Reply on RC2

- **Black text** indicates reviewer comments.
- **Blue text** indicates author responses.

In the manuscript (submitted to GMD) "InsNet-CRAFTY v1.0: Integrating institutional network dynamics powered by large language models with land use change simulation", Yongchao Zeng, together with his collaborators, has developed a very interesting and powerful technique to use multiple institutional agents each with its own large language model (LLM) prompt history, together with a land-use change model (the latter based upon the CRAFTY model). For the entire European region, they can simulate the inter-institutional dynamics, with unstructured text (i.e., bullet-point recommendations) and numerical output being passed from one institutional agent to another, driving both the changing meat production and the changing percent of land that is a protected area. The agents that are defined are as diverse as a lawyer agent that is familiar with European law, to lobbyist agents that take the side either of agriculture or of environmental advocacy, and further to a high-level institution agent that has long-range goals in mind and that integrates the advice of other agents and prompts the other agents to try to achieve its goals. I am particularly impressed with this paper, never having imagined LLM chatbots that talk to each other, and furthermore never having imagined that these LLM chatbots can be defined with the prompt engineering to groom them as specialist institutional chatbots that can drive a land-use simulator. The writing (grammar, structure, etc.) is of very high quality. I only ask for minor revisions, which I enumerate below.

Dear Reviewer,

We are delighted that you find our work interesting and valuable. Your recognition of the novelty of our work is greatly appreciated.

We also appreciate your constructive comments. We have addressed each of them in the responses below. We believe these refinements will further enhance the clarity of our study.

Sincerely,

Yongchao Zeng (On behalf of all authors)

Lines 65-66: Holzhauer et al. (2019): This reference is missing in the list of references.

Thank you for pointing out this. We have added the reference in the revised manuscript.

Line 251: Why not SSP3 or SSP5 for the changing climate? SSP1 has little change climate-wise from the current time.

The SSP1 scenario assumes a relatively modest climate change alongside gradually improving socio-economic conditions, including steady economic growth. Under these conditions, the

CRAFTY land-use model produces a gradual increase in ecosystem service supply, providing a straightforward baseline for us to investigate the impact of LLM agents.

We previously used this scenario in our publications to evaluate different approaches to autonomous policymaking agents. While this choice ensures consistency, we acknowledge that exploring alternative scenarios, such as SSP3 and SSP5, would offer rich insights into institutional and land-use dynamics.

We will address this point in the discussion section.

Line 286: how long is an iteration in days or months or years?

Here one iteration indicates a year. We will mention this explicitly in the revised manuscript.

Line 296: What are the differences between the definitions and between performance of Llama-3-70b-8192 and gpt-4o, listed below in Table 1?

Both LLMs used in our experiments were among the most advanced models available for handling textual data at the time. A key distinction is that Llama-3-70b-8192 (developed by Meta) is open-source, while GPT-4o (developed by OpenAI) is closed-source. Both of the models were accessed via APIs that require internet connection. We chose to use both models instead of one due to rate limits on token usage and API calls per minute, which could disrupt the simulation. By distributing the API requests across both models, we mitigated the risk of errors caused by exceeding rate limits. Additionally, this approach helped reduce token costs, as Llama was accessed via the Groq platform, which offers free usage within a certain range.

We will expand on this point in the revised manuscript for further clarity.

Line 301: Table 1: Maybe “Wiring” needs to be defined?

Thanks for noting this. This is a typo. We have corrected it to “Writing” in the revised manuscript.

Line 306: Is this amount of output for the whole period of 2016-2076? Or is it per iteration? I'm a bit surprised that the amount of output is so small. If you're simulating land use over all of Europe with a 5-arcminute spatial resolution, I would expect a lot more output, especially if different countries have different policies.

Thank you for your question. The CRAFTY land-use model is driven by numerous rule-based agents representing different types of land users and produces a large amount of data per iteration. However, the output mentioned in Line 306 refers only to the text generated by LLM agents between 2016 and 2076. Since the LLM agents were activated every ten iterations, text generation occurred only six times during the entire simulation period, with each activation producing an average of approximately 3,300 words.

In this study, we did not model country-specific policies for each EU nation. Our primary objective was to assess the feasibility and proof-of-concept of integrating LLM agents with the rule-based land-use model, focusing on their contextual awareness and logical coherence, which requires

manual examination of textual outputs. Incorporating individual country policies would significantly increase data volume, making manual analysis impractical.

We acknowledge the importance of country-specific institutional agents and the challenges associated with scaling up to integrate them effectively. To address this, we are collaborating with experts in political science, and this will be a key focus of future research.

We will reflect on this point in the discussion section to provide further clarity.

Line 371: What does a “link” between nodes signify in a word graph?

In a word graph, a link between two nodes indicates the two connected words appear within a specified proximity to each other in the text. Proximity can be understood as a window that identifies two words represented by a pair of connected nodes in the graph. The window size was set to 4 words in our analysis, indicating we connected the words within the proximity of four words. The thickness of a link means the frequency of two words coexisting in the same window across the whole text.

We will expand on this briefly in the revised manuscript.

Line 500: If you don't discuss this elsewhere here in this paper, it might be useful to know: how much your computers or the LLM computers need to work to produce these results? And how long from start to finish does a simulation take?

Thank you for your interest in the technical details. We did not record the exact runtime, but a full simulation could take several hours on a laptop with 32 GB RAM and 12 CPU cores (Intel i7-1260P). This is because we used an emulator to run the land-use model, which simplifies the computation required by the original software. One challenge was that the LLM-generated output did not always conform to a valid JSON format, occasionally requiring manual corrections before the hard-coded land-use model could process the data.

As LLM technologies have rapidly evolved, output reliability has significantly improved since our experiments. However, applying LLMs to simulate a large number of agent interactions (such as those involving over 20,000 land users) remains computationally expensive in both time and cost.

Also, in addition to the graphs, I would be particularly interested in seeing (for example) a time-ordered list of bullet points that are output by the various institutions. (This is to get more of a flavor of what messages are being passed between agents.)

Thank you for your suggestion. We agree that a structured, time-ordered summary of institutional messages would help illustrate agent interactions. However, given the length and verbosity of the original textual outputs (which are already in bullet point format) condensing the outputs into a readable list without losing key details is challenging. For a more comprehensive view, we would suggest readers refer to the uploaded dataset, which contains the complete textual outputs. That said, we are open to exploring ways to present these results more effectively. We appreciate your feedback and will reflect on how best to integrate this in the revision.