Review: Exploring the opportunities and challenges of using large language models to represent institutional agency in land system modelling

In this manuscript, the authors describe their work inducing a large language model to "role-play" as various kinds of policy decisionmakers in an agent-based land use model. While a human operator needs to stay in the loop to keep the LLM on task and producing output in the correct format, the agents—when properly prompted—are capable of producing policy actions that achieve their goal. As befits such a novel method, the authors do more than just using the policy actions output by the model; they also dig in to the apparent "reasoning" behind its actions.

This is a fascinating piece of research. The paper is composed logically, well-written, and the figures are clear. However, I do have a number of comments, the most important of which relate to the manuscript's eliding of how LLMs actually work. Once these are addressed, though, it will stand as an important, foundational contribution to the use of LLMs in agent-based land use modeling.

Major comments

The biggest comment I have is that the paper needs to be more precise in how it discusses the LLM agents. I appreciate that anthropomorphizing them to some extent can be helpful, but the paper needs a better grounding in what is actually happening in the LLM. It's not until the Discussion that the actual mode of operation is actually discussed—these models generate a series of words based on a corpus of training data as well as the "conversation" that has occurred between the operator and model. This is an important enough concept that it should be described when LLMs are first mentioned in the Introduction and used to frame the results and discussion.

As a result, these models are not capable of "reasoning," a term which appears throughout the paper; they simply generate "the most likely next word." (See Bender et al., 2021, "On the dangers of stochastic parrots.") However, the paper frequently seems to buy in to the illusion. For example, at lines 314-316:

> While LLM models are often perceived as opaque, LLM-powered agents can offer the compelling ability to articulate human-comprehensible reasoning for their actions, providing a window into the decision-making processes that drive their behaviour.

It certainly is compelling and articulate, but the LLM is neither reasoning nor making decisions. I'm not just trying to be pedantic: This over-anthropomorphization of the model seems in some places to bleed into the authors' interpretation of their results. For example, at lines 280-281: "S1.2 seems to provide reasoning in hindsight to justify a decision made in the absence of such reasoning." It's not "reasoning" or "justifying" anything. It's simply producing the most plausible string of text given that it's already said "+2 tax increase." This

is why the results can differ so much based on the order of "decision" vs. "explanation" (Agents S1.1 vs. S1.2) and should not have been surprising.
The authors also must add discussion of the potential biases that are possible with this kind of setup. One is that it'd be possible for the human in the loop to introduce their own bias when interacting with the LLM; I hardly consider this a dealbreaker, but it should be at least mentioned. More importantly, the outputs of an LLM will reflect any biases in its training dataset. It might thus be difficult to spur an LLM to enact policies that defy political-economic orthodoxy. It's going to be biased toward "conventional wisdom," making me skeptical of statements like "[the use of LLMs] provides an opportunity to search for novel insights into human behaviour" (lines 71-72) and "modellers can get useful inspiration from this communication" (line 498).

Minor comments

- Fig. 1:
    - Does not seem to be referenced at all in text.
    - "Use LLM to assist with refinement" does not seem to be mentioned anywhere in the text.
- Lines 141-143: I thought the agent was supposed to represent an institution, but here it says there's an "institutional environment within which the agent operates." Are those *other* institutions? Are there non-institutional agents as well?
- Lines 209-210: Add text explaining that the policy actions represent *change in* tax level.
- Table 1:
    - What "experience" is being referred to by "experiential learning"? Is the agent looking at what its policy was for previous years and considering what changes to make?
    - Second sentence of Description box for Agent Q is unclear. I don't know what sequencing the roles" or "conversational endpoint setting" are.
- Lines 217-220: For reproducibility, please give more details of the genetic algorithm setup in an Appendix or Supplement.
- Fig. 4:
    - Please avoid using red and green on the same figure, as these are hard to distinguish for people with the most common form of color blindness.
    - What are the Y-axis units?
    - What is "demand force"?
- Lines 242-246: Is the average error in Fig. 5a (Agent B1) just a re-presentation of the data from Fig. 4? If so, please mention that in the text here. If not, please explain.
- Sect. 3.2.1 (performance of Agent S1.1:
    - You're right that the policy actions are "generally understandable," but one weird thing is the drop to baseline taxation levels around years 35-50. Any idea why that happened? … I see now that this is discussed in Sect. 3.3, Action III. Please add a reference to that in Sect. 3.2.1.

    - Why are non-negative changes in taxation “plausible”?
    - Line 257: “Significant”—was there a statistical test? If not, please rephrase for clarity. If so, please explain.
- Fig. 5: Having errors be negative when there’s *too much* meat production feels counterintuitive. This is obviously just a personal preference, but in any case, the chosen convention should be mentioned in the figure caption.
- Lines 348-349: The way this is phrased makes it sound like the agent was given two goals: maintaining supply levels and matching supply to demand. In reality, it seems like it just made up the latter. Right?
- Lines 450-451: What “conventional methods” do you mean? Hard-coded agent behavior can’t produce (the appearance of) reasoning, but in that case it doesn’t need to—you already know the rules governing agent behavior.
- Lines 460-461: “As the policy objective nears realisation, Agent S1.1 judiciously reduces tax levels to mitigate potential over-adjustment.” Does it? I thought policy actions represented *changes* in tax levels. Because those aren’t negative after the first few years, this means that taxes never go down.
- Lines 490-492: Please explain why conventional modeling techniques can’t represent these interactions.

Technical corrections / typos

- Line 199: “plausibility” is probably not the right word. “Desire for”?
- Line 327: Word missing (of?) in “investigation the large”.
- Line 444: “conversions” should be “conversations”.
- Line 472: Missing period.