Reply on RC2

(Note: The symbol “➢” denotes the beginning of each response.)

Major comments

Reproducibility

- Please make all code and data used to produce these results available in a public repository with accompanying doi. In my view it is not enough to have code available“on request”. This repository should have sufficient meta data and installation instructions to make your results fully reproducible on a range of operating systems.

➢ We appreciate your interest in the source code and data of our work. We will upload the code to a public repository with a DOI, ensuring it is accessible. We will also provide installation instructions. The code is implemented in Java and Python, both of which are cross-platform languages with APIs that should work across different operating systems. The repository will be open to public contributions, allowing others to adapt and extend the code, including testing and improving its compatibility with various systems. The author team will be glad to help with this.

- Lines 173-175; great that there is a CRAFTY emulator. However, I don’t think this has been published. As such, to use it here I need you to provide details about its design andevaluation of how well it reproduces CRAFTY outputs as an SI. Please make all code andsupporting data required to run it publicly available in a repository.

➢ Thank you for your comment. We are pleased to inform you that the CRAFTY emulator is publicly available on both GitHub and Zenodo, along with our other papers online. We will also upload a version of the source code that is coupled with the LLM agents for this paper to Zenodo with documentation.

The emulator has been designed to replicate the functions of the original CRAFTY model by the same team, and its behaviour has been evaluated to ensure alignment with the core dynamics of CRAFTY. As per your suggestion, we will add a description of the emulator's design and a comparison of its outputs with the original CRAFTY model in the SI of the revised manuscript.

- Overall, I think a supplementary information expanding lines 179-191 to be really clear on what you did would be very helpful and make the work more reproducible. E.g. whichRCPs / SSPs did you use? I *think* from Figure 5 that you are showing means across the scenarios, but this isn’ttotally clear. It would also be very interesting to see how much variance there is by scenario.

➢ Thank you for your suggestions. In this study, we used a single scenario, SSP1, to test the agents. The purpose of this paper is not scenario-specific; rather, it is intended as a proof of concept to explore the applications of LLM agents in land system modelling. By using a simplified setting, our aim is to observe the internal logic consistency and the contextually relevant behaviours of different agents rather than to test their performance across multiple RCP-SSP scenarios.

We acknowledge that testing the method across a range of scenarios could provide valuable insights, particularly if future research shifts toward a more empirical focus. In the revised manuscript, we will explicitly mention the scenario used in the experimental settings and clarify that the results are based on this single scenario.

We will expand the description in the Supplementary Information section to detail the coupled model works. We will also revise Figure 5 and its caption to make it clear that the results are specific to one scenario.

- Lines 126-152: do you have a complete set of prompts that you tried? Presumably this ispossible if the dialogue is held in the agent memory (Line 139). I think this would be fascinating to see (as an SI) and would help with reproducibility.

➢ Thank you for your valuable suggestion. The dialogues were stored temporarily as a list of Python dictionaries in the program during runtime. We did not record how the prompts evolved. However, we can provide the very beginning prompt (a draft actually) in the SI. While this does not capture the full evolution of the prompts, it might effectively illustrate the differences between the initial and final versions, which will be useful for understanding what changes have been made.

- If not, then at lines 131-132 I think it would be good to more clearly state your criteria for assessing the LLM output. Perhaps in a table. IE were these very strict criteria (basic functionality) the only ones you used? Or did you also, subjectively, select for those thatseemed to make sense? How many people evaluated the LLM outputs during the human in the loop stage to check you were applying any criteria consistently? I think a degree of subjectivity is inevitable, but worth being transparent about.

➢ The primary focus of our prompt refinement was on the functional aspects of the agents, such as ensuring that their outputs were correctly formatted to avoid disrupting the existing programmed model, aligned with the simulation's structural requirements, and operationally smooth within the coupled system. This process was more about fixing problems in existing prompts rather than creating entirely new strategies for the agents.

For example, in one scenario, the agent repeatedly tried to adjust the meat supply based on meat demand rather than aligning with the policy goal. To address this, we modified the prompts in different ways, such as adding notes to the prompts to emphasize the real target, attempting to clarify that meat demand was only reference information, not the target to pursue. However, this misunderstanding persisted. We eventually excluded meat demand information from the input entirely for some agents, such as Agent S1.1 and Agent S1.2.

You may notice examples of the required format for agents' outputs and notes such as, "Don’t fake interaction with the policymaker if there is no interaction yet." All such information was derived from the human-in-the-loop (HIL) process. Each note reflects a functional issue that was identified and addressed during the HIL process.

Given this functional focus, only one person was involved in the prompt refinement process. Having more people engaged would have been desirable to enhance consistency and reduce potential subjectivity, given that the process was time-intensive. However, the problematic outcomes were relatively straightforward to identify (e.g., ill-formatted outputs or repeated misunderstandings of context). A single person was able to carry out the necessary refinements effectively.

We will include this clarification in the revised manuscript and emphasize that while our approach was practical and focused on functionality, future research could benefit from broader collaboration in refining and evaluating prompts, particularly as more complex tasks are introduced.

Scaling

You note scaling issues in your discussion; however I think some further information and detailswould be helpful for the reader to gauge the extent of these challenges.

- It would be helpful to get a sense of run-time of one human-in-the-loop prompt session per institutional agent type (setup, LLM thinking time etc). E.g. Suppose we wanted to dosets of runs with differing levels of policy targets to assess consistency of answers. How feasible is this?

➢ Yes, this is highly feasible, particularly as more reliable and faster APIs are being developed.

For example, GPT-4o, released in May this year, is significantly faster than GPT-4 used in this paper. Additionally, open-source LLMs accessed via platforms like Groq API offer specialized infrastructure that can speed up LLM inference, further enhancing feasibility.

During our experiments, the runtime for LLM agents' "thinking" was slower, but the major challenge was the occasional failure of API responses, which required re-sending requests. This made the runtime less predictable. Another time-consuming aspect was the iterative process of trial and error in refining prompts. While this approach allowed us to address specific issues, it did not guarantee consistent progress with every iteration.

It is worth mentioning that recent developments in frameworks, such as LangGraph's integration with Pydantic, help automate data validation and improve output structuring. While these tools are not yet perfect or fully mature, their ongoing development greatly facilitates the creation of AI agents and makes outputs adhere to the required format. These technologies could significantly reduce the time and effort required for prompt refinement and data format validation in future research.

- Similarly, perhaps I missed this, but if you repeat one of your policy scenarios multiple times, how much do the resulting outputs diverge? Is this computationally prohibitive?

➢ The results generated by LLMs can be reproducible under specific conditions. Reproducibility depends on factors such as setting a fixed random seed, if the system or API allows, using the exact same model version and configuration (e.g., temperature, max tokens), ensuring that input text remains identical.

With the same settings and running the API on the same device, the model can reproduce the same results in principle. This is what we have observed during our experiments. However, exact reproducibility might not always be feasible because users cannot control when or how LLM providers might update the underlying model. These updates can result in subtle differences in output even when the same inputs and parameters are used.

This is an important caveat for reproducibility, which we will highlight in the revised manuscript. We also suggest that our focus on evaluating the internal logical consistency and contextual relevance of the LLM agents' outputs is expected to be relatively robust to these changes.

- Further, if you have stochasticity in the underlying model, how much can this lead to unpredictable policy pathways? You provided historical data to the LLM – I take it this was observations rather than historical CRAFTY runs? Did you then spin-up the CRAFTY AFT distribution to match the observations? Otherwise might our initial LLM choices be sensitive to the initial conditions?

➢ A good point; stochasticity exists in both the LLMs and the CRAFTY model, although there is no spin-up and the model starts from the same set of conditions each time. As above, we will discuss the existence and influence of stochasticity in the revision. It's worth noting though that the data received by the LLM agents are updated periodically during the simulation as CRAFTY runs. This ensures the agents are informed by dynamic, real-time simulation outputs rather than relying solely on static, pre-observed (historical) data. We will revise this phrasing in the paper to eliminate any potential confusion.

While both models are able to produce understandable behavioural patterns, which is why they are useful and can be coupled meaningfully, stochasticity can be very important. In the case of LLMs, a straightforward way to increase unpredictability is by adjusting the temperature parameter. Higher temperatures make outputs more diverse and therefore more unpredictable across runs. However, in our experiments, the temperature was set to 0 by default, which is intended to ensure consistent outputs across runs.

Even if some unpredictability exists, it might not impact the goals of this research. Our primary focus is on the believability, contextual awareness, and logical consistency of the LLM outputs, as well as their potential to mimic human decision-makers with understandable behaviours. If unpredictability enhances the diversity of the LLMs without compromising the quality of their outputs, they remain valuable and worthy of further study.

Generalisability

You observe that stakeholder disagreement & subsequent contested policy spaces lead generally to slower decision-making. This is an important and fundamental insight, with somegrounding in the literature. Some questions and comments below on how universal / generalisable such a finding may / may not be.

– Lines 308-312

The setup of agent Q overall seems good and appropriate. I think it is worth being careful to remind readers that you are explicitly mimicking policymaking processes in a European context with broadly democratic norms and systems. The text here seems to discuss multi- stakeholder policymaking in abstract terms, but the setup of the multi-stakeholder network would presumably have to vary substantially in other policymaking systems. For example, in more authoritarian government systems, we may have "industrialisation from above" with very rapid changes, or one group of stakeholders' rights and views being cut out of decision-making.

➢ Thank you for your insightful comment. We agree with your observation and will clarify in the text that the setup of Agent Q reflects a political system modelled on broadly European Union (EU)-like democratic norms and systems, rather than an authoritarian framework. We acknowledge that the dynamics of multi-stakeholder policymaking would differ significantly in other political contexts, such as in authoritarian systems where rapid "industrialization from above" or exclusion of certain stakeholder groups may dominate the decision-making process. We will ensure this distinction is explicitly stated in the paper to avoid potential misunderstanding.

– Lines 407-414

Here and elsewhere you state that slow and or incremental policy changes are more realistic /more in line with expectations than the optimisation algorithm. A few more references to support this would be good, particularly to clarify whether this is primarily a feature of western democratic systems or a more general phenomenon.

That said, let us assume that incrementalism is a broadly realistic simulated policymaking approach. I wonder if, in a subsequent paper, one could demonstrate this empirically? E.g. could we take countries' stated climate targets, and review concrete progress / policy implementation towards them vs what an economically-optimal trajectory towards achieving them might look like. If it could be clearly demonstrated that these simulated policy responsesare closer to real-world choices than optimisation-based modelling that would be a tremendously important finding, I think. Not only to evaluate your model, but also for wider consideration of institutional constraints on rates of environmental land use change.

➢ We agree that incrementalism, as a feature of policymaking, could benefit from stronger contextual support in the paper. Incremental policy change has been well-documented as a characteristic of western democratic systems, particularly in literature on policy sciences. We will incorporate additional references to highlight these points.

Your suggestion to empirically validate incremental policymaking approaches by comparing real-world policy trajectories to simulated and optimization-based trajectories is highly valuable. We see significant potential for future research in this direction. It can be envisioned that computational expense might be a critical challenge in applying optimization algorithms to policymaking and land-use modelling. When institutional models or land-use models are highly complex or involve large parameter spaces, optimization can become infeasible due to excessive computational requirements. Moreover, if the optimization process is not sufficiently robust, it may underperform heuristic or rule-based decision-making approaches in both accuracy and speed, given the time required for computation.

In this paper, the optimization algorithm operates within a limited action space, which enables it to work effectively. However, this is a simplified scenario designed specifically for this research's conceptual focus. Expanding the optimization to a broader or more realistic problem space would introduce significant complexity. Reaching a balance between problem complexity and computational feasibility is critical, and defining the optimization problem well is a prerequisite to ensure it is neither too complicated to yield effective solutions nor too simplistic to preserve the essence of empirical policymaking processes.

Minor comments

– Line 67: I think it would be worth flagging here that the majority of this training text is in English and noting how this may culturally skew the "thinking" of the LLM. This sets up the issues around generalisability.

➢ Thank you for pointing this out. It is true that the richness and diversity of training text in different languages significantly affect the performance of large language models (LLMs). We will note in the paper that English is one of the richest and most extensively represented resources in LLM training datasets. This can lead to a cultural and linguistic skew in the "thinking" of the LLM, with potentially different outputs and text generation patterns. This highlights an important limitation and sets up the issue of generalisability, which we will explicitly discuss in the revised manuscript.

– Lines 67-74; I'm not going to argue that we need to move beyond the paradigm of economic rationality. That said, some citations here would be good -> can we be explicit about why economic rationality may not produce nuanced representations of human decision making?

➢ We agree that it is important to explicitly address why economic rationality may fall short in representing the nuances of human decision-making. While economic rationality assumes that individuals always act in ways that maximize utility based on stable preferences and perfect information, human decision-making is often influenced by factors such as bounded rationality, cognitive biases, emotional responses, and social or cultural norms. These factors lead to behaviours that may deviate from purely economically rational decisions.

We will incorporate citations to support this perspective. Foundational works such as Simon's *bounded rationality* (1997) and Kahneman and Tversky's *prospect theory* (1979) will be included to contextualize this limitation of economic rationality. This will help clarify why alternative approaches, such as those explored in our model, are valuable for capturing the complexity and nuance of human decision-making.

– Lines 87-95; this is good, but I want several citations here to point users not familiar with the LLM literature in the right direction. If I wanted to replicate your method, I would want first to consult the underlying methodological literature on this topic.

➢ Sure, we will add relevant literature in the text for the reader to reference.

– Similarly, lines 95-119; it isn't clear to me as a non-expert how far this is methodological innovation or a very common approach to prompt development. I would like much more reference to the underlying literature. Please also add these to Figure 1 as 1,2,3, etc with refs in a key.

- ➢ The prompt development framework proposed in this paper provides a systematic process for developing prompts specifically for LLM agents that are coupled with existing programmed systems. While some of the techniques may align with common practices in prompt engineering, our contribution lies in presenting a streamlined and structured process tailored to this particular task. This framework aims to assist future researchers in accelerating their prompt development by providing clear steps.

  At the time this paper was composed, there were numerous papers and technical documents discussing prompting techniques in general. However, resources specifically addressing the development of prompts for LLMs integrated with external programmed systems were very limited, if not nonexistent. To our knowledge, this framework represents a novel contribution. It is not entirely new in every individual element, but rather it synthesizes key steps derived from the engineering practices developed in this research.

  This framework is more of a practical engineering summary than a theoretical model supported by established evidence. Given the rapid developments in this field, we will review recent literature to identify relevant works that can be incorporated as references for the paper.

- – Figure 2: This is good, but please can it be formatted so that I understand better where it begins and ends & which steps follow which? Either by numbering stages or having it more clearly as atop-to-bottom process flow? Figure 3 is much better in this regard.

- ➢ Of course, we will improve Figure 2 to make the processes clearer.

- – Figure 5: Please point the reader to Table 1 for definition of your agent types in the caption.

- ➢ Very constructive suggestion. Thank you. We will mention Table 1 in the figure caption to better guide readers.

References

Simon, Herbert Alexander. *Models of bounded rationality: Empirically grounded economic reason*. Vol. 3. MIT press, 1997.

Kai-Ineman, D. A. N. I. E. L., and Amos Tversky. "Prospect theory: An analysis of decision under risk." *Econometrica* 47, no. 2 (1979): 363-391.