Social Simulation Conference - SSC2026 - 24-28 August 2026

Large Language Models as Supervised Extraction Assistants

Lowering the Barrier to Documentation Standard Adoption in Agent-Based Modelling

Peer-Olaf Siebers. School of Computer Science. Nottingham University. UK
Christopher Frantz. Norwegian University of Science and Technology (NTNU), Norway

Last update: 29/08/2026

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

LLM4ABM

The Problem: Many Standards, Limited Adoption

- ABM credibility depends on transparent reporting of methods and data use

- Reporting tools exist, but adoption remains weak
  - ODD Protocol (Grimm et al 2006)
  - ODD+D Protocol (Müller et al 2013)
  - RAT Standard (Achter et al 2022)
  - …

- Evidence and surveys consistently point to effort as the main barrier
  - Bruce Edmonds @ SSC2026 "What we need is minimal but most helpful work"
  - Documentation is often treated as supplementary, not core

A Tragedy of the Commons

- Collective benefits of reporting standard and protocol[1] adoption are substantial…
  - Greater rigour in presenting models
  - Faster credibility assessment via traceability
  - Greater likelihood of reuse and adaptation
  - Strengthened credibility of social simulation as a field
- … but the individual's trade-off is poor
  - High effort (time/training) to comply with a standard
  - Benefit mostly collective and delayed, not immediate for the individual contributor

[1] In the following, when we refer to standard we mean 'reporting standard and protocol', unless we provide specific names

Result: Rational individuals under-contribute, even though the field as a whole would be better off if everyone did contribute

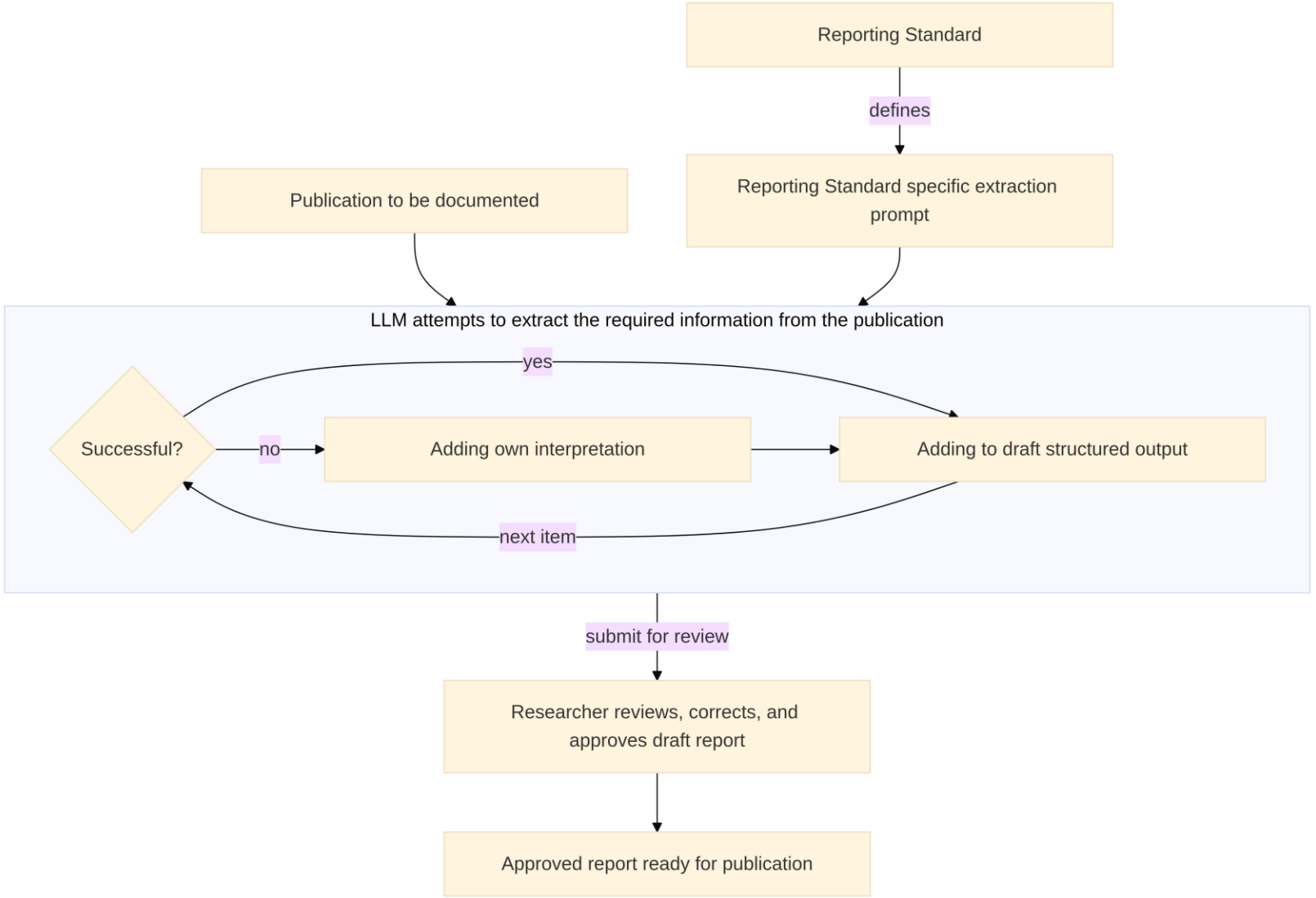
The Position: Supervised Extraction at Scale

Human-supervised LLMs can make the adoption of standards practical at scale

- Supervised extraction: The LLM performs initial data collection
- The researcher validates, corrects, and approves the output

Researcher role shifts from producer to reviewer

High Level Process Flow



Case and Methodology: A Feasibility Study

- Goal
  - A controlled feasibility test, not a generalisable claim
- Focus
  - RAT-RS: Question-centred and therefore well suited to LLM interaction
- Test case (Siebers & Aickelin 2011)
  - Modelling Staff Proactiveness in Retail Simulation Models
    - An existing RAT-RS example provides a human-authored benchmark

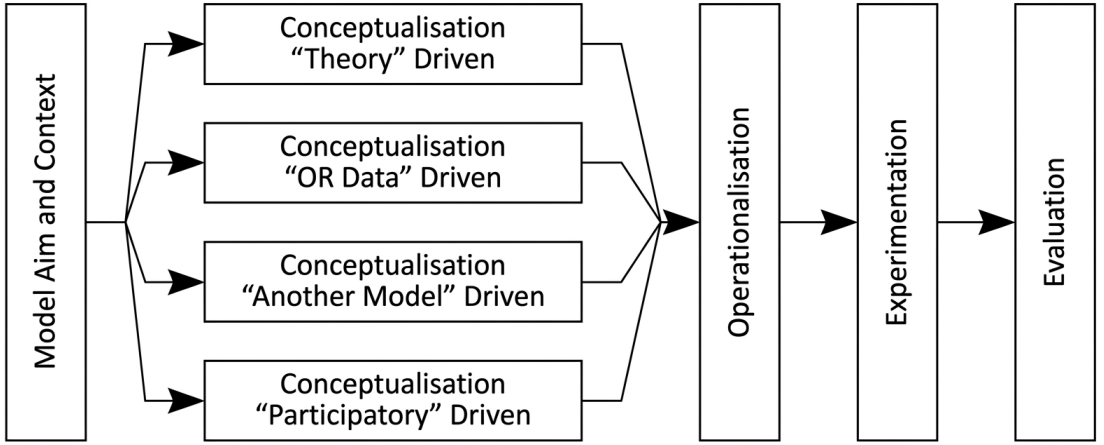
Case and Methodology: A Feasibility Study

- Execution
  - Same prompt and publication given to different LLMs via chat interfaces

- Analysis
  - Cross human/LLM & LLM/LLM comparison
  - Intra-model consistency across repeated runs

The RAT-RS

- Rigour And Transparency Reporting Standard (Achter et al 2022)
  - Improves data-use documentation across the ABM cycle
  - Organised as a Question Suite (QS) toolbox
  - Four model-development 'flavours':
    - Theory-driven
    - OR-data-driven
    - Another-model-driven
    - Participatory-driven



Test Case

- Our feasibility study uses the OR-data-driven flavour

- Original RAT-RS: 39 questions, five QSs, covering the full ABM cycle
- Extended RAT-RS: 44 questions, six QSs: original standard plus five meta-evaluative

  - QS1. Model Aim And Context Questions
  - QS2. Conceptualisation Questions: What And Why?
  - QS3. Operationalisation Questions: How And Why?
  - QS4. Experimentation Questions
  - QS5. Evaluation Questions
  - QS6. Reporting Quality Assessment (Supplementary LLM-driven QS)

1 MODEL AIM AND CONTEXT QUESTIONS

- Q1.1 If this RAT-RS use is related to a specific publication, please provide a reference to that publication.
- Q1.2 What is the purpose of the model? (prediction, explanation, description, theoretical exploration, illustration, analogy, social interaction, or other (please specify); for an explanation of model purposes, see Edmonds et al (2019) http://jasss.soc.surrey.ac.uk/22/3/6.html)
- Q1.3 What domain does the model research?
- Q1.4 What (research) question(s) is the model addressing? (in general / in this publication)
- Q1.5 What is the MAIN driver for your initial model development step? (theory(s), empirical evidence, existing model(s), participatory modelling data)
- Q1.6 Explain why this MAIN driver was chosen.
- Q1.7 What is the target system that this model reproduces? (briefly describe the target system and its boundaries)
- Q1.8 Explain why this target system and these boundaries were chosen.

2 CONCEPTUALISATION QUESTIONS

- Q2.1 What key mechanism(s) of the target system is/are used as driver(s) in this model? (examples: queuing system; physical/social/political network)
- Q2.2 Why is/are this/these key mechanism(s) used?
- Q2.3 What are the key elements of the target system?
- Q2.4 Which of these key elements were mapped into model elements? (distinguish (at least) between agents, environment, and relationships/interactions among any combination of these)
- Q2.5 Explain why key elements of the target system were included, excluded, or changed in the model.
- Q2.6 Explain why a model element was added when this was not included in the target system.
- Q2.7 Describe the procedures and methods used to conceptualise the target system key elements as model elements. How did you make use of the evidence? What other sources did you utilise to conceptualise model elements?

1 MODEL AIM AND CONTEXT QUESTIONS

- Q1.1 If this RAT-RS use is related to a specific publication, please provide a reference to that publication.
- Q1.2 What is the purpose of the model? (prediction, explanation, description, theoretical exploration, illustration, analogy, social interaction, or other (please specify); for an explanation of model purposes, see Edmonds et al (2019) http://jasss.soc.surrey.ac.uk/22/3/6.html)
- Q1.3 What domain does the model research?
- Q1.4 What (research) question(s) is the model addressing? (in general / in this publication)
- Q1.5 What is the MAIN driver for your initial model development step? (theory(s), empirical evidence, existing model(s), participatory modelling data)
- Q1.6 Explain why this MAIN driver was chosen.
- Q1.7 What is the target system that this model reproduces? (briefly describe the target system and its boundaries)
- Q1.8 Explain why this target system and these boundaries were chosen.

2 CONCEPTUALISATION QUESTIONS

- Q2.1 What key mechanism(s) of the target system is/are used as driver(s) in this model? (examples: queuing system; physical/social/political network)
- Q2.2 Why is/are this/these key mechanism(s) used?
- Q2.3 What are the key elements of the target system?
- Q2.4 Which of these key elements were mapped into model elements? (distinguish (at least) between agents, environment, and relationships/interactions among any combination of these)
- Q2.5 Explain why key elements of the target system were included, excluded, or changed in the model.
- Q2.6 Explain why a model element was added when this was not included in the target system.
- Q2.7 Describe the procedures and methods used to conceptualise the target system key elements as model elements. How did you make use of the evidence? What other sources did you utilise to conceptualise model elements?

Question Type Analysis

- 44 RAT-RS questions classified into five types by lexical/syntactic cues
    - Factual: verifiable, single correct answer (1)
    - Descriptive: 'what/which', answer found in text (23)
    - Explanatory: causal reasoning, 'why/explain' (13)
    - Binary: yes/no, often with follow-up (2)
    - Evaluative: meta-level appraisal of the paper (5)

Prompt Design and Architecture

- Iterative design grounded in current prompt-engineering research

- Five modules
  - Role
  - Instructions
  - Output Format
  - Questions
  - Terminology

- **[INTERPRETIVE] questions:** Synthesise only the reasoning the authors have explicitly stated, in your own words, drawing on evidence from across the paper. If the authors provide no stated rationale, return [NOT AVAILABLE]. Do not construct plausible-sounding justifications from domain knowledge or general convention.

Step E. For every response, assign one of four evidential basis signals:
- **[DIRECTLY STATED]** — The information is directly and clearly stated in the paper.
- **[LOGICALLY INFERRED]** — The information is not stated explicitly but can be directly and logically derived from a specific passage in the paper (cite that passage). This signal must never be used to supply reasoning the authors did not articulate. If you cannot cite a specific passage, use [NOT AVAILABLE] instead.
- **[NOT AVAILABLE]** — The information is genuinely absent from the paper. State this clearly and briefly note why it may be absent (e.g. authors do not discuss validation; no stakeholder engagement reported).
- **[NOT APPLICABLE]** — This question does not apply to this paper (e.g. a conditional question whose premise is not met, or a practice not relevant to the study type). State the reason in one sentence.

Q2.1 What key mechanism(s) of the target system is/are used as driver(s) in this model? (examples: queuing system; physical/social/political network)
Q2.2 Why is/are this/these key mechanism(s) used?
Q2.3 What are the key elements of the target system?
Q2.4 Which of these key elements were mapped into model elements? (distinguish (at least) between agents, environment, and relationships/interactions among any combination of these)
Q2.5 Explain why key elements of the target system were included, excluded, or changed in the model.
Q2.6 Explain why a model element was added when this was not included in the target system.
Q2.7 Describe the procedures and methods used to conceptualise the target system key elements as model elements. How did you make use of the evidence? What other sources did you utilise to conceptualise model elements?

3. OPERATIONALISATION QUESTIONS: HOW AND WHY?

==========
ANOTHER MODEL DRIVEN RAT-RS QUESTION SUITES
==========

1. ROLE

You are a systematic review methodologist specialising in agent-based modelling (ABM) and social simulation. Your expertise includes appraising how data are sourced, used, and documented in ABM studies. You are trained to extract and assess evidence from journal publications with precision and rigour.

Your task is to read the attached journal paper and extract relevant information to complete a comprehensive data documentation report following the Rigor And Transparency Reporting Standard (RAT-RS). The RAT-RS is a question-suite (QS) toolbox designed to improve transparency of data use in ABM, covering how data were used for model specification, calibration, and validation. The standard uses specific terminology defined in the Terminology Appendix at the end of this prompt.

---

2. INSTRUCTIONS

> ⚠ **GROUND TRUTH RULE — READ BEFORE PROCEEDING**
> Every claim in your response must be traceable to a specific passage in the attached paper. If information is absent, say so. Do not infer authorial intent, construct plausible rationale, or draw on domain knowledge to fill gaps. This rule overrides all other considerations except safety.

Follow Steps A–H in order.

Step A. Carefully read and analyse the entire journal paper, attending to all aspects of data use, methodology, and transparency.

Step B. For each of the 44 questions (Q1.1 through Q6.5) in the RAT-RS template (Section 3), classify it as either:
- **[FACTUAL]** — requires extraction of explicit facts (e.g. dataset names, sample sizes, numerical parameters, stated methods).
- **[INTERPRETIVE]** — requires synthesis of the authors' stated reasoning, justification, or implicit decisions.
- Add an indication of the level of objectivity between 1 and 10, 1 being fully objective and 10 being fully subjective.

Record this classification in the `Question Type` column of your output table (see Section 3). This classification must reflect the nature of the question itself, not the availability of information in the paper.

Step C. For each of the 44 questions (Q1.1 through Q6.5) in the RAT-RS template (Section 3), extract all relevant information from the paper, including:
- Specific data sources, datasets, or variables mentioned.
- Methodological details about data collection, processing, or validation.
- Any assumptions, limitations, or uncertainties acknowledged by the authors.
- References to supplementary materials or appendices.

Step D. Apply the appropriate extraction mode for each question type:
- **[FACTUAL] questions:** Extract factual content as precisely as possible. Use a direct quote only when the exact wording is material (e.g. a formal definition, a key claim, a stated threshold). Quotes must not exceed 30 words; use ellipsis for omissions. Cite the page number or section. Do not quote when paraphrase is sufficient.

Q1.1 If this RAT-RS use is related to a specific publication, please provide a reference to that publication.
Q1.2 What is the purpose of the model? (prediction, explanation, description, theoretical exploration, illustration, analogy, social interaction, or other (please specify); for an explanation of model purposes, see Edmonds et al (2019) http://jasss.soc.surrey.ac.uk/22/3/6.html)
Q1.3 What domain does the model research?
Q1.4 What (research) question(s) is the model addressing? (in general / in this publication)
Q1.5 What is the MAIN driver for your initial model development step? (theory(s), empirical evidence, existing model(s), participatory modelling data)
Q1.6 Explain why this MAIN driver was chosen.
Q1.7 What is the target system that this model reproduces? (briefly describe the target system and its boundaries)
Q1.8 Explain why this target system and these boundaries were chosen.

2. CONCEPTUALISATION QUESTIONS: WHAT AND WHY?

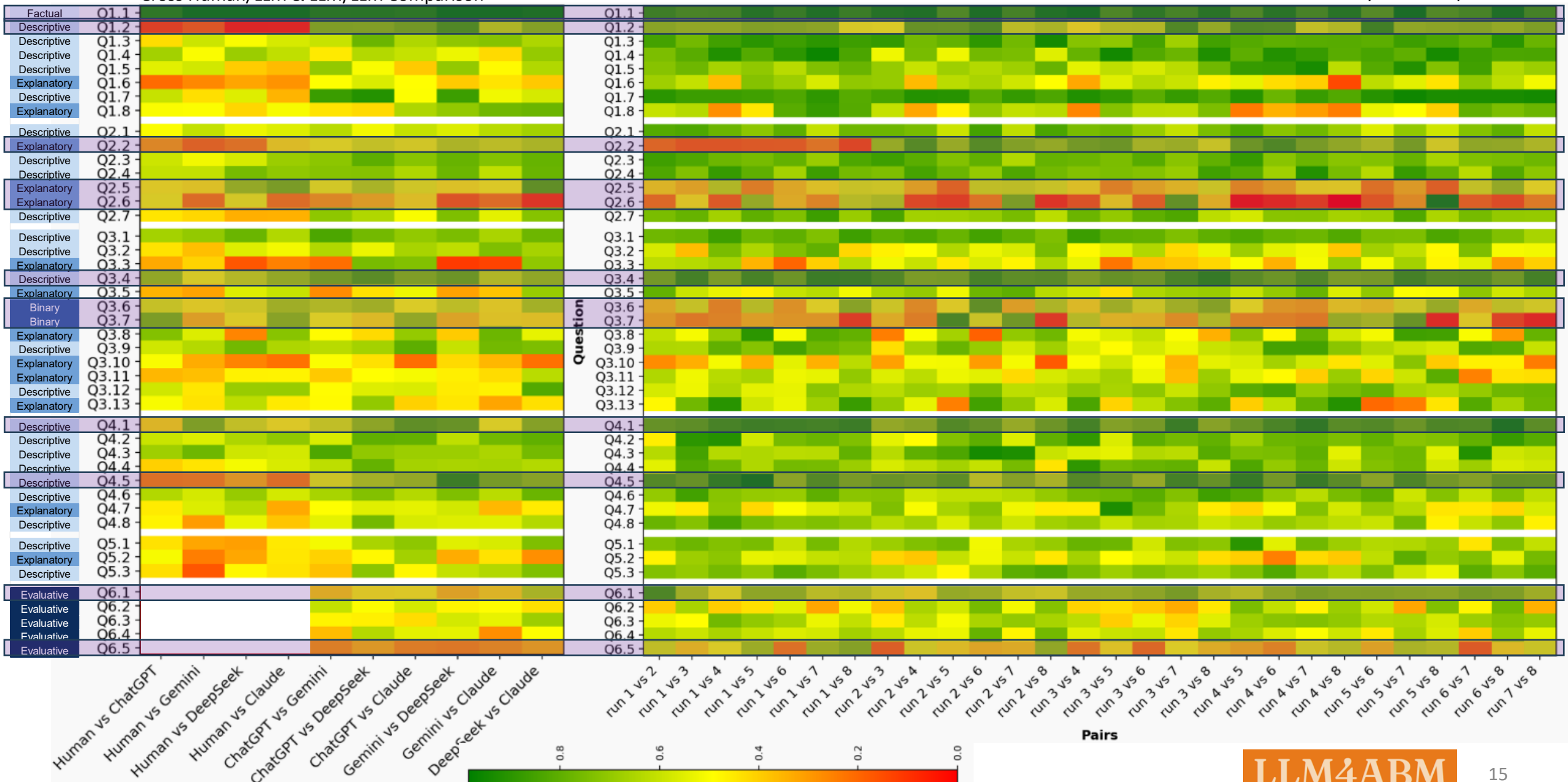
Question Suite (QS)	A structured set of questions in the RAT-RS focusing on a distinct aspect of data use (specification, conceptualisation, operationalisation, experimentation, evaluation).
Target system	Those aspects of the real-world system that are studied in order to gain knowledge about the phenomenon.
Theory	Collective term for theory, theories, or theoretical constructs; includes any theoretical construct, not only mature or formalised theories.

https://github.com/PeerOlafSiebers/ssc2026/tree/main/prompts

Response Similarity Comparison: Heatmap

Cross Human/LLM & LLM/LLM Comparison

Internal Consistency Across Repeated Runs



LLM4ABM

Response Similarity Comparison: Emerging Patterns

- General
  - Low response similarity may indicate that information is not clearly articulated or missing in the paper
  - Low LLM/LLM response similarity indicates that model-specific interpretation differs (with Gemini being the least aligned model (obvious for Q2.6 and Q3.3)
- Factual Questions
  - Q1.1: Factual information is easy to extract (if it is provided in the document)
- Descriptive Questions
  - Q3.4 and Q4.1: Response similarity of descriptive questions is at a decent level
  - Q1.2 and Q4.5: Questions might be misinterpreted either by human or LLM, if human/LLM similarity is low but LLM/LLM and intra-LLM similarity is high
- Explanatory Questions
  - Q2.5 and Q2.6: Explanatory questions generally have low similarity scores, even for repeated runs <<<why?>>>
  - Q2.2: Highlights the importance of doing repeated runs, as an LLM can generate completely different answers. Here, run 1 should be ignored.
- Binary Questions
  - Q3.6 and 3.7: Such decisions are usually not reported directly in the paper, and if there is a dependency between multiple binary questions it can only get worse
- Evaluative Questions
  - Q6.1 and Q6.5: These vary from medium to low response similarity between LLMs but also between repeated runs of the same LLM. This is expected as these questions are not related to extracted information ask about a judgement of the LLM about the paper as well as some suggestions for improvement

Response Similarity Comparison: Statistics

- Cross Human/LLM Comparison
  - Average similarity: 0.497 (SD = 0.174)
  - Consistency: Factual 0.969. Binary 0.562. Descriptive 0.515. Explanatory 0.420. Evaluative N/A
- Cross LLM/LLM Comparison
  - Average similarity: 0.586 (SD = 0.166)
  - Consistency: Factual 0.959. Descriptive 0.666. Binary 0.528. Explanatory 0.474. Evaluative 0.459
- Internal Consistency Across Repeated Runs
  - Average similarity: 0.664 (SD = 0.175)
  - Consistency: Factual 0.921: Descriptive 0.750: Evaluative 0.586: Explanatory 0.551. Binary 0.474

Descriptive Response Example

- Q3.2 Are these data elements implemented with the help of qualitative or quantitative data or further models?
  - Human: Mix of quantitative (service time, arrival rates, probabilities), qualitative (states + transitions, proactive strategy, archetypes), and model-based (queuing behaviour, network) data.
  - ChatGPT: They used both quantitative and qualitative evidence. Quantitative data fed distributions and parameters, while qualitative observation and interviews supported state charts and rule design.
  - Gemini: Data elements are implemented using both quantitative data (rotas, company reports) and qualitative insights from observations and interviews (p.5, §3; p.6, §3).
  - DeepSeek: Quantitative data: frequency distributions (triangular), probability distributions, company transaction numbers. Qualitative data: customer type definitions from observations. No further models used.
  - Claude: Both. Quantitative: company transaction data, rota data, numerically-coded observation timings, survey results. Qualitative: semi-structured interviews and participant observation informing behavioural rules and agent state designs. No further sub-models are used as inputs. (p.5, §6; p.10).

Descriptive Response Example

- Q3.2 Are these data elements implemented with the help of qualitative or quantitative data or further models?
  - Human: Mix of quantitative (service time, arrival rates, probabilities), qualitative (states + transitions, proactive strategy, archetypes), and model-based (queuing behaviour, network) data.
  - ChatGPT: They used both quantitative and qualitative evidence. Quantitative data fed distributions and parameters, while qualitative observation and interviews supported state charts and rule design.
  - Gemini: Data elements are implemented using both quantitative data (rotas, company reports) and qualitative insights from observations and interviews (p.5, §3; p.6, §3).
  - DeepSeek: Quantitative data: frequency distributions (triangular), probability distributions, company transaction numbers. Qualitative data: customer type definitions from observations. No further models used.
  - Claude: Both. Quantitative: company transaction data, rota data, numerically-coded observation timings, survey results. Qualitative: semi-structured interviews and participant observation informing behavioural rules and agent state designs. No further sub-models are used as inputs. (p.5, §6; p.10).

Conditions, Caveats, and Limitations

- Three conditions for reliable use of LLM-assisted RAT-RS completion
  - Humans must stay involved
  - LLM use must be disclosed
  - Extraction prompt must be versioned and shared
- Review requirements
  - Factual and descriptive questions need minimal review
  - Explanatory, binary, and evaluative questions need thorough review
- Study limitations
  - Single-paper study; no claim of generalisability
  - Cosine similarity captures topical overlap only, not correctness

Tragedy of Commons: Answers to the Key Questions

- Q1: How can we reduce the effort of using a standard, to boost its adoption?
  - Use of LLMs as Supervised Extraction Assistants (swap your role from producer to reviewer)
- Q2: What immediate individual benefit could adopting a standard offer?
  - Identifying missing or vague information in the paper
  - Gaining confidence about the paper's rigour and transparency

Together, lower individual cost (Q1) and an individual benefit realised at the point of use rather than afterwards (Q2), is what turns standard adoption from a collective-action problem into an individually rational choice

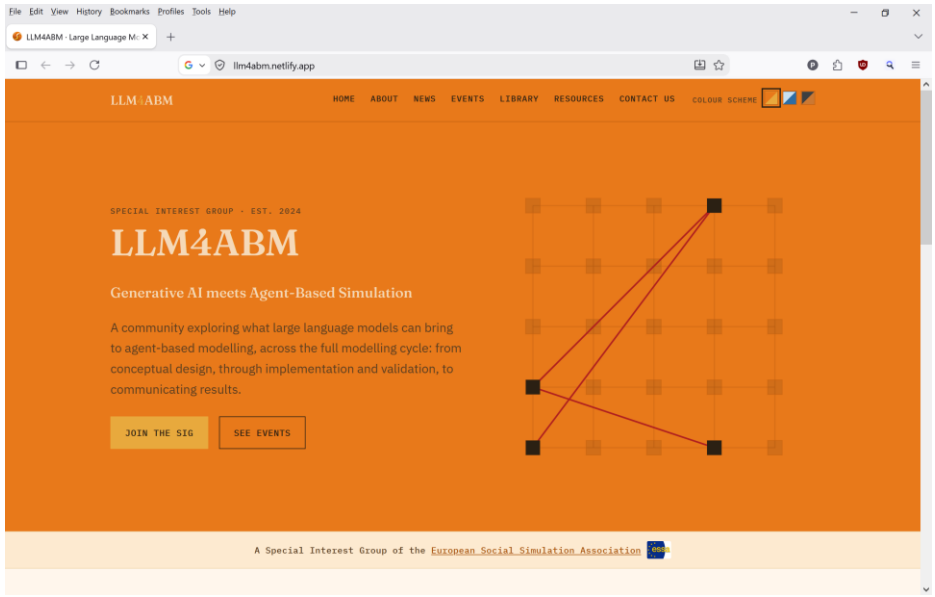
Summary of Benefits from using an LLM-driven Supervised Extraction Assistant

- Fast draft generation
  - A draft report is returned in minutes, cutting the overall workload drastically
- Validation opportunities for users and standard developers alike
  - For users: Identifies missing or vague information in the paper
  - For developers: Identifies shortcomings that could lead to instructions being misinterpreted
- Advice for improvement of reporting
  - QS6 assesses the quality of information provision in the paper and suggests how to improve it
- Training for using standards
  - Generated reports also double as training material

Call to Action

- Next steps for the community
  - Adopt the prompt templates for supervised LLM-assisted extraction on your own papers
  - Contribute validation cases to a shared repository
  - Update standards like ODD and RAT-RS to recognise LLM-assisted extraction
  - Define acknowledgement 'reporting standards' for AI-assisted reports
  - Rethink question design to improve consistency for humans and LLMs
- Next steps for us
  - Develop a RAG system to improve information gathering
  - Test LLM-assisted extraction with other standards and protocols
  - Automating processes and analysis (including information fusion)
  - Drive and contribute to the community efforts mentioned above

Questions?



GitHub

arXiv

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

peer-olaf.siebers@nottingham.ac.uk

LLM4ABM

References & Resources

- References
  - Achter, S., Borit, M., Chattoe-Brown, E., & Siebers, P. O. (2022). RAT-RS: a reporting standard for improving the documentation of data use in agent-based modelling. International journal of social research methodology, 25(4), 517-540.
  - Grimm, V., Berger, U., Bastiansen, F., Eliassen, S., Ginot, V., Giske, J., ... & DeAngelis, D. L. (2006). A standard protocol for describing individual-based and agent-based models. Ecological modelling, 198(1-2), 115-126.
  - Müller, B., Bohn, F., Dreßler, G., Groeneveld, J., Klassert, C., Martin, R., ... & Schwarz, N. (2013). Describing human decisions in agent-based models–ODD+ D, an extension of the ODD protocol. Environmental Modelling & Software, 48, 37-48.

- Resources
  - LLMs as Supervised Extraction Assistants - Paper Preprint: https://arxiv.org/abs/2606.13749
  - LLMs as Supervised Extraction Assistants - GitHub Repository: https://github.com/PeerOlafSiebers/ssc2026

QS1: MODEL AIM AND CONTEXT QUESTIONS

Q1.1 If this RAT-RS use is related to a specific publication, please provide a reference to that publication.

Q1.2 What is the purpose of the model? {prediction, explanation, description, theoretical exploration, illustration, analogy, social interaction, or other (please specify); for an explanation of model purposes, see Edmonds et al (2019) http://jasss.soc.surrey.ac.uk/22/3/6.html}

Q1.3 What domain does the model research?

Q1.4 What (research) question(s) is the model addressing? {in general / in this publication}

Q1.5 What is the MAIN driver for your initial model development step? {theory(s), empirical evidence, existing model(s), participatory modelling data}

Q1.6 Explain why this MAIN driver was chosen.

Q1.7 What is the target system that this model reproduces? {briefly describe the target system and its boundaries}

Q1.8 Explain why this target system and these boundaries were chosen.

QS2: CONCEPTUALISATION QUESTIONS: WHAT AND WHY?

Q2.1 What key mechanism(s) of the target system is/are used as driver(s) in this model? (examples: queuing system; physical/social/political network)

Q2.2 Why is/are this/these key mechanism(s) used?

Q2.3 What are the key elements of the target system?

Q2.4 Which of these key elements were mapped into model elements? {distinguish (at least) between agents, environment, and relationships/interactions among any combination of these}

Q2.5 Explain why key elements of the target system were included, excluded, or changed in the model.

Q2.6 Explain why a model element was added when this was not included in the target system.

Q2.7 Describe the procedures and methods used to conceptualise the target system key elements as model elements. How did you make use of the evidence? What other sources did you utilise to conceptualise model elements?

QS3: OPERATIONALISATION QUESTIONS: HOW AND WHY? 1/2

Q3.1 What data element(s) did you include for implementing each key model element in the model's scope?

Q3.2 Are these data elements implemented with the help of qualitative or quantitative data or further models?

Q3.3 Explain how data affected the way you implemented each model element and why. {i.e. explain your choice of data elements}

Q3.4 What are the data elements used for in the modelling process: specification, calibration, validation, other?

Q3.5 Why for this use and not another one?

Q3.6 Did required data exist?

Q3.7 [Conditional on Q3.6 = yes] If it existed, did you use it?

Q3.8 [Conditional on Q3.7 = no] If you did not use it, why not?

QS3: OPERATIONALISATION QUESTIONS: HOW AND WHY? 2/2

Q3.9 For the existing data you used, provide details about data sources, sampling strategy, sample size, and collection period. For any data you collected, provide details about how it was collected, sampling strategy, sample size, and collection period.

Q3.10 Justify your data-gathering decisions from Q3.9.

Q3.11 If you needed to analyse the data before including them in the model (regardless of whether you collected data yourself or used existing data), what data analysis did you do and why did you choose this specific analysis?

Q3.12 In what format was the data implemented? {e.g. look-up table; distribution}

Q3.13 Why this way?

QS4: EXPERIMENTATION QUESTIONS

Q4.1 Describe the calibration process you followed, stating which parameters you calibrated, their ranges, your reasons, and the similarity you achieved.

Q4.2 Describe the experimental design process you followed, stating your reasons and the methods you used for the different steps. {e.g. calculating warm-up period, run length, and number of replications; sensitivity analysis; robustness analysis}

Q4.3 What type(s) of experiments did you run? {e.g. calibration; empirical validation; sensitivity analysis; performance optimisation}

Q4.4 For each experiment, name the purpose (objective).

Q4.5 Describe the parameters you used to set up the experiments.

Q4.6 Describe the data output that the model was designed to produce, your reasons for producing this output, and the data type of the output (qualitative or quantitative).

Q4.7 Describe the (statistical) analysis that you used on the output data and why.

Q4.8 Did you discuss the output with the stakeholders? What did you discuss? Why? What effect did it have on the model?

QS5: EVALUATION QUESTIONS

Q5.1 In validation, what similarity measures did you use and why? What similarity did you get? What would you consider a good similarity and why?

Q5.2 How do the data outputs support an answer to the research question?

Q5.3 Did you discuss the validation results with the participants? What did you discuss? Why? What effect did it have on the conclusions?

6. REPORTING QUALITY ASSESSMENT (Supplementary QS)

Q6.1 Overall, how transparent and complete is the data reporting in this paper? (score 1–5, where 1 = very poor, 5 = exemplary). Justify your score.

Q6.2 What are the main strengths of the data reporting in this paper?

Q6.3 What are the main weaknesses or gaps in the data reporting?

Q6.4 What specific recommendations would you make to the authors to improve data use, sourcing, and documentation?

Q6.5 Are there aspects of data use that appear inconsistent or contradictory within the paper? If so, describe them.