citation integrity, though it is confounded somewhat by the shorter outputs the few-shot prompts produce.

4.3 Zero-shot vs. Few-shot Outcomes

Because the exemplars come from unrelated domains, any exemplar-specific material in an insider-threat answer is an instruction violation with known provenance. The longest word sequence shared with the burglary exemplar rises from a mean of 2.14 words when the exemplar is not provided to 4.48 when it is; of 42 few-shot runs, 19 (45.2%) show carryover, 15 ontological and 4 verbatim. The verbatim cases are systematic: all three llama3.2-3b P3 runs reproduce the exemplar’s interactions clause intact: “recent successes raise local repeat and near-repeat victimization risk,” merely substituting employees and contractors for offenders and targets. The ontological cases are worse: awareness space, a crime-pattern-theory construct denoting an offender’s habitual geography, appears as a declared agent attribute in nine P3 runs, appearing in every replicate of three model families.

4.4 Output Selections

Below is an example of the agent output from one run of the gemm4-e4b model. It captures human insider actors, resources of IT systems, and security of the organization.

Table 1. Agents and Attributes Generated by gemm4-e4b P3 Run

Agent	Attributes
Insider_Actor	"access_level", "motive_strength", "frustration_level", "awareness_space", "knowledge_gaps"
System_Resource	"data_sensitivity", "data_volume", "access_controls", "logging_depth"
Organization_Security	"detection_efficiency", "response_speed", "monitoring_coverage", "policy_strictness"

The attribute “awareness_space” is the aforementioned contamination from the P3 few-shot exemplar of crime. Otherwise, this generation adheres to the prompt instructions and might provide value to the modeler.

This output from llama3.1-8b provides a meaning for each attribute, making the results more immediately intelligible and useful.

Table 2. Employee Agent, Attributes, and Meanings Generated by llama3.1-8b

Agent	Attribute	Meaning
Employee	Access Level	Employee's level of access to sensitive resources
	Motivation	Employee's reasons for insider threat behavior (e.g., financial gain, revenge)
	Seniority	Employee's length of service within the organization

Table 3. Selection of Generated Stylized Facts

Model	Stylized Fact	Category	Citation	Disposition
gemma4-e4b	"The peak risk of insider threat activity often occurs during periods of organizational transition, such as departmental restructuring, mergers, or immediately following an employee's notice of resignation."	Temporal	Brown, T.L.	Fabricated
qwen3.5-9b	"Malicious insider campaigns typically follow a sequence: reconnaissance → privilege escalation → lateral movement → exfiltration, with an average dwell time of 180 days before detection."	Sequence	Thompson, R., et al. (2021)	Fabricated
deepseek-r1-8b	"There is a strong positive correlation between user privilege level and the potential damage of insider threat incidents."	Relational	ENISA 2020	Verified
llama3.1-8b	"Insider threats tend to follow a power-law distribution, with a few high-impact incidents and many low-impact incidents."	Distributional	Barabási, A. L. (2002)	Verified
gemma4-e4b	"The highest concentration of privileged access and sensitive data is typically found within specific, structurally defined units (e.g., R&D, Executive Leadership, Finance), creating high-value, localized hot spots for threats."	Spatial	Brown, T.L.	Fabricated

5 Discussion

As a tool to assist with structured ideation for the construction of agent-based models, generating candidate agent types, attributes, behaviors, and interaction rules, these smaller models running on personal hardware can be of value to the modeler. 5 of the 7 models returned schema-valid JSON output on every run, while the remaining results were parsable with some cleaning. 81.5% of references were unsubstantiated, no generated DOIs were resolvable, model-reported confidence values were uninformative, and generated numeric ranges unsourced. The instruction to refrain from inventing sources was largely ignored. As stylized facts are meant to represent some empirical regularities, our inability to corroborate them in most cases by tracing them back to a primary source significantly weakens their value to the modeler and fails to engender trust.

Analogous-domain exemplars tended to poison model output. Using a worked example from another field to convey output format bought no compliance improvement while contaminating 45% of few-shot runs, importing a construct with no value to the Insider Threat domain into nine agent definitions across three model families, though a content-neutral skeleton may have provided format without contamination.

6 Conclusion and Future Work

Small, open-weight models are competent at structure. They follow a demanding JSON schema reliably, fill fields appropriately, adhere to a supplied taxonomy, and reconstruct a coherent insider-threat agent ontology. They are much less competent at supplying evidence for their stated empirical regularities: 62.1% of the citations they produce name works that do not exist and only 3.6% are verifiable, yet they declare high confidence regardless. A mitigation may be to use small local models to generate modeling choices, and treat every claim attached to them as an unverified until verified.

This study looked specifically at small local LLMs capable of running on modest personal hardware. This is not to say that LLMs are, in general, ill fit for these tasks: it is possible that larger models with more parameters may perform these tasks better, and that adding in RAG functionality for external search or specially curated, domain-specific corpora may prove valuable as well. Supplying a curated corpus and requiring every claim to cite items within it would convert open-ended generation into constrained attribution and would avoid the problem of fictitious citations.

Acknowledgments. This research was supported through the Minerva Research Initiative, in partnership with Office of Naval Research under grant #N000142114012; the Office of Naval Research under grant #N00014241256; and by the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University. The views and conclusions are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research or the U.S. Government.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Costa, D.: CERT Definition of “Insider Threat” - Updated, <https://insights.sei.cmu.edu/blog/cert-definition-of-insider-threat-updated/>, (2017).
2. Sokolowski, J.A., Banks, C.M., Dover, T.J.: An agent-based approach to modeling insider threat. *Comput Math Organ Theory*. 22, 273–287 (2016). <https://doi.org/10.1007/s10588-016-9220-6>.
3. Ostertter, L.J., Carley, K.M.: Modeling Interventions for Insider Threat. In: *Social, Cultural, and Behavioral Modeling: 13th International Conference, SBP-BRIMS 2020, Washington, DC, USA, October 18–21, 2020, Proceedings*. pp. 55–64. Springer-Verlag, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-030-61255-9_6.
4. Moore, A.P.: Modeling the Influence of Positive Incentives on Insider Threat Risk Reduction. Presented at the 35th International Conference of the System Dynamics Society and 60th Anniversary of System Dynamics Celebration (2017).
5. Moore, A.P., Mundie, D.A., Collins, M.L.: A System Dynamics Model for Investigating Early Detection of Insider Threat Risk. In: *Conference Proceedings of the 31st International Conference of the System Dynamics Society* (2013).
6. Moore, A.P., Cappelli, D.M., Caron, T.C., Shaw, E., Spooner, D., Trzeciak, R.F.: A Preliminary Model of Insider Theft of Intellectual Property: *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*. 2, 28–49 (2011). <https://doi.org/10.22667/JOWUA.2011.03.31.028>.
7. Ostertter, L.J., Carley, K.M.: A Comprehensive Bibliometric Analysis of Insider Threat Scholarship. *IEEE Transactions on Computational Social Systems*. 1–17 (2026). <https://doi.org/10.1109/TCSS.2026.3705576>.
8. Glasser, J., Lindauer, B.: Bridging the Gap: A Pragmatic Approach to Generating Insider Threat Data. In: *2013 IEEE Security and Privacy Workshops*. pp. 98–104 (2013). <https://doi.org/10.1109/SPW.2013.37>.
9. Arroyo Abad, L., Khalifa, K.: What are stylized facts? *Journal of Economic Methodology*. 22, 143–156 (2015). <https://doi.org/10.1080/1350178X.2015.1024878>.
10. Vanh  e, L., Borit, M., Siebers, P.-O., Cremades, R., Frantz, C., G  rcan,   ., Kalvas, F., Kera, D.R., Nallur, V., Narasimhan, K., Neumann, M.: Large Language Models for Agent-Based Modelling: Current and Possible Uses Across the Modelling Cycle. In: Verhagen, H., Chappin, E., and Scholz, G. (eds.) *Advances in Social Simulation*. pp. 601–619. Springer Nature Switzerland, Cham (2026). https://doi.org/10.1007/978-3-032-31712-4_43.
11. Frydenlund, E., Mart  nez, J., Padilla, J.J., Palacio, K., Shuttleworth, D.: Modeler in a box: how can large language models aid in the simulation modeling process? *SIMULATION*. 100, 727–749 (2024). <https://doi.org/10.1177/00375497241239360>.
12. Mart  nez, J., Llinas, B., Botello, J.G., Padilla, J.J., Frydenlund, E.: Enhancing GPT-3.5’s Proficiency in Netlogo Through Few-Shot Prompting and Retrieval-Augmented Generation. In: *2024 Winter Simulation Conference (WSC)*. pp. 666–677 (2024). <https://doi.org/10.1109/WSC63780.2024.10838967>.

13. Möltner, T., Manzl, P., Pieber, M., Gerstmayr, J.: Creation, evaluation and self-validation of simulation models with large language models. *Neurocomputing*. 663, 132030 (2026). <https://doi.org/10.1016/j.neucom.2025.132030>.
14. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W.X., Wei, Z., Wen, J.: A survey on large language model based autonomous agents. *Front. Comput. Sci.* 18, 186345 (2024). <https://doi.org/10.1007/s11704-024-40231-1>.
15. Gao, C., Lan, X., Li, N., Yuan, Y., Ding, J., Zhou, Z., Xu, F., Li, Y.: Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanit Soc Sci Commun*. 11, 1259 (2024). <https://doi.org/10.1057/s41599-024-03611-3>.
16. Vezhnevets, A.S., Agapiou, J.P., Aharon, A., Ziv, R., Matyas, J., Duéñez-Guzmán, E.A., Cunningham, W.A., Osindero, S., Karmon, D., Leibo, J.Z.: Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia, <http://arxiv.org/abs/2312.03664>, (2023). <https://doi.org/10.48550/arXiv.2312.03664>.
17. Plaat, A., Duijn, M. van, Stein, N.V., Preuss, M., Putten, P. van der, Batenburg, K.J.: Agentic Large Language Models, a Survey. *Journal of Artificial Intelligence Research*. 84, (2025). <https://doi.org/10.1613/jair.1.18675>.
18. Ng, L.H.X., Carley, K.M.: Modeling Social Dynamics with an LLM-Enabled Agent Based Network-Dynamic (LAND) Model, <http://arxiv.org/abs/2608.00929>, (2026). <https://doi.org/10.48550/arXiv.2608.00929>.
19. Ng, L.H.X., Kang, B.N.Y., Carley, K.M.: AuraSight: Generating Realistic Social Media Data, <http://arxiv.org/abs/2509.08927>, (2025). <https://doi.org/10.48550/arXiv.2509.08927>.
20. Ng, L.H.X., Carley, K.M.: BotSim: Mitigating the Formation of Conspiratorial Societies with Useful Bots. *JASSS*. 29, 4 (2026).
21. Walters, W.H., Wilder, E.I.: Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Rep*. 13, 14045 (2023). <https://doi.org/10.1038/s41598-023-41032-5>.
22. Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., Hooi, B.: Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. *International Conference on Learning Representations*. 2024, 23650–23678 (2024).
23. Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., Zettlemoyer, L.: Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?, <http://arxiv.org/abs/2202.12837>, (2022). <https://doi.org/10.48550/arXiv.2202.12837>.
24. Meyer, M.: How to Use and Derive Stylized Facts for Validating Simulation Models. In: Beisbart, C. and Saam, N.J. (eds.) *Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives*. pp. 383–403. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-319-70766-2_16.
25. Meyer, M.: Bibliometrics, Stylized Facts and the Way Ahead: How to Build Good Social Simulation Models of Science? *JASSS*. 14, 4 (2011).

26. Janssen, M.A., Ostrom, E.: Empirically Based, Agent-based models. *Ecology and Society*. 11, (2006).
27. Brantingham, P.L., Brantingham, P.J.: *Environment, Routine, and Situation: Toward a Pattern Theory of Crime*. In: *Routine Activity and Rational Choice*. Routledge (1993).
28. Rama Cont: Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*. 1, 223 (2001). <https://doi.org/10.1088/1469-7688/1/2/304>.
29. Ollama: Ollama: Get up and running with large language models, <https://ollama.com>, (2026).