Reply on RC1

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

Summary of the review

Zeng et al. developed an innovative LLM model that simulates interactions between institutional agents that can mimic reasoning, planning and action. The model is novel because it addresses key challenges that learning and memory and polycentricity and because it is linked to an agent-based model that simulates changes in land use and livelihoods. The development of the LLM model described in the paper is ambitious and challenging. Certainly, it cannot be expected that all issues and challenges are yet addressed and that it completely functions as intended. The authors describe the challenges they occur and how they may solve them. It is impressive that the authors make sure that everything becomes available open access.

I read the paper with pleasure. It is generally well-written, novel and informative. However, there are a number of things that need improving in my opinion. That is why I recommend major revisions. In the attached pdf file, the authors can find detailed comments. Below they can find a summary of the issues that are in, my opinion, important to address.

The experimental set up:

The intention of the paper is to test the model and simulate institutional actor's behavior in the land system. Many different types of policy goals can be tested and different types of actors with different types of profiles and ways in which they interact can be chosen. The choices made in the experimental set up affect the outcome of the experiment. At the moment, limited rationale is provided for the experimental set up by the authors. There is no rationale provided for the choice of the SSP-RCP scenario. Additionally, limited rationale is provided for the choice of starting conditions and the types of policies that are considered. Limited rationale is provided for the choice of the combination of agents in the experiment. As this all influences the outcomes of the experiment, it is important that such rationales are provided. There is limited rationale provided for focusing only on the response of institutional agents to EU land system dynamics, without considering effects other regions in the world may have had on the results. Additionally, it is important to discuss how the experimental set up could have influenced the outcomes in unintended ways and which limitations of the model could have been accidentally missed because they did not come into play because of the way the experiment was set up. It would be great if the authors could address this thoroughly in the paper, so that the value of doing this particular experiment, but also its limitations, becomes clearer. At the moment, it was difficult for me to judge if the model is sufficiently tested using through this one experiment to run other types of scenarios with other policy targets, other institutional agents etc. Or whether more tests and sensitivity analyses are necessary for the model to be used more broadly. Especially since the outcomes of budget surplus for PAs and budget deficit for agriculture are a bit counter-intuitive in my perception and seem to reveal an overreliance of agents on policy documents.

Errors and robustness:

The authors speak of error proneness, error tolerance and robustness but these terms are not defined and the process of testing for this is not explained in the methodology. Usually, these terms are used in modelling literature in the context of quantitative sensitivity analysis but here they are used to refer to some unexpected or undesirable behavior of institutional agents. I find this personally a bit confusing, as I do not see so well how the error and robustness of the model could be derived from a qualitative assessment of the agents' behavioral patterns. Therefore, I would recommend to either use different terms, such as undesired or unexpected agent behavior or to really well define the terms around error well and thoroughly describe in the methodology how the authors assessed the errors. If the authors would really like to emphasize error proneness and robustness in the more traditional modelling sense, I would recommend the authors to do additional analyses. For example, to add a sensitivity analysis with different starting conditions, different environmental and social goals or different combinations of agents with different profiles, etc.

Information lacking to interpret results well:

As the LLM model is linked to CRAFTY, it is of course not possible to describe every dynamic of both models in detail in this paper. However, to understand the results and discussion some fundamental modelling assumptions were missing from the main paper, such as the way the budget is modelled and how the agents, for example know how much budget is needed etc. It would be great if the authors could provide more detailed descriptions of such assumptions and modelling choices, so that the results can be more easily interpreted. Or to very explicitly refer the appendices that are adjusted in such a way that the reader can understand the results and their interpretation easily after reading them.

Writing:

Although the model is intended to mimic real-life situations, the description of the model and the results, as well as the discussion of the results remains very high-level and abstract. I would highly recommend including real-life examples of agents in the context of the EU and to discuss the results in context of dynamics at play in the EU. This would all make it a bit more tangible. In particular I would recommend including a discussion of the outcome of the model in context of what happened in the EU in the past and what has been found in previous studies.

Writing style:

The paper is generally well-written. Yet, in some parts of the paper jargon is used and quite some terms that would be up for interpretation remain undefined. It would be good to more specifically define some of the terms, so that the model and results are easier to interpret by readers of different disciplines. This is important because the model can be used in interdisciplinary settings and, when linked to other models, such as CRAFTY can influence land use modelling, which is a different field again altogether. I have put comments throughout the paper that are hopefully helpful to address this.

Dear Reviewer,

Thank you very much for your thorough and detailed review of our manuscript. We greatly appreciate the time and effort you have invested in providing your feedback. For ease of reviewing and revision, we have categorized and numbered your comments. Below, we provide a general overview of your comments and our corresponding approach, with detailed point-by-point responses following in the subsequent pages. We have addressed many of the comments in the revised manuscript and will upload the completed revision upon Editor Müller's approval. The manuscript with your numbered comments is attached to the end of the responses.

- **Definitions and Terminologies:** We appreciate you highlighted the importance of providing clear definitions and conceptual clarity, especially as this research spans multiple disciplines, which may pose challenges for readers' comprehension if the terms are not adequately explained. We will make every effort to improve clarity through additional explanations and illustrative examples where appropriate, including adding a glossary in the appendix if necessary.
- **Additional Examples:** We recognize the importance of concrete examples in enhancing reader comprehension. We will incorporate additional examples where feasible to further support our arguments.
- **Rephrasing Suggestions:** We appreciate your suggestions for improving readability and will incorporate them accordingly.
- **Text Reorganization:** We will carefully consider your recommendations regarding the restructuring of text and implement changes where they improve logical flow and clarity.
- **Model's Working Mechanisms:** Many of your comments pertain to details provided in the appendices. While we attempted to move some of this information into the main text, we found that doing so often led to redundancy or a loss of contextual coherence due to the interdependency of the equations. To address your concerns, we have followed your suggestion to explicitly reference the relevant appendices wherever needed in the main text to ensure clearer guidance for readers.
- **Experimental Settings:** As noted, our model builds upon previous research, and many of the land use model settings have been documented in prior publications. We will strive to strike a balance between providing sufficient detail and avoiding unnecessary repetition. Additionally, we would like to mention that our framework is directly derived from an empirical study of the EU political system. However, it is intentionally designed to be relatively abstract and stylized, as the current experiments mainly serve proof-of-concept and exploratory purposes. The experimental settings of the AI agents are currently hypothetical and designed to be simple and straightforward. Rather than aiming to derive operable policies for real-world implementation, the primary focus is on investigating the internal logic coherence and contextual awareness of the AI agents, as well as exploring their potential integration with existing rule-based land use models.
- **Punctuation and Typographical Issues:** We greatly appreciate your meticulous attention to detail. We will carefully review and correct all punctuation and typographical errors.
- **Others:** We acknowledge these and will strive to address them carefully to improve the overall quality of the manuscript.

Sincerely,

Yongchao Zeng

On behalf of all authors

Page 1:

1. Maybe revise the first sentence. Understanding and modeling environmental policy interventions does not directly contribute to land use and management. Perhaps you can turn it around and state something like: To foster sustainable land use and management an increased understanding of ways in which policy interventions can contribute etc. etc. Modeling such processes can help derive this understanding.
Good suggestions. This sentence has been improved in the revised manuscript.

2. ...that, while...
Done.

3. This term seems a bit vague to me, the tolerance part I mean (but perhaps it's jargon from the field I don't know). Could you state something like that the network is robust to.... Or just low error or something.
Thanks. The sentence has been rephrased.

4. Very long sentence, quite difficult to follow.
Done.

Page 2:

5. ...topics, such as...
Done.

6. Reference missing.
Reference added.

7. Could you give an example here of actors and how they relate to each other (e.g., further building upon the EU example)?
The formation and implementation of land use policies are the product of complex institutional dynamics and can involve a wide range of actors with differing objectives and powers (Davidson et al., 2024). For instance, within the European Union (EU), multi-level governance systems play a significant role. At the EU level, institutions like the European Commission propose policies such as the Common Agricultural Policy (CAP), which sets broad objectives for sustainable land management, rural development and food security. These policies are then negotiated with the European Parliament and the EU Council, which consists of ministers from each EU country. National governments, in turn, are responsible for tailoring these policies to their domestic contexts, often collaborating with regional governments, local municipalities, and non-governmental organizations (NGOs). For example, an NGO advocating for biodiversity conservation might work alongside local authorities to implement EU directives, such as the Natura 2000 framework, ensuring compliance with broader policy objectives while addressing local needs. See Box 1 (at the end of the responses) for more details.

8. Could you define institutions and/or institutional dynamics? In other disciplines institutions can mean a range of things

Here we define institutions as organizations or governing bodies involved in policy-making, such as government agencies, research institutions, or NGOs. Institutional dynamics encompass the interactions, adaptations, and power relations among these institutions over time, which influence how policies are formulated, negotiated, and implemented.

We will add the above definitions to the manuscript.

9. It feels to me a bit like jumping to a conclusion. Could you explain a bit before how actors can shape land use systems and how their interactions and the influence this has on policies matters? Maybe using the example on the EU?
Referring to the CAP as an example once again, national and local government implementation can vary based on economic, social and environmental priorities. Local actors, including farmers and regional governments, further influence land use through on-the-ground practices and lobbying efforts. These interactions—whether cooperative or contentious—can result in policy outcomes that either advance or hinder environmental goals. For instance, tensions between biodiversity conservation objectives and agricultural production have led to debates over subsidies and land management practices, demonstrating how institutional dynamics can shape the land system and its environmental implications.

10. Would it be possible to move this paragraph after the next one (so after the paragraph that's currently the third one)? I think it is easier perhaps to understand what it requires to have a holistic representation of institutional actors. Hereafter, this second paragraph can provide information on how things were done before and then moving on to the LLMs. By doing this you likely also regard some of the comments I provided before and after this paragraph.
Thank you for the suggestion. We would prefer to keep the order as is, with the proposed additions to the paragraphs. As such, we would first state the current research to identify gaps, and then propose new approaches (i.e., the LLM method) to address these gaps.

11. The importance is not really well described. Please refer to my previous comments. In my opinion there is a bit more context and some examples needed to make this point.
Thanks, and yes, we believe the additional context and examples will allow us to better describe the importance here.

12. How is this defined?
We have added the institutions and institutional dynamics definitions above, in response to Comment 8. We will also rephrase the text to “networks of institutions” so it is clearer what we are referring to.

13. Better not to refer to previous paragraph like this at the start of a new one.
Thanks, the sentence will be revised accordingly.

14. Maybe better to not yet use "we" at this point in the introduction.
We will revise for consistent usage throughout.

Page 3:

15....model, InsNet-CRAFTY, and...
Done.

16. Adopted
Done.

Page 4:

17. Potentially two spaces here between maintains and Gonzalez.
Thanks, we have corrected it.

18. I am a bit confused about the terms "agents" and "actors". It seems you use these terms interchangeably in some cases but in other cases they seem to be used in a distinct way (e.g., in case of CRAFTY agents). Could you define these two terms and be consistent?

Yes, we will check for consistency.

"Agent" means a computational entity within the model that represents an institution, stakeholder, or decision-making body. Agents in the institutional network are powered by LLMs and autonomously make decisions, process information, and interact with other agents. Agents in the CRAFTY land use model represent various types of land users.

"Actor" is a general term for entities (individuals or organizations) involved in decision-making processes. Actors can be both real-world stakeholders (e.g., policy-makers, lobbyists) and their simulated counterparts (LLM-powered agents).

Page 5:

19. Could you give an example of a high-level institution in real life? And could you do the same for all other actors and agents? This would make things a bit more tangible and less abstract.
A good suggestion, thanks - we will do so. The level here is a relative term, and so real-world institutions can be at any level depending on the context being considered. We only need to know whether the role is relatively higher in the structure, and it is not the front-line institutions that execute specific policies. For example, the high-level institution can be the European Commission, which is a supranational policymaking entity. The high-level institution sets the overall policy ambitions and constraints (e.g., budgets) that affect the decisions of the Member States. Within the EU, the European Commission develops and proposes policy frameworks, such as the European Green Deal, aiming to achieve long-term goals like climate neutrality by 2050, based on the information provided by the operational institutions, research suppliers, lobbyists, and law consultants. For more details, please see Box 1.

20. Could you be more specific and/or give an example of this data?

We will specify here yes, by referring to the description of the data in the experiment setting section, i.e., "the data used here is described in detail in section 2.3".

Page 6:

21. I see you are still providing the example here of PAs and meat supply. I didn't realize up to this point that you were still providing an example. Could you make it more clear? I suddenly thought this was a specific focus or the only thing that could be modeled.
We will make it clearer in the first mention of meat and PAs that this is just an example. We will also modify the text to explain the importance of this example:

"For instance, CRAFTY can produce information indicating that the supply of certain food products (e.g. meat) and the coverage of protected areas (PAs) need to be improved to achieve better food security and nature conservation. This is an important example, as land is a finite resource and both meat production and PAs compete for this land."

22. ...instruments, such as...
Done.

23. Would it be possible to expand slightly on this? How asynchronous is the nature of agent decision-making across levels typically?
I am also wondering how this works with the long-term and short-term memory. Perhaps I am not fully understanding something in that paragraph. Would this mean the agent is getting a lot of short-term memory in its prompt and this is saved until the agent acts? Could you comment on that in the paragraph about memory?
We will expand on this yes - please see the response to comment 31 regarding the memory mechanisms of LLM agents.
In the EU governance system, decision-making is often asynchronous, with different institutions operating at different levels. The European Commission, as the supranational governing body, primarily focuses on long-term strategic policy goals (e.g., the European Green Deal, CAP reforms, etc.), setting overarching frameworks for Member States. In contrast, national and regional institutions make more frequent policy adjustments to ensure effective policy implementation and adaptation within their specific contexts.

24. You write the LLM have a polycentric structure. From the text I can't completely visualize this and understand why it is a polycentric. It seems to me, for example, that quite a lot of power still lies with the high-level institution, so I don't see the polycentric structure completely myself when observing figure 1. Could you perhaps describe why you consider the structure to be polycentric?
Although the high-level institution does imply some hierarchical decision-making, the structure represents polycentric governments in the real-world. The operational institutions are relatively autonomous bodies that can make decisions on their own and operate independently, yet interact with each other. The design of the model's polycentric structure was influenced by the real-world example of the European Union, which has also a high-level institution (European Commission), but decisions on implementation are done through autonomous member-state government departments. To improve clarity in the text, we added Box 1 describing the link from the conceptual model (seen in Fig. 1) with the example of the European Union.

25. Could you define this? Does "brain" equal framework?
"Brain" here is used as an analogy to the cognitive architecture of an LLM agent. As you raised this question, it seems this word caused confusion instead of the clarity initially intended. We

have deleted this analogy in the revised manuscript to avoid potential confusion.

26. Could you provide a rationale for using this specific framework? What is the benefit of this one?
We can say a little more, but fundamentally this framework is used because it is useful and sufficient for our purposes, containing extensive elements and being able to represent a range of agent cognitive architectures from simple to complex.

27. Is what you mean by "brain"?
Please see the response to Comment 25.

28. "Synthesizing results" (with a capital for consistency).
Done.

29. Could you define tools? Either in the text or in the caption?
Update: I see you define it later on. Could you, to be able to better read the figure, also define it in the caption of the figure?
Yes, done.

Page 7:

30. ...information, such as...
Done.

31. This is the part where my question on the frequency versus memory comes in.
Memory here is not directly related to policy adjustment frequencies. It is more about the technical approaches that determine how an LLM agent handles information. For instance, a prompt with updated information that can be directly input into an LLM constitutes short-term memory. In contrast, information stored in external containers or external files—such as datasets containing historical land use outcomes or a knowledge base with a subset of EU policies—is typically considered long-term memory. Long-term memory reduces reliance on the context window (the maximum number of tokens an LLM can process at once) by enabling the storage and retrieval of vast amounts of information beyond what the model can handle in a single input. We will expand the explanation here and include a citation to a comprehensive survey on the memory mechanisms of LLM agents in the revised manuscript for the reader's reference.

32. How does the agent evaluate this exactly? I am not sure if I understand well the interaction between, I assume, quantitative output from CRAFTY, that shows whether or not policy goals are being reached etc. and the translated qualitative information the agents receive in which this is reported to them.
For instance, if there is a policy goal on expanding protected areas to 30% of EU land cover, would the agent receive a qualitative "true" or "false" based on a calculation of the extent of protected areas in CRAFTY? For clarity perhaps it is useful to provide an example.
We did not instruct the agents on how to know whether the policy goals were achieved in the experiments. We intended to give the agents high autonomy to analyze and interpret the data in order to let the agents expose their "intelligence" level while keeping the prompts simple. Tables B4 and B5 cited in the experimental setting section have shown the prompts for the agents. As shown in the results, the agents did not do well in this. They compared the average

values of the time series of the land use outcomes with the policy targets, which is not correct. However, this can be avoided by simply prompting it to use the latest data points.

33. In my opinion, the structure of this Section can be improved. I have provided comments that could help guide this process. I general, I suggest to put in one paragraph per agent/actor type and discuss the targets and other parameterization for these agents in that paragraph instead of splitting it up.
Thank you. We will revise the manuscript as per your suggestions.

34. It would be helpful if CRAFTY-EU and the RCP2.6-SSP1 parameterization summarized in the supplementary materials.
We have included the reference to the original publication for information, but can summarize in the supplementary materials.

35. Could you reflect on the potential limitations that may arise from focusing on the EU only? Many social-environmental dynamics in regions outside the EU affect agents and actors as well as policies in the EU. What is the consequence of not taking interactions with regions outside the EU into account? Or is this considered? If it is considered could you please describe this and how this influences the experiment?
Applying this state-of-the-art approach to broader regions will surely be meaningful. But at this exploratory stage, we are not aiming at drawing very region-specific conclusions to guide real-world policymaking. Instead, the focus of this research is on exploring the potential of using AI agents to simulate institutional actors within the existing land use model. We focus on building AI agent workflow, AI agents' capability of autonomous decision-making, exchanging unstructured data, and constructing complex relational structures in institutional networks, as well as how they should be coupled with programmed systems. The evaluation of behaviours is focused on their contextual awareness and logic consistency. We use EU as it provides a nice example for multiple levels of governance, and relates to the spatial coverage of the CRAFTY model used. This research can of course be applied to other areas in future research.

36. Why this scenario type? Why is this a useful scenario to experiment with?
Please see the response to Comment 39.

37. Why simulate until 2076?
This is purely because of data availability.

38. You forgot a dot at the end of this sentence.
Dot added.

39. I do not completely understand the rationale behind the specific experimental settings and parameterization you describe here. Likely this is very understandable of course because you build on previous studies. However, it would help to include a bit more rationale for the choices you make. Why is it useful to consider PAs and meat production? Is it because they have contrasting lobbying groups and operational actors with opposing interests? Why is it useful to consider RCP2.6-SSP1? What is the effect choosing this particular experimental setting over another? I personally think the paper would be a lot stronger if this is better explained and understood by the reader. How can I interpret the outcome of this experiment in context of other types of experiments that could be done and other types of RCP-SSP combinations? Etc.

Thanks for the suggestion. The scenario was chosen really to provide a background to initialize the land use model. The 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 relatively simple and straightforward baseline for us to investigate the impact of LLM agents. The focus here is more methodological and conceptual rather than empirical. In addition, the agents' believable behaviours should be independent of specific scenarios. The agents are expected to perform reasonable, logically-consistent, and contextually-aware actions to steer the specific model outcomes towards the target level. We choose to incorporate meat production and PA expansion because this provides a conflicting environment. Also, meat production and PA expansion represent two distinct policy instruments – economic incentives and area-based regulations.

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 add clearer rationales throughout.

40. What might work well here to improve clarity and structure is if you split up the pieces per agent type up in the same way as you do in Section 2.1 with italic agent types followed by the experimental setting for that agent, including also immediately the target setting and other parameterization. Of course not necessarily in the same order as Section 2.1 but in an order that works best.
Thanks for the suggestion; we will consider rearrangement for clarity.

41. In the context of this experiment, what would be real-life equivalents of these institutions? Could you give examples? I think it would help to make things less abstract if the reader can keep examples of existing institutions in mind.
An example that illustrates the institutions and governmental structure of the model is the European Union. The operational institutions could be seen as member-state governmental departments. We added Box 1 including a reference to the operational institutions as follows "National scale implementation is done through various government departments that are usually responsible for a specific policy sector, e.g. agriculture, environment, research, etc. Within the model, these government departments are represented by the operational institutions, and the multiple instances of the operational institutions represent the different policy sectors.".

42. ...instruments, such as...
Done.

43. Long sentence, a bit difficult to follow. I suggest to split up or shorten.
Done.

44. Replace "They" with "The operational institutions" for clarity.
Done.

45. You mention the targets later on. I was immediately wondering about the specific targets.

It would help the reader if you move the piece of text about the targets to this paragraph.
Done.

46. Could you provide a real-life example in the context of this experiment?
For example, the European Livestock Voice group advocate for policies that support the sustainability and economic viability of the European meat industry, whilst organizations like EUROPARC seek to secure funding and favourable policies for the conservation and sustainable management of protected areas in Europe.

47. I am a bit confused here whether this piece about the high-level institution applies only to the lobbyist you discuss in this paragraph or whether this is a more general statement that concerns all agents. If the former is the case, could you specify? If the latter is the case, perhaps you can put this information in a seperate paragraph?
Good suggestion. Done.

48. Like for the operational institutions, could you perhaps provide a real-life example of high-level institutions that could play a role in the context of the experiment? This would make everything less abstract.
A real-life example of the high-level institution is the European Commission. We have added a reference to the high-level institution in Box 1 as follows: "The executive body of the EU is the European Commission, which is responsible for enacting new legislation. This is equivalent to the high-level institution in the model. The European Commission proposes new Directives and other policy measures that are ratified by the European Parliament, but which are then implemented by the member states (national governments)."

49. Which agents? Only the lobbyists discussed in this paragraph or more broadly?
We have divided the original paragraph into two to make things clearer.

Page 8:

50. Merge with the paragraph on lobbyist.
Additionally, could you describe the experimental settings of the research supplier? (I guess this is considered an "advisor" actor?)
51. Suggestion to put this sentence completely at the end, after you have described the experimental settings of all actors that are involved.
52. Move to part about operational institutions.
Response to Comments 50 - 52: Thank you. We will follow your suggestions in the revised manuscript.

53. You mean the operational actors here right?
Yes. We have modified the text here to improve clarity.

54. What is the rationale behind this? Could you describe this? In real life it is not usually the case, so I am wondering how realistic this setting is.
I can imagine setting an unequal budget allocation could really influence the results and also the reliability of outcomes. Could you reflect on this in the discussion?
We can reflect this in the discussion yes. The main reason is that an unequal initial budget allocation would be harder to justify at any specific level than an equal one under the hypothetical experimental settings while making the results harder to interpret. In future work,

we hope to work alongside policy makers and other stakeholders to parameterize the budget allocation using real-world data, to be able to simulate how funding disparities impact the achievement of policy targets. This is not trivial, as budgets for individual policies are not clearly documented and funding to reach specific targets tends to come from multiple sources.

55. Transfer this statement to a place in the previous text where you talk about budgets.
Good suggestion. Done.

56. I did not know what this is and what is meant by it. Perhaps other readers might also not know it. Could you use different terms to describe this?
Done. Tokens are the basic units encoded to train large language models. Tokens can be derived from words, sub-words, characters, etc. They are normally shorter than complete words.

57. I am guessing "Llama-3-70b-8192" and "gpt-4" refer to things in this library. Could you expand on them here?
Done. They are different versions of models. We have changed the title of the column named "LLM" to "LLM version" to improve clarity.

58. ...behaviour, as well...
Done.

59. A dot is missing at the end of this sentence.
Done. Dot added.

60. It would help perhaps if you put in one extra column in which you specify the agent type (operational, lobbyist, advisor etc.).
Good suggestion. Done.

61. Sometimes you use "Goals" and sometimes "Goal", while it is often not super clear whether what is described is one or multiple goals. I would suggest to be consistent.
Good suggestion. We have changed the word to Goal to improve consistency.

62. I see that indeed quantitative CRAFTY data is transformed into qualitative text. I am still wondering a bit how, for example, things like "we reached the policy goal" is communicated qualitatively. Is this very consistent for each target (e.g., TRUE / FALSE output from the tool? Or is reaching the policy target up for interpretation by the agent based on more vague text?
As mentioned in our response to Comment 32, we did not provide specific instructions on how the agents should determine whether the goals were met. Instead of directly converting data into text, the data were processed using code written by the LLM agents. This code produced numerical results, which were then incorporated into the prompt for the LLM agent to interpret. The agents had full autonomy in deciding how to process and interpret the data, using any relevant computational approach. For example, an agent might calculate the gap between policy goals and actual supply to assess progress. However, the interpretation was not strictly constrained to numerical analysis, as we did not specify in the prompts whether the outputs should include quantitative elements.

63. Why are there different LLMs? Could you describe a bit what is done in each of them and how they differ?
Because Llamma is free but gpt-4o costs money, and both of them have a rate limit per minute imposed by the LLM providers. If we call the LLMs too frequently, the API will report an error, which will disrupt the simulation. So, switching between models can avoid the rate limit and save costs. We will explain this more explicitly in the revised manuscript.

64. Could you define this? What does intermediate output mean and what is done with it?
When an AI agent is solving a problem, it often breaks the task into steps instead of giving an answer right away. The results from these steps are called intermediate outputs. Figure 3 illustrates a sample of these intermediate outputs. We will also explain this concept explicitly in the revised manuscript.

Page 9:

65. Perhaps make sure the column width is adjusted to the width of the words.
Sure. We will adjust the column width to better present the words.

66. How is it known what budget is needed? Does this change over time etc.?
Update: I have found information in the Appendix on budgeting. I suggest to move information to the main paper as it is so fundamental to understand the methods and results.
Thanks for noting this. We found that incorporating the budgeting information from the appendix into the main text would make the manuscript unnecessarily lengthy and potentially disrupt the reader's experience. The equations and symbols are interdependent, and relocating the budgeting details would require moving all related technical content as well. To maintain clarity and accessibility, we aim to keep the main text more conceptual while consolidating the technical details in the appendix. This approach ensures that readers with varying levels of interest in the technical aspects can engage with the material at their preferred depth. We appreciate your suggestion and hope this explanation clarifies our rationale. We will refer to the Appendix wherever needed to better guide the readers' attention.

67. For all other institutions (e.g., operational institutions) names were provided, such as agricultural institution. It would be useful if you do the same here for the high-level institution. It would also be helpful if you could put some examples in the text of what a real-life example is of such institution in the context of the experiment.
We have added Box 1 to provide an example of a high-level institution, such as the European Commission. However, directly naming this agent the European Commission could be misleading, as real-world supranational decision-making bodies in the EU context are much more complex than the current agent intends to represent. To maintain flexibility and avoid misinterpretation, we have chosen to keep the institution's name generic throughout the text, allowing future, more empirically focused research to incorporate necessary details.

68. In my opinion, it would be good to describe general results of CRAFTY before going into the other outputs. I find it difficult the interpret the results of the textual outputs without having a context on what was happening with LU in CRAFTY and how this was influenced by the actors. For instance, in the last paragraph of lobbyist it is stated that: "Its discourse primarily focused on two aspects: the environmental threats posed by meat production and the critical importance of conservation efforts. This concern was underscored by the research supplier's data interpretation showing a widening gap between meat demand and supply." I see here that there is a link to the output of CRAFTY but such things are only sporadically mentioned so it

is difficult to interpret the output for the institutional actors in the context of LUC and how they influenced this process in CRAFTY (hopefully this is clear enough to understand what I mean..).
Thanks, we understand your questions. We will highlight CRAFTY results as suggested where essential to understanding the results presented here, and otherwise refer more clearly to previous publications where extra details can be found.

69. Please define what this means. I would suggest you discuss in detail how you test for this in the methodology. Or highlight what criteria you use to qualitatively interpret the behaviors etc.
Erroneous behaviours means any unintended or incorrect actions performed by an LLM agent. In the manuscript, erroneous behaviour includes misinterpretations of data, incorrect formatting, logical inconsistencies, or failures in workflow execution. We examined all the agents' output manually, intending to capture mistakes or errors generated by the agents. There is no standardized or established method to evaluate agents' outputs especially when their outputs are unstructured and dependent on the logical validity. In the Results and Discussion sections, such behaviours have been highlighted. Some mistakes are straightforward. For instance, the research supplier agent failed to generate a final output because its "thinking" process used too many iterations or too much time; The high-level institution agent occasionally failed to follow the instructions to consider all other agents' input. Some mistakes are less obvious. For example, the operational institution agents were found using the mean values of the times series rather than the latest model outcome to evaluate how close the current outcome is to the policy goals. This mistake requires us to examine the intermediate outputs carefully to see how the data were used.

70. Please define (see previous comment)
We will amend and explain more explicitly.

71. Here you split up nicely between the institutional types. I would suggest to also do this in the experimental setting section. Perhaps even in this exact order.
We will reorganize the text in the experimental setting section while avoiding repeating the information in Table 1.

72. This advisor is not so much described in experimental settings. Would be useful to have a bit more background.
Table 1 has highlighted the input, action, output, and goal of the research supplier agent. The agent's prompt (Table B1) provided well-formatted information to define what the agent's role is. We will cite Table B1 in Table 1 to better guide readers' attention.

Page 10:

73.This seems to me something that belongs in the methodology. It is not described there. I would suggest to describe it in the methodology and only repeat it shortly here.
Done.

74 ....mitigation".
Done.

75. To me it was not clear that other laws and policies that those directly regarding PAs and meat could influence things. This makes the experiment a bit broader that I expected. Could you state this explicitly in the methodology? To which extent do other policies not directly related to PAs and meat have influence? I see here the agent is interpreting a nature restoration policy in the context of PAs even though it does not directly apply to PAs. In this case it makes sense to a certain extent, even though I am not sure how this would influence PA expansion efforts eventually (would it shift to protecting degraded areas with the aim to restore them and would this then reduce the area of, for instance, intact forest to be restored?).
Thank you. This is an interesting point. It is true that the influence of policies beyond those directly targeting PAs and meat production has not been explicitly stated. In the model, institutional agents, particularly the law consultant and high-level institution, can interpret a range of policies and regulations in ways that extend beyond their immediate scope. This reflects real-world policy-making, where broad legislative frameworks (e.g., biodiversity strategies, restoration policies, and climate regulations) can indirectly shape sector-specific decisions.

Indirect policies can shape the agents' narratives and lead them to propose actions beyond what is explicitly represented in the land-use model. However, the agents have limited direct influence on the land-use model's operations: the high-level institution can only adjust budget allocations and policy goals, while operational institutions can only impact meat supply or PA coverage. As a result, policy actions suggested in the agents' textual outputs do not translate into direct interventions that alter the land-use model's behaviour. A clear example, as noted in the manuscript, is the high-level institution's proposal to allocate 30% of the total budget to "other initiatives and programs", a policy action not represented in the model.

Assessing how textual variations in inputs affect the numerical results of the land-use model is inherently challenging, given the vast number of possible word combinations. As a general guideline, it is advisable to select only the most relevant policies for agents to reference, as excessive text may dilute the LLM's attention and hinder reasoning performance (Levy et al., 2024).

We will highlight this point in the discussion.

Page 11:

76. This is quite logical right? This is inherent to the experimental setting described in Table 1. Or does the high-level institution have the power to create new "EU laws, policies, regulations, etc."?
Correct, this agent is not able to create new laws. It can only select a subset of policies in the static knowledge base. The law consultant agent behaved as expected because this agent did not receive enough updated information. The law consultant agent is designed to offer a set of constraints that may modify the high-level institutions' behaviour. A more sophisticated law consultant agent could be built in future studies with more dynamism.

77. Sentence seems incomplete. ... in other years seems to be missing. I suggest rephrasing.
We think this sentence is adequate as it is, but will check for clarity in the context of the revised version.

78. It would help the reader if this figure is positioned after the paragraph or the piece of text that refers to this figure (last paragraph of Section 3.1.2).

Done.

79. This is not explained in the methodology. I see you explain it in the results below. I suggest to explain such things in the methodology.
We will try to follow your suggestion and add a description of these analysis methods to the methodology section without disrupting the text flow.

Page 12

80. Much of this belongs in the methodology.
We will reorganize the text accordingly.

81. In the above result section often such statements are made but I don't understand so well where they come from and how to interpret them in the context of what is described in the methods and in the context of the experiment. As far as I understood, the input to the lobbyist agents is the research supplier's output and the profile. In that case, is it specified in the profile that the lobbyist always advocates for increased budgets? It is described that the agents compete for budget but not how the budget is used to implement policies (how expensive are things, how do agents know that they don't have enough budget, or do they always want more budget anyway?)
There are more such statements all throughout the result section. In my opinion, it would be helpful to go through the statements and see if all information needed to interpret them is provided in the method section.
Please see the response to Comment 82 below.

82. I see here also most output concerns budget challenges. Could you describe this a bit more in the experimental settings? How do budget challenges arise? And how is it determined how much it costs to implement certain policies etc. I expect this data comes from CRAFTY?
In general, it would be good to describe budget allocation and use. How, for example, can a budget be inefficiently allocated within the context of the model?
Response to Comments 81 and 82:
Tables B1 and B2 show the prompt templates of the lobbyists. They are not instructed to advocate for increased budgets but to prompt the high-level institution to prioritize nature conservation or meat industry development.

Regarding the budget challenges, the prompts in Tables B4 and B5 instruct operational institutions to seek increased priorities and financial support for meat production or protected area establishment, and to prompt the high-level institution to set policy goals that align with budget allocations. Consequently, agents are motivated to request additional budget if policy goals are unmet.

The costs of policy adjustments are determined via Equations C2–C7, which evaluate policy performance and the resources required for subsequent adjustments. Table C1 provides the parameter settings for policy-resource functions, indicating the budget needed for corresponding policy adjustments. When one institution consistently experiences a surplus while another faces a deficit, it signals inefficiency. This can be addressed by reallocating funds or adjusting policy goals to better align with available resources, improving overall financial efficiency and policy effectiveness.

We will explicitly explain and cite the information from appendices in the main text to better guide readers.

Page 13:

83. Here again policies come into play that do not relate directly to PAs. It would be good to understand how the agents combine the policies that directly apply to PAs to those indirectly applying to PAs. In principle, net-zero targets are not direclty related to PA expansion in EU, protection of biodiversity is.
Please see the responses to Comments 75 and 104. We will discuss this in the revised manuscript.

84. Is this what is meant by "potentially erroneous agent behaviours"?
Yes, this is one of the erroneous agent behaviours.

85. What about the researcher agent? Did this agent make the same mistake? Did this lead to inconsistent supply of input to the high-level institution?
Good point. Unlike the operational institution agents' outputs, there is no numerical information explicitly mentioned in the agent's interpretation of data analysis. This might be because these agents used different large language models with different prompts, which could influence the agents' "intelligence". However, it is reasonable to suspect that the researcher agent could also make the same mistake. A safe way to address this issue is to explicitly instruct these agents to use the latest data points to determine the current state of the land use outcomes.

86. You mean here that the agent did not change their conclusions based on this error?
Yes. Gaps exist between the policy goals and actual outcomes (meat supply and PA coverage) in most years. Thus, the mean values of the actual outcomes are lower than the policy goals, which is qualitatively consistent with comparison results using the latest data points. This mistake needs careful examination of the intermediate outputs. Otherwise, it would be more evident to capture and easier to correct during prompt design.

87. involved
Corrected.

88. Could this be related to the short-term vs long-term memory?
This could be caused by the failure of the LLM to follow a very lengthy prompt. Please refer to the response to Comment 31.

89. I see here the policy actions and outcomes are described. It would be quite informative to put these results prior to the ones on the institutions output to provide background necessary to interpret the output.
Or perhaps the PA changes, LU changes and meat production changes can be discussed prior to the institution output, and the feedback with the institutions and the CRAFTY output either in the institutions output section or a separate section below.
I feel though quite some pieces in this section arguably belong in the discussion. I have highlighted them.
Thanks for the suggestions; we will revise for flow and clarity.

90. Which assumptions were made for PA allocation? Would be good to specify.

We have cited the corresponding text in appendices to better guide the reader's attention.

91. This seems to belong in the discussion.
We will consider moving this to the discussion or removing it.

92. This paragraph contains a combination of some results with a lot of discussion. I suggest to move it to the discussion. And expand more broadly on the results for PAs, meat production and LU prior to the institution outcomes.
While we have limited room to expand all the points raised in the reviews, we will try to improve this section including by moving text to the discussion.

93. Repetition.
We will check both of these points for repetition.

94. Repetition
As above.

Page 14:

95. Could you define "Gap" in the figure caption?
Done. The gap means the difference between the policy goal and the actual outcome (e.g., meat supply and PA coverage).

Page 15:

96. Maybe rephrase as: ... resulting in a sum of the budget allocation ration of 0.7.
Good suggestion. Done.

97. This seems more fitting for the discussion.
The textual output and numerical analysis are matched as the latter is the outcome of the former. We will rephrase the text here.

98. In order for the reader to judge whether behavior of LLM agents is believable, "believable behavior" needs to be described and defined. I suggest to do this in the methodology (see earlier comments). Additionally, there seems to be a mix of terms: believable behavior, erroneous behavior, counter-intuitive behavior etc. It may be useful to be consistent. As none of them are defined in the text the readers are left to interpret them for themselves, in my case resulting in different interpretations and confusion.
We will clarify their meanings.

Believable behaviour means that an agent’s actions and decision-making processes resemble realistic human behaviour.

Erroneous behaviour indicates unintended or incorrect actions performed by an LLM agent. In the manuscript, erroneous behaviour includes misinterpretations of data, incorrect formatting, logical inconsistencies, or failures in workflow execution.

Counter-intuitive behaviour refers to an emergent pattern or decision-making outcome that deviates from conventional expectations or common sense. In the context of the manuscript, counter-intuitive behaviour occurs when the high-level institution's decisions do not align with typical policy-making norms, such as the unexpected budget allocation, which may contradict the assumption that the percentage of the total budget should be 100%. This behaviour does not necessarily indicate an error but rather an unanticipated outcome driven by the agents' decision-making.

Page 16:

99. ...tasks, such as...
Done.

100. ...challenges, such as...
Done.

101. Ok now I understand. Please describe this in the methodology. How is the high-level agent making these decisions? How does it know how expensive things are etc.?
This has been described in the methodology (please see Table 1 and Table B7). The decision-making details are presented in the equations in Appendix C. We will cite them explicitly in the methodology section.

102. This sentence does not flow very well. I suggest rephrasing it.
Done.

103. Perhaps I don't have enough knowledge on EU dynamics, but this outcome does not seem very intuitive to me. In general there is not enough funding to protect nature worldwide (I am not an expert on EU though, so perhaps I am very wrong in this context).
Would it be possible to discuss the outcome in context of past dynamics and prior modeling studies? To discuss whether it is realistic that the environmental institutions had so much influence?
It seems almost like the opposite of real life is happening, where policies are very influential (even though that is often not the case, for instance in EU already around 1 billion trees out of the 3 billion should have been planted but we are at 22 million) and the wishes of stakeholders on the economic and livelihood side of the spectrum have almost no power. In the case of The Netherlands, the opposite occurs currently, where the state has taken very limited action on, for instance reducing nitrogen, not reaching the target by far.
Perhaps you can discuss the findings in the context of such examples (but more fitting ones)?
Please see the response to Comment 105 below.

104. Here the other indirect policies come into play. Could you further discuss on the interactions between direct and indirect policies related to PAs and meat production? Do the agents understand the difference?

Thank you for your question. Understanding the interactions between direct and indirect policies related to PAs and meat production is a highly complex task. Analyzing such interactions in depth would require a dedicated study, as it involves effort beyond the current scope of our research. While our model captures how institutional agents process and interpret various policies, it does not explicitly distinguish the causal relationships between direct and indirect policy impacts. Future work could explore this issue more systematically by designing experiments that isolate specific policy interactions and their cascading effects on land use changes.

The extent to which LLM-powered agents "understand" the distinction between direct and indirect policies requires careful consideration. Our approach does not assume that LLMs possess genuine comprehension but rather utilize their ability to process textual patterns and generate human-like reasoning. Below, we clarify how our agents operate in relation to distinguishing policies:

LLMs and policymakers can distinguish textual patterns. LLMs are pre-trained on vast amounts of text and good at recognizing textual patterns. If the policy instruments and targets have strong textual relevance, LLMs should be able to differentiate between direct and indirect policies, just as human policymakers do. Policymakers rely on experience, legal frameworks, and historical precedents, while LLMs rely on learned associations from their training data.

Textual relevance alone is insufficient for effective policymaking. However, effective policymaking is not just about distinguishing textual relevance, It is about predicting and evaluating policy outcomes. Understanding the real-world impact of policies, including their unintended consequences, is a highly challenging task in complex socio-ecological systems. Given these complexities, we did not expect current LLMs to outperform human policymakers in making policy impact assessments.

LLMs do not truly comprehend, but they mimic human reasoning. While some researchers have suggested that sophisticated reasoning capabilities in LLMs could emerge through reinforcement learning (often referred to as the “aha moment” (Guo et al., 2025)) current evidence indicates that LLMs still lack true comprehension of policies and their effects. Instead, they utilize contextually meaningful natural language to mimic human reasoning. While this approach allows them to generate plausible and logically structured arguments, it is not free from hallucination (factually incorrect or fabricated outputs). That said, LLM technology is evolving rapidly, and the rate of hallucination has significantly decreased, which also suggests that the logical coherence within their outputs has improved over time (please see the Hallucination Leaderboard).

Fine-tuning or prompt engineering could improve distinction, but we chose autonomous decision-making. LLMs can be fine-tuned or prompted by end-users to better distinguish direct and indirect policies, improving their reliability in specific policy contexts. However, in this study, we did not manually engineer prompts to enforce such distinctions. Instead, we allowed the LLM agents to autonomously decide how to interpret policies, as the goal of our research was not to design highly optimized LLM agents but rather to evaluate both their potential usefulness and their limitations in institutional decision-making.

We will discuss these points in the revised manuscript.

105. In real life you often see exactly the opposite. It seems to me that environmental policy documents have too much power and are too influential in decision making about budget allocation. Could you reflect more on this using prior studies on European context?
The numerical results show that the environmental institution in our simulations was over-funded, while the agricultural institutions were struggling with an inadequate budget. This

diverges from intuitive expectations based on real-world dynamics and is caused by the parameterization assumed (see Table C1 for details): the budget settings are hypothetical and are roughly set to make the budget needed for different policies the same scale and comparable. Future efforts could focus on calibrating the model with empirically accurate data as the research evolves toward empirical analysis and the development of actionable policy recommendations.

Globally and within the EU, environmental policies and initiatives often face significant funding constraints. For instance, the EU's ambition to plant 3 billion trees by 2030 has seen limited progress, with only 24 million planted thus far (*3 Billion Trees initiative*). Similarly, challenges in reducing nitrogen emissions in countries like the Netherlands highlight the limited influence of environmental policies in the face of economic and political barriers.

This imbalance was also prompted by the environmental institution and the research supplier misleadingly informing the high-level institution that PA coverage was positively correlated with budget surplus. Indeed, both the two operational institutions insisted that their respective policy targets (PA coverage and meat production) should be increased because those targets were positively correlated with other desired outcomes. While these mechanisms aim to capture real-world lobbying behaviours, they may overestimate the environmental institution's effectiveness in securing resources.

The influence of environmental concerns might also come from the biases in the LLMs' training data, as text containing social norms favouring environmental protection over economy might play an important role in the training process. LLM biases have been well documented in the literature (see e.g., Zhou et al., (2024)) and can be rectified by prompt design and fine-tuning (Taubenfeld et al., 2024; Tao et al., 2024).

To align the findings with real-world contexts, future iterations of the model could incorporate a more nuanced representation of political and economic pressures. This would include explicitly modelling the comparative lobbying power of agricultural stakeholders and the structural barriers faced by environmental policies and organizations, in part through unequal budget distribution. By doing so, the model could offer a more balanced reflection of institutional dynamics, providing deeper insights into the challenges of achieving sustainable policy outcomes.

We will reflect on these points in the discussion section.

106. Could you please expand on this? I think this is fundamental and highlights one of the "counter-intuitive or potentially erroneous agent behaviours". Could you for example reflect in more detail on how an over-reliance on policy documents to inform decision-making by the high-level agent could be prevented?
Yes, we will discuss LLM biases in the revision.

107. I don't fully understand why you conclude that there is a lack of decisive action by the high-level institution. The budget allocation significantly changed and there was a lot of change in PAs in response to the actions of the high-level institutions (at least this appears to be the case from what was written in the results). Could you reflect on this and indicate what you mean with a lack of decisive action in this context?
It is true that there were significant changes in PAs over time. However, these changes were

primarily driven by the availability of surplus funds rather than strategic intervention by the high-level institution. The environmental institution had no budgetary pressure, allowing it to continuously expand PAs. In contrast, the agricultural institution consistently experienced a budget deficit, which should have prompted the high-level institution to reallocate more funds towards meat production to balance resource use.

Despite these circumstances, the high-level institution did not make sufficiently large budget reallocations to correct the imbalance. For most of the simulation (40 years), the budget distribution remained 60% for PAs vs. 40% for meat supply, and only in the final 10 years did the ratio shift slightly in favour of meat production (55% vs. 45%). However, this modest adjustment was insufficient to reverse the long-standing budget surplus for PAs and the deficit for meat supply. A truly decisive action in this context would have been a more prominent reallocation of funds—such as shifting the majority of the budget toward the underfunded agricultural sector at an earlier stage.

Thus, when we refer to a lack of decisive action, we mean that the high-level institution failed to make bold, corrective adjustments to its budget allocation, instead following an incremental approach. This behaviour aligns with path-dependent decision-making, where institutions adjust policies gradually rather than making radical shifts, even when inefficiency persists.

108. I am personally not very convinced that the behavior is believable (also here the term "believable" is rather abstract and not defined). A big budget surplus for environmental issues compared to agriculture seems quite counter-intuitive to me. As far as I am aware, the opposite is usually the case.
Could you discuss this please in context of past and current dynamics in the EU and prior studies?
Please see the response to Comment 105 for budget surplus issue.
The believability of agent behaviours in our model is determined by whether they capture aspects of real-world decision-making within the given experimental context. The experimental results indicate that the high-level institution was unable to sufficiently reallocate funds from the over-funded environmental institution to the under-funded agricultural institution. This occurred because the institution had to balance multiple, conflicting stakeholder inputs, leading to incremental rather than decisive policy shifts. Instead of effectively correcting the budget imbalance, the high-level institution followed a decision-making process constrained by the necessity of considering multiple institutional perspectives. This behaviour reflects some aspects of real-world policy-making in democratic systems, where competing interests, bureaucratic inertia, and consensus-driven decision-making often limit drastic policy changes (Lindblom, 1959; Jones, 2003). The model thus reflects bounded rationality, where policymakers operate within cognitive and informational constraints, and incrementalism, where policies evolve gradually rather than through radical shifts. While the model may have overrepresented the shift toward environmental funding, it captures the challenges of redistributing resources in multi-actor governance settings, such as those found in the EU's Common Agricultural Policy (CAP), where competing priorities frequently lead to compromise-based rather than optimal budget allocations (Daugbjerg & Feindt, 2017).

Page 17:

109. Please define.

Done.

110. What does this mean? Please define.
Done.

111. Reference is missing.
Reference added.

112. Could you perhaps reflect on whether this is realistic? I can imagine this happens quite a lot in real life? Is there an example of this happening in the EU?
We will try to add an example.

113. large-scale land use models
Done.

114. How does this work in case of the high-level institution which operates at lower frequency? Is more often the maximum number of tokens reached? If so, how does this influence the results?
The frequency here does not have direct connections with token numbers. The frequency is introduced to simulate the time lags of policy adjustments, as policymakers in the real world normally do not respond to land use changes immediately (Watts et al., 2020).
The maximum number of tokens that can be processed by an LLM like GPT-4o used in this paper can reach 128000 (see the documentation of the models here), approximately 96000 words estimated based on the rule of thumb that 100 tokens equals 75 English words. Technically, the context window is generous. However, research shows that the reasoning performance of LLMs drops notably as the length of prompts increases even if the length is far less than the technical maximum (Levy et al., 2024), which means the best practice is to use concise prompts; otherwise, the LLM outputs might be more prone to mistakes.

115. It seems useful to put this in the methodology. This is not described there it seems.
We will add to the methods.

116. Perhaps this should be put in the methodology and the bounded rationality argument used as a rationale for the limited size of the context window?
We think this works best in the discussion because it relies partly on results generated here, and so would be less clear earlier in the manuscript.

117. Reference missing.
Added.

118. I am not sure if I would personally consider these errors. I interpret them as inherent aspects of the model and assumptions made, parameterization etc. Perhaps you can call them limitations?
Yes, we can change this.

119. This is useful information for the methodology. It is not described there as such.
We will check that and the point below is covered in the methodology.

Page 18:

120. This is all methods that is not included in the methodology. It would be very helpful to

put it there, it is answering some of the questions that came up. When you put it in the methodology, please expand on it. For instance, what is the human-in-the-loop exactly?
Thanks. We will expand on this in the revised manuscript.

121. How exactly is this quantitative data translated into qualitative data by the agents?
Please see the responses to Comments 32 and 62.

122. Would it be possible to discuss the limitations of the modelling approach in terms of the limited number of policy targets? And in general the experimental set up and whether different dynamics between environmental and agricultural agents and lobbyists etc. would be expected when they are fed different policy targets (e.g., 1 billion tree planting instead of PAs). Do you think the same issues would arise?
In other words, following this assessment, is the model sufficiently tested to run other types of scenarios with other policy targets, other institutional agents etc.? Or are more tests and sensitivity analyses necessary?
We can briefly touch on this, but as initial proof-of-concept research, the modelling presented here is not really intended to establish all of these outcomes. Such work would require a range of specific inputs as well as targeted tests. We will discuss the generalisability of our findings, however.

123. I don't fully understand how the robustness is of the modeling approach is tested. There is no sensitivity analysis with different starting conditions and no comparison between scenarios etc. So I personally don't think you can speak of model robustness in the sense as it is a term usually implying some sensitivity analyses etc.. Additionally, the terms "error-proneness" and "error-tolerance" both imply some sort of quantification of the error by the model. In principle it seems that the dynamics that are called "errors" are inherent to the model structure and assumptions and the way agents retrieve information from text. Perhaps you can just refer to them as limitations or "unrealistic dynamics" or something like that?
Good suggestion, thanks - we think this will indeed help to clarify the work. We’ll check terminology throughout with these points in mind as well.
We’d like to clarify that robustness here refers to the ability of the institutional network model to function effectively despite errors, incomplete information, or unexpected disruptions. It is understood from the perspective of system science. A robust system maintains functionality even when agents produce erroneous outputs or misinterpret data. This type of robustness does not imply statistical robustness (which requires sensitivity analysis) but rather operational resilience, meaning that the model does not collapse or produce entirely unrealistic behaviours due to minor errors.

124. Didn't the environmental agents underestimate PA extents and wouldn't this have lead potentially to more power to influence budget allocation to increasing protected areas?
Underestimating PA coverage alone does not inherently grant environmental agents more power. Since actual PA coverage is also underestimated, the gap between policy goals and reality may not necessarily be larger. The influence of environmental agents depends on multiple factors, including the narratives they present, the counterarguments from competing agents, and how the high-level institution weighs inputs from other LLM agents.

125. I already wrote a comment on this before. It does not seem to me that the high-level institution is making balanced decisions, particularly not on the budget division part. Perhaps though I don't know what you mean by balanced in this context. Perhaps you could define what

balanced means to you?
In this context, "balanced decisions" do not imply the final budget surplus/deficit accumulated in different operational institutions. Instead, it means that the high-level institution avoided extreme reallocations because of the conflicting inputs from different stakeholders, which reflects incrementalism and path-dependent decision-making, common in real-world governance (Lindblom, 1959; Jones, 2003).

We have rephrased the sentence in the revised manuscript to improve clarity: The high-level institution’s tendency to seek compromise among competing policy priorities contributed to the error-tolerance of the institutional network.

126. I do not fully understand how you derive this conclusion about distrust in simulations.
This does not refer to distrust in simulations, but between policy actors; we are simply drawing a parallel to a real-world issue here.

127. Perhaps replace the arrow with writing?
Done.

128. This is why a definition of error is needed. In my opinion the LLM output is not correct nor incorrect (unless there is a bug or something fundamentally wrong with the simulations). The model simulates into the future, so it can never be correct in that general sense. Additionally, if "correct" means that agents never make mistakes, this would not be realistic.
In the context of the discussion here, “correct” means logically correct. It means the resultant policy decisions or actions are derived logically from the reasoning process conducted by the agents. In the revised manuscript, we have made this explicit.

Page 25:

129. There is quite some information in this section that could help clarify things for the reader in the main text. If possible, I would suggest to transfer some of this information so that that reader can more easily follow the methodology and interpret the results.
Please see the response to Comment 130.

Page 26:

130. Could you transfer this to the main text? I was missing information like this to interpret the results.
Thank you for noting this. This is very difficult to do without creating further confusion, as much of the text depends on other sections for meaning. We will cite the equations here explicitly in the revised manuscript for clearer guidance as per your suggestion in *Summary of the review*.

Page 27:

131. Perhaps it is possible to refer a bit more explicitly to these sections of the SI by stating for instance that you can find information about the policy implementation for PAs in this part of

the SI in the methodology. This would be very helpful because then the reader would know that the information is provided somewhere. Now this is not included in the experimental settings.
Good suggestion. Done.

Page 31:

132. References
Done.

Box 1. *How the stylised model and experiments map onto real-world policy institutions in the context of the European Union.*

The stylised model and experimental design presented here were inspired by the real-world mechanisms for policy delivery within the European Union (EU). Whilst the EU reflects a specific set of policy institutions and policy instruments, many of these concepts are transferable to other parts of the world with similar governance modes. In this box, we outline the relationships between the model components, especially the modelled agents, and their real-world counter-parts that are outlined in Fig. 1.

The executive body of the EU is the European Commission, which is responsible for enacting new legislation. This is equivalent to the *high-level institution* in the model. The European Commission proposes new Directives and other policy measures that are ratified by the European Parliament, but which are then implemented by the member states (national governments). National scale implementation is done through various government departments that are usually responsible for a specific policy sector, e.g. agriculture, environment, research, etc. Within the model, these government departments are represented by the *operational institutions*, and the multiple instances of the *operational institutions* represent the different policy sectors.

Beyond the basic mechanism for policy implementation described above, policy institutions are influenced by a number of external bodies. Within the model, these are the *lobbyists* and the *advisors*. In the European context, lobbyists could include land owner associations with responsibility for the economic well-being of their membership. They also include environmental Non-Governmental Organisations (NGOs) that lobby for stronger environmental protection. *Advisors* can include lawyers who support the legal aspects of policy development and implementation as well as scientific researchers who provide policy institutions with knowledge to support policy development (at least in principle). It should be noted that the European Commission, a *high-level institution,* is also a very large research funder, providing financial support for policy-relevant research in universities and other research organisations across the EU.

The stylized experiments applied to the model, also reflect a real-world context. Protected Areas (PAs) currently cover 26% of the EU's terrestrial land surface, and so represent a major land use by area. Likewise, the livestock sector in the EU uses large areas of pasture and cropland to provide livestock feed for meat and milk production. Consequently, livestock farming and nature conservation compete for finite land resources, and policy interventions are a major contribution to resolving this competition. Competition processes are central to land system models such as CRAFTY.

Terms and definitions within the context of this research	
Term	Definition
Institution	An organization or governing body involved in policy-making, such as government agencies, research institutions, or NGOs.
Institutional Dynamics	The interactions, adaptations, and power relations among these institutions over time, which influence how policies are formulated, negotiated, and implemented.
Institutional Network	A structured system of interconnected institutions that interact within the policy-making landscape.

Agent	A computational entity within the model that represents an institution, stakeholder, or decision-making body. Agents in institutional network are powered by LLMs and autonomously make decisions, process information, and interact with other agents. Agents in the CRAFTY land use model represent various types of land users.
Actor	A general term for entities (individuals or organizations) involved in decision-making processes. Actors can be both real-world stakeholders (e.g., policy-makers, lobbyists) and their simulated counterparts (LLM-powered agents).
Long-term/Short-term Memory	Memory components of an LLM-powered agent that influence its decision-making process. Short-term memory refers to immediately available contextual information embedded in the agent's prompt. Long-term memory is stored information retrieved when needed, using data retrieval techniques such as RAG, allowing agents to reference past information and knowledge base.
Tools	External functions or programming scripts that LLM agents can execute to complete tasks beyond text generation. Here, tools include Python scripts for data analysis, information retrieval, and numerical computations.
Tokens	The fundamental units of text processed by an LLM, representing words, characters, or sub-words. Token usage is a consideration in computational cost and efficiency when running LLM-powered agents. LLM providers normally charge API users by calculating input/output tokens.
Intermediate Output	Partial results generated by an LLM agent before reaching a final decision. The intermediate output allows agents to iterate on their reasoning, refine calculations, and update their responses before producing a definitive action.
Erroneous Behaviour	Any unintended or incorrect actions performed by an LLM agent. In the manuscript, erroneous behaviour includes misinterpretations of data, incorrect formatting, logical inconsistencies, or failures in workflow execution.
Gap	The difference between policy targets and actual outcomes. For example, in the manuscript, gaps exist between projected meat production levels and actual supply, or between budget allocations and their intended impact.
Believable Behaviour	The extent to which an agent's actions and decision-making processes resemble realistic human behaviour. Believable behaviour in the manuscript refers to whether the LLM agents mimic the constraints, trade-offs, and reasoning expected of real-world institutional actors.
Hallucination	The generation of plausible-sounding but factually incorrect or logically flawed responses by an LLM agent. In the manuscript, hallucinations occur when agents produce misleading data interpretations, misrepresent numerical values, or invent non-existent policies and regulations.
Robustness	Robustness refers to the ability of the institutional network model to function effectively despite errors, incomplete information, or unexpected disruptions. A robust system maintains functionality even when agents produce erroneous outputs or misinterpret data. This type of robustness does not imply statistical robustness (which requires sensitivity analysis) but rather operational resilience, meaning that the model does not collapse or produce entirely unrealistic behaviours due to minor errors.
Counter-intuitive behaviour	Counter-intuitive behaviour refers to an emergent pattern or decision-making outcome that deviates from conventional expectations or common sense. In the context of the manuscript, counter-intuitive behaviour occurs when the high-level institution's decisions do not align with typical policy-making norms, such as the unexpected budget allocation, which may contradict the assumption that the percentage of the total budget should be 100%. This behaviour does not necessarily indicate an error but rather an unanticipated outcome driven by the agents' decision-making.

References

Blanco González, V. (2017). Modelling adaptation strategies for Swedish forestry under climate and global change.

Daugbjerg, C., & Feindt, P. H. (2017). Post-exceptionalism in public policy: transforming food and agricultural policy. Journal of European Public Policy, 24(11), 1565-1584.

Davidson, M.R., Filatova, T., Peng, W., Verbeek, L. & Kucuksayacigil, F. (2024). Simulating institutional heterogeneity in sustainability science. Proceedings of the National Academy of Sciences, 121 (8), e2215674121.

Greer, A. (2017). Post-exceptional politics in agriculture: an examination of the 2013 CAP reform. Journal of European Public Policy, 24(11), 1585-1603.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., ... & He, Y. (2025).Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

Jones, B. D. (2003). Bounded rationality and political science: Lessons from public administration and public policy. Journal of Public Administration Research and Theory, 13(4), 395-412.

Levy, M., Jacoby, A., & Goldberg, Y. (2024). Same task, more tokens: the impact of input length on the reasoning performance of large language models. arXiv preprint arXiv:2402.14848.

Lindblom, C. E. (1959). The science of "muddling through." Public Administration Review, 19(2), 79-88.

Matthews, A. (2013). Greening agricultural payments in the EU's Common Agricultural Policy. Bio-based and Applied Economics, 2(1), 1-27.

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F.(2024). Cultural bias and cultural alignment of large language models. PNAS nexus, 3 (9), 346.

Taubenfeld, A., Dover, Y., Reichart, R., & Goldstein, A. (2024) Systematic biases in LLM simulations of debates. arXiv preprint, arXiv:2402.04049.

Watts, K., Whytock, R.C., Park, K.J., Fuentes-Montemayor, E., Macgregor, N.A., Duffield, S., McGowan, P.J.K. (2020). Ecological time lags and the journey towards conservation success. Nature Ecology & Evolution 4, 304-311.

Zhou, H., Feng, Z., Zhu, Z., Qian, J., and Mao, K. (2024).UniBias: Unveiling and Mitigating LLM Bias through Internal Attention and FFN Manipulation, arXiv preprint, arXiv:2405.20612.

https://doi.org/10.5194/egusphere-2024-2661
Preprint. Discussion started: 21 October 2024

InsNet-CRAFTY v1.0: Integrating institutional network dynamics powered by large language models with land use change simulation

Yongchao Zeng[1], Calum Brown [1,2], Mohamed Byari[1☆], Joanna Raymond[1☆],Thomas Schmitt[1☆], Mark Rounsevell[1,3,4]

[1] Institute of Meteorology and Climate Research, Atmospheric Environmental Research (IMK-IFU), Karlsruhe Institute of Technology, 82467 Garmisch-Partenkirchen, Germany

[2] Highlands Rewilding Limited, The Old School House, Bunloit, Drumnadrochit IV63 6XG, UK

[3] Institute of Geography and Geo-ecology, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany

[4] School of Geosciences, University of Edinburgh, Drummond Street, Edinburgh EH8 9XP, UK

☆ These authors contributed equally to this work.

Correspondence: Yongchao Zeng (yongchao.zeng@kit.edu)

Abstract: 1 Understanding and modelling environmental policy interventions can contribute to sustainable land use and management but is challenging because of the complex interactions among various decision-making actors. Key challenges include endowing modelled actors with autonomy, accurately representing their relational network structures, and managing the often-unstructured information exchange. Large language models (LLMs) offer new ways to address these challenges through the development of agents that are capable of mimicking reasoning, reflection, planning, and action. We present InsNet-CRAFTY (Institutional Network – Competition for Resources between Agent Functional Types) v1.0, a multi-LLM-agent model with a polycentric institutional framework coupled with an agent-based land system model. The numerical experiments simulate two competing policy priorities: increasing meat production versus expanding protected areas for nature conservation. The model includes a high-level policy-making institution, two lobbyist organizations, two operational institutions, and two advisory agents. Our findings indicate 2 that while the high-level institution tends to avoid extreme budget imbalances and adopts incremental policy goals for the operational institutions, it leaves a budget deficit in one institution and a surplus in another unresolved. This is due to the competing influence of multiple stakeholders, which leads to the emergence of a path-dependent decision-making approach. Despite errors in information and behaviours by the LLM agents, the network maintains overall behavioural believability, demonstrating 3 error tolerance. 4 The results point to both the capabilities and challenges of using LLM agents to simulate policy decision-making processes of bounded rational human actors and complex institutional dynamics, such as LLM agents' high flexibility and autonomy, alongside the complicatedness of agent workflow design and reliability in coupling with existing programmed land use systems. These insights contribute to advancing land system modelling and the broader field of institutional analysis, providing new tools and methodologies for researchers and policy-makers.

https://doi.org/10.5194/egusphere-2024-2661
Preprint. Discussion started: 21 October 2024

1. Introduction

Scientists have developed various models to study land systems, given their critical role in exploring key topics such as climate mitigation pathways (Duffy et al., 2022), carbon storage (Ekholm et al., 2024), human fire use (Perkins et al., 2024), and land cover change (Calvin et al., 2022; Chen et al., 2019). Land systems encompass both natural and human factors, with policy interventions playing a pivotal role in shaping land use dynamics. These interventions serve as critical mechanisms for addressing climate change, preserving biodiversity, and ensuring food security (Broussard et al., 2023; Guo et al., 2024; Qi et al., 2018). The formation and implementation of land use policies are the product of complex institutional dynamics and can involve a wide range of actors with differing objectives and powers (Davidson et al., 2024), as well as multi-level governance systems, such as that of the European Union (EU) (González, 2016). Understanding how these actors interact and public policies evolve is crucial for assessing how changes in policy can influence the land system in the future, and what this can mean for environmental goals.

Despite the importance of being able to simulate the effects of institutional dynamics on land systems, and despite ample empirical evidence highlighting interconnectivity among institutional actors (Ariti et al., 2019; Díez-Echavarría et al., 2023; Tesfaye et al., 2024), there is a scarcity of land use models which incorporate institutional networks, due to the challenges of representing heterogeneous, autonomous institutional decision-makers. Among the few studies that have explicitly modelled institutional actors within the land system are González (2016) and Holzhauer et al. (2019). In these examples, institutional agents are rule-based and programmed to take limited actions in response to specific land use changes. To strengthen the connection between modelling and real-world policy-makers, Zeng et al. (2024b) developed an endogenous institutional model using a fuzzy logic controller mechanism that can integrate real-world policy-makers' knowledge as IF-THEN rules. Other studies employ the network of action situation (NAS) approach (Kimmich et al., 2023), which is developed from action situation and game theory (McGinnis, 2011), allowing for systematic integration of a wide range of empirical evidence. However, NAS is still in its infancy (Tan et al., 2023), and it does not yet offer an approach to create formalized models.

These studies have advantages in modelling specific aspects of policy institutions. However, we contend that advancing the holistic representation of institutional actors in formal models needs to overcome three key challenges: agent autonomy, complex relational structures, and unstructured data. Firstly, modelling institutional actors' autonomy requires accounting for heterogenous behaviour (Dakin and Ryder, 2020), involving learning and memory (Nair and Howlett, 2017) together with bounded rationality (Jones, 2003; Simon, 1972). Secondly, there are both horizontal and hierarchical structures in the policy-making process, which can result in complex relationships between institutional actors and a lack of clarity in the policy formulation process (Cairney et al., 2019). For example, within the EU, there are multiple scales and layers of governance and authority, existing alongside NGOs and lobbyists (González, 2016). Thirdly, modelling institutional networks is confounded by the unstructured nature of the data that are available to policy actors (Lawrence et al., 2014). Data can be textual, and come in diverse formats, such as policy documents, grey literature, and research papers, which require institutional actors to understand natural languages including technical language. These challenges are not unique to this field; the simulation of human behaviour or ecological dynamics in the land system is similarly complicated, and solutions applied in these cases might be relevant here. Another similarity is in the value of such solutions, which cannot render a complicated system fully predictable but can reveal important dynamics stemming from behavioural processes (Davidson et al., 2024).

Large Language Models (LLMs), a form of artificial intelligence (AI), are based on numerous parameters that have been pre-trained on massive textual data and are designed to conduct natural language processes to understand and generate human-like text. The transformer architecture based on neural networks enables the LLMs to process sequences of text and contextual relationships between words (Vaswani et al., 2017). The text that LLMs produce is usually broken down into tokens, representing characters, sub-words or words (Minaee et al., 2024). LLMs have demonstrated strong language understanding and generation abilities and have emergent abilities such as multi-step reasoning that breaks down complex tasks into intermediate reasoning steps (Minaee et al., 2024). Hence, LLMs can be a powerful cognitive engine for autonomous agents that are able to sense the environment and act with regard to their own prescribed agenda (Wang et al., 2024). LLM agents' ability to process and understand natural language allows them to synthesize information from various sources including unstructured data.

LLM agents provide high flexibility in modelling complex interactions between multiple decision-makers. Park et al. (2023) simulated an artificial village with 25 villagers powered by LLMs. The simulated villagers had heterogeneous persona's and could interact with one another and their environment. These artificial villagers displayed believable, human-like behaviour and were able to organize a Valentine's Day party proposed by a user-controlled villager agent. Similarly, Qian et al. (2024) used LLM agents to simulate different roles in a software development team that is able to produce software cooperatively via a waterfall model. Further frameworks for dealing with many interacting agents have been emerging (see e.g., AutoGPT (Yang et al., 2023), AutoGen (Wu et al., 2023), AgentLite (Liu et al., 2024), MetaGPT (Hong et al., 2023)), which indicate the power of LLM agents in modelling complex relationships.

The aim of this study is to present a newly developed model InsNet-CRAFTY and[15] explore the potential of modelling institutional networks in the land system using a state-of-the-art LLM agent approach. First, we identify the conceptual framework for implementing the institutional model and its coupling with a land use model. Specific tasks are assigned to the institutional agents to facilitate the interpretation and evaluation of the model. We analyse the agents' textual output and numerical output to evaluate the believability of their decisions and the resultant performance of their actions. We identify both opportunities and challenges for LLM agent applications in modelling institutional networks within the land system, which may provide useful insights into both model conceptualization and implementation for future research. This study also contributes to the broader field of institutional analysis in socio-ecological modelling, offering novel tools and methodologies for researchers and policy-makers.

2. Methodology

2.1 Model framework of InsNet-CRAFTY v1.0

We adopt[16] the conceptual framework of a stylized, polycentric institutional network from González (2016), which offers a generic framework based on empirical evidence (e.g., peer-reviewed and grey literature) for Swedish forestry institutional actors. The key decision-makers included in the

https://doi.org/10.5194/egusphere-2024-2661
Preprint. Discussion started: 21 October 2024

conceptual framework are the government, research suppliers, environmental NGOs, (forest) owner associations, and supranational institutions. The government has three levels, namely national, regional, and local authorities. González (2016)'s framework features both hierarchical and horizontal structures, offering rich components of a polycentric institutional structure while maintaining parsimony for computational modelling.

We further adapt González's (2016) framework through generalisation and abstraction to obtain the conceptual framework for this analysis (see Fig. 1). The framework maintains González's (2016) structural features, but the hierarchical governments are abstracted into two layers with one comprising a high-level institution and the other several independent operational institutions (representing different policy sectors) leading to greater governmental polycentrism. Additionally, two new agents are included - a law consultant and a narrative injector. A description of all of the LLM agents follows here.

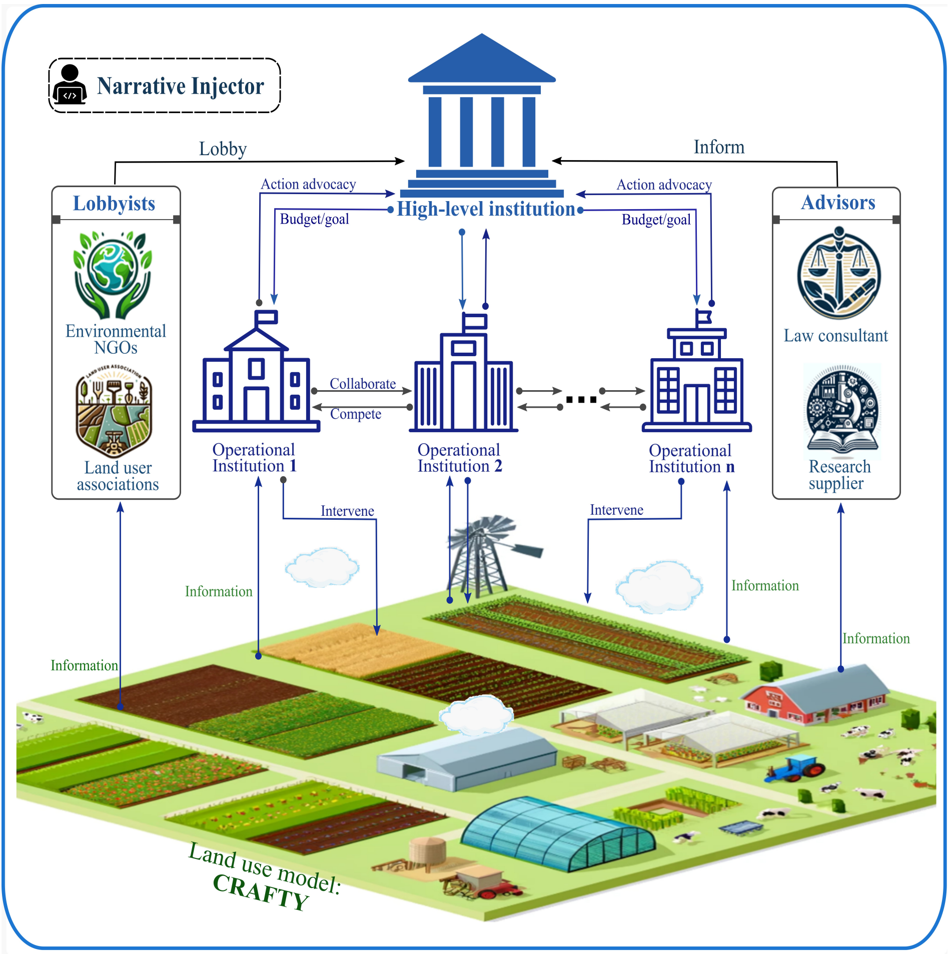


Figure 1: Conceptual framework for InsNet-CRAFTY v1.0. The institutional network model is adapted from Gonzalez et al. (2016) and coupled with the CRAFTY land use model (Brown et al., 2019). The hierarchical governments are abstracted into two layers with one comprising a

high-level institution and the other several independent operational institutions to achieve greater governmental polycentrism.

19

High-level institution: The high-level institution sets the overall policy ambitions and constraints (e.g. budgets) that affect the decisions of the operational institutions. The high-level institution aims to achieve mid to long-term policy goals based on the information provided by the operational institutions, research suppliers, lobbyists, law consultant, and narrative injector.

Operational institutions: Operational institutional agents represent different policy sectors (e.g., agriculture, nature conservation, forestry, transport). They adopt and execute concrete policy instruments to influence the decisions of land user agents in order to achieve specific policy goals. Operational institutions can also submit action advocacies to the high-level institution to obtain budgets or permissions for certain policy actions.

Lobbyists: Lobbyist agents represent professionals who advocate for specific interests or causes (e.g. environmental NGOs and land use associations). Lobbyists in the model can observe the state of the land use system and form their own opinions about what should be changed to reach their own objectives. Their advocacy is considered by the high-level institution when making policy adjustments.

Advisors: Advisory agents can inform the high-level institution's policy-making using professional knowledge and skills. The framework considers two types of advisors: research suppliers and law consultants. The research suppliers observe land use changes and provide a description of the current and possible future land use states. They analyse and interpret both numerical and textual data to support the high-level institution's decision-making. Law consultants offer information about existing laws, regulations, policies, etc., that legally underpin the high-level institution's policy actions; here we use EU policy documents to define these.

Narrative Injector (optional): An actor whose absence does not affect the functioning of an institutional network but can introduce highly unstructured exogenous disruptions into the model simulations through narratives (e.g., protest, war, unexpected disasters). The narratives can interact with all actors in the model and can be injected at any point during the simulation. The narrative injector provides the means to explore the impact of shock and extreme events on the functioning of the institutional model.

Together with these institutional agents, we apply the CRAFTY land use model (Brown et al., 2019; Murray-Rust et al., 2014) to simulate land use changes in response to the institutional agents' interventions and potentially other drivers of change, e.g. socio-economic and climate change. The LLM agents form a stylized polycentric institutional model that can be implemented in a sequential order. For instance, CRAFTY can produce information indicating that both meat supply and protected areas (PAs) need to be improved to achieve better food security and nature conservation. 20 Then, the research supplier, operational institutions, and lobbyists collect and analyse the relevant data generated from CRAFTY. The data analysis serves as a basis for these agents to form different narratives that fit their distinct roles. The law consultant references policy and law documents to extract relevant information. The narrative injector may output a piece of news about an emergent incident. All these agents' output is eventually fed to the high-level institution, which considers

https://doi.org/10.5194/egusphere-2024-2661
Preprint. Discussion started: 21 October 2024

the different stakeholders' positions and strives to make balanced decisions. The high-level institution has concrete actions to influence the behaviour of the operational institutions, such as budget allocations and policy goal adjustments. The operational institutions utilize their budgets to formulate policy instruments such as subsidies, taxes, and administrative measures to steer meat supply and PA coverage towards the target levels. It is worth noting that the high-level institution does not have to be activated at the same frequency (in time) as the operational institutions, reflecting the asynchronous nature of agent decision-making at different levels. Appendix A provides extra details and a technical description of the model's sequential processes.

2.2 LLM agent framework

To implement the institutional network model, the agents have to be equipped with a powerful "brain". Because of the extremely rapid evolution in the LLM field, a variety of ways to create LLM agent "brains" have been emerging (Sumers et al., 2024). Here we use the framework in Fig. 2 to represent the cognitive architecture of an LLM agent, which derives from the unified framework proposed by (Wang et al., 2024), and the LangChain framework (LangChain, 2024).

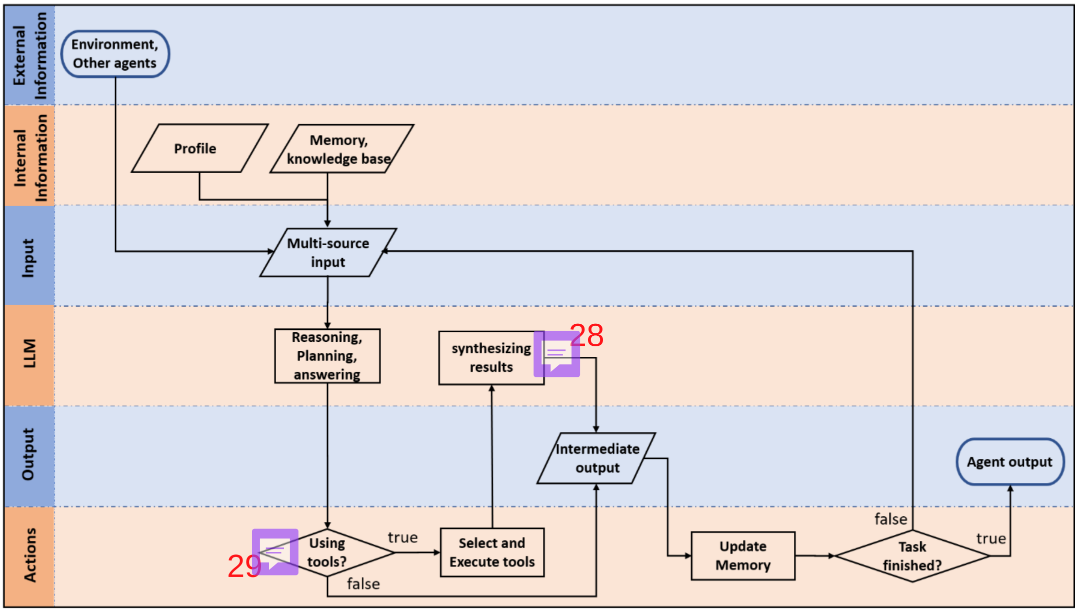


Figure 2: The cognitive architecture of an LLM agent. The core procedures of a LLM agent include the input, output, and the LLM. The agent's capability can be enhanced by integrating sophisticated work-flows such as memory, tool use, and reflection.

Although the complicatedness of different agents' "brains" varies, the core of a LLM agent consists of a LLM and the LLM's input and output. The functionality of the LLM agent can be enriched by incorporating more information into the input. Besides receiving external information from the modelled environment and other agents' responses, the LLM agent integrates internal

https://doi.org/10.5194/egusphere-2024-2661
Preprint. Discussion started: 21 October 2024

information such as a profile describing its identity, objectives, decision guidance, etc. An agent can also incorporate memory and a knowledge base into its input. Memory is divided into short-term memory and long-term memory. Short-term memory with high temporal relevance is embedded directly into the agent's prompt (the input directly received by an LLM). Knowledge and long-term memory relevant to a given decision-making context are extracted using Retrieval Augmented Generation (RAG) (Fan et al., 2024). This multi-source information forms the input to prompt the LLM to generate reasoning and planning, or to answer specific requests. If the agent is given a task to complete, the LLM helps to divide the task into small and achievable sub-tasks.

The capabilities of an LLM agent extend beyond text generation; it can actively execute sub-tasks and make decisions about the necessity of tools for task completion. In this context, "tools" refers to functions coded in programming languages such as Python. For instance, a function might perform calculations that current LLMs struggle to handle reliably on their own. An agent selects and employs appropriate tools to advance a task, as required. These tools process and organize results, which the LLM then synthesizes and outputs in natural language. Initially, these outputs are considered intermediate. The agent updates its memory by organizing and storing relevant inputs and outputs as necessary. Subsequently, it evaluates whether the tasks are complete to decide whether to produce the final output or to continue processing with updated memory.

2.3 Experimental Settings

The CRAFTY land use model is a crucial component of the simulation environment, within which the institutional agents operate. We set up the land use model according to CRAFTY-EU (Brown et al., 2019) and parametrized it with the data for the RCP2.6-SSP1 climatic and socio-economic scenario (Brown et al., 2019). The CRAFTY-EU model uses a map of European countries at a 5-arcminute resolution. The scenario simulation covers the period from 2016 to 2076. The data are available on Zenodo (Zeng et al., 2024c)

To enhance the focus of the experiments and facilitate the analysis, we narrowed the scope of the modelled actors by specifying their roles and responsibilities. Instead of integrating a diverse array of operational institutions with a wide range of policy objectives and tools, we incorporated two distinct operational institutions focused on different policy sectors: an environmental institution and an agricultural institution. The former prioritises environmental protection with a specific aim of expanding protected areas (PAs) for nature conservation, while the latter focuses on meat production to ensure food security using economic policy instruments such as subsidies. Since meat consumption is a major driver of deforestation, greenhouse gas emission, climate change, and biodiversity loss (Djekic, 2015; Machovina et al., 2015), and its consumption continues to increase (Petrovic et al., 2015), this experimental setting creates a conflicting context for the two institutions. They compete for limited budgets to fulfil their respective policy objectives.

Lobbyists actively seek to influence the high-level institution by advocating for increased financial support to either enhance the PAs or develop the meat industry. The research supplier analyses and interprets the data generated by the CRAFTY model; while the law consultant uses RAG to retrieve relevant information from a selected set of EU policies. The data are available on Zenodo (Zeng et al., 2024c). Consequently, the high-level institution is tasked with managing the interplay and potential conflicts between these agents, striving to balance budget allocations with the practical achievement of policy goals. The specific experimental purpose is to explore how these

https://doi.org/10.5194/egusphere-2024-2661
Preprint. Discussion started: 21 October 2024

agents reason and make plans in favour of their positions and to evaluate their performance in policy goal adjustments/achievements and budget allocation.

To improve the performance of the simulations, the lobbyists are allowed to use the output from the research supplier to strengthen their arguments. We chose not to incorporate the narrative injector agent in the results reported here for simplicity and in order to maintain the system's full autonomy. We followed an AI-assisted prompt development procedure depicted in Zeng et al. (2024a) and sought to use straightforward language to form the prompt templates. The prompt templates are given in Table B1 – B7.

As previously stated, the high-level institution and the operational institution are not synchronous. Here, the high-level institution is activated every ten iterations, while the operational institutions adjust their policies every two iterations, representing a more frequent response in policy adjustments. This frequent adjustment reflects the agility of the operational institutions compared to the slower, more deliberative pace of the high-level institution.

We set the initial target meat supply as 1.2 times the initial meat production level, and the target of PA coverage as 10% of the total land area. These parameters give the institutional actors slightly higher but achievable initial targets to pursue. The initial budget allocation is equally divided between the operational institutions.

To implement the model, we used a combination of different LLMs to power the agents to improve the token cost. The LLMs agents with actions were built using the LangChain library. The agents' features are summarized in Table 1. The equations that describe the high-level and operational institutions' non-LLM behaviour as well as related numerical settings can be found in Appendix C. The code is available on Zenodo (Zeng et al., 2024d)

Table 1: The experimental settings of the LLM agents

Agent	Input	Action	Output	LLM	Remarks
Law consultant	1)Document containing EU laws, policies, regulations etc. 2) Profile	Using RAG.	Unstructured text to inform the high-level institution's decision-making	Llama-3-70b-8192	Goals: Extracting relevant information from a knowledge base to inform the high-level institution's legal actions.
Research supplier	1) CSV file containing data from CRAFTY 2) Profile	Wiring and executing Python code to analyse the data.	1) Unstructured text to inform other agents 2) Intermediate output	gpt-4o	Goal: Analysing and interpreting the data generated by CRAFTY.

Environmental NGO	1) Research supplier's output 2) Profile	None	Unstructured text to lobby the high-level institution	Llama-3-70b-8192	Goal: Lobbying the high-level institution to prioritise nature conservation.
Land user associations	1) Research supplier's output 2) Profile	None	Unstructured text to lobby the high-level institution	Llama-3-70b-8192	Goal: Lobbying the high-level institution to prioritise meat industry development.
Environmental institution	1) CSV file containing data from CRAFTY 2) Profile	Wiring and executing Python code to analyse the data.	1) Unstructured text to inform the high-level institution 2) Data from CRAFTY code	Llama-3-70b-8192	Goal: Striving to acquire budget to support PA expansions to reach the target PA coverages. Policy instrument: PA designation.
Agricultural institution	1) CSV file containing data from CRAFTY 2) Profile	Wiring and executing Python code to analyse the data.	1) Unstructured text to inform the high-level institution 2) Data from CRAFTY code	Llama-3-70b-8192	Goal: Striving to acquire budget to support meat production to reach the target meat supply level. Policy instrument: economic measures (e.g., taxes and subsidies)
High-level institution	1)Unstructured text from all other agents 2) Profile	None	1) Unstructured text 2) Policy goal and budget allocation adjustments in JSON structure	gpt-4o	Goal: Making policy adjustments based on multiple stakeholders. Policy Instrument: Administrative orders to adjust the operational institutions' policy goals; financial measures to allocate budget between the operational institutions.

3. Results

3.1 Textual output

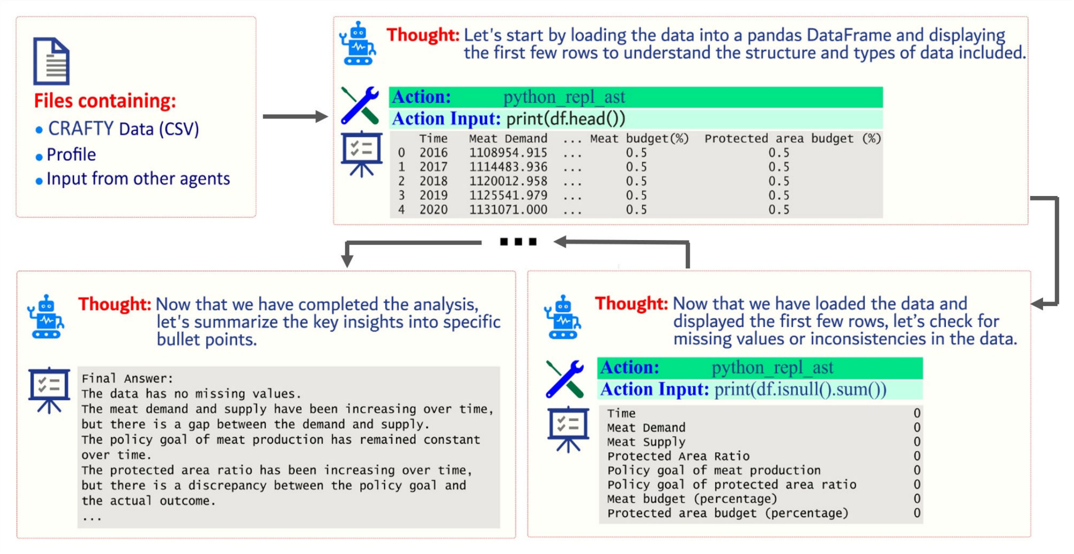
The LLM agents' output contained 19808 words (28778 tokens) and 48 plots. We summarised the textual output that demonstrates the behavioural patterns of the agents, while also highlighting counter-intuitive or potentially erroneous agent behaviours. This allows the agents' general behavioural regularities and occasional malfunctions to be displayed simultaneously.

3.1.1 The advisors' output

The research supplier completed various tasks, including checking missing values in the data, interpreting the trend of meat supply and demand, analysing the discrepancy between policy goals

https://doi.org/10.5194/egusphere-2024-2661
Preprint. Discussion started: 21 October 2024

and actual outcomes, budget allocation, and calculating the correlation between different time series. Figure 3 briefly illustrates the work-flow of the research supplier agent. It made plans, executed the plans step by step and interacted with tools. The tools offered returned values to form intermediate output, which was fed back to the LLM. A final output was produced based on the intermediate outputs. In some of the final outputs, the agent attached a note at the end of the output as a reminder of the applicable scope of the analysis, e.g., "Note: The above insights are based on the analysis of the provided data and may not be generalizable to other contexts". In 2066, the agent encountered an error -- "Agent stopped due to iteration limit or time limit" – indicating the actions the agent needed to take exceeded the set maximum execution time.



Thought: Let's start by loading the data into a pandas DataFrame and displaying the first few rows to understand the structure and types of data included.

Action: python_repl_ast
Action Input: print(df.head())

```
   Time   Meat Demand   ... Meat budget(%)  Protected area budget (%)
0  2016   1108954.915   ...       0.5                    0.5
1  2017   1114483.936   ...       0.5                    0.5
2  2018   1120012.958   ...       0.5                    0.5
3  2019   1125541.979   ...       0.5                    0.5
4  2020   1131071.000   ...       0.5                    0.5
```

Thought: Now that we have loaded the data and displayed the first few rows, let's check for missing values or inconsistencies in the data.

Action: python_repl_ast
Action Input: print(df.isnull().sum())

```
Time                                   0
Meat Demand                            0
Meat Supply                            0
Protected Area Ratio                   0
Policy goal of meat production         0
Policy goal of protected area ratio    0
Meat budget (percentage)               0
Protected area budget (percentage)     0
```

Thought: Now that we have completed the analysis, let's summarize the key insights into specific bullet points.

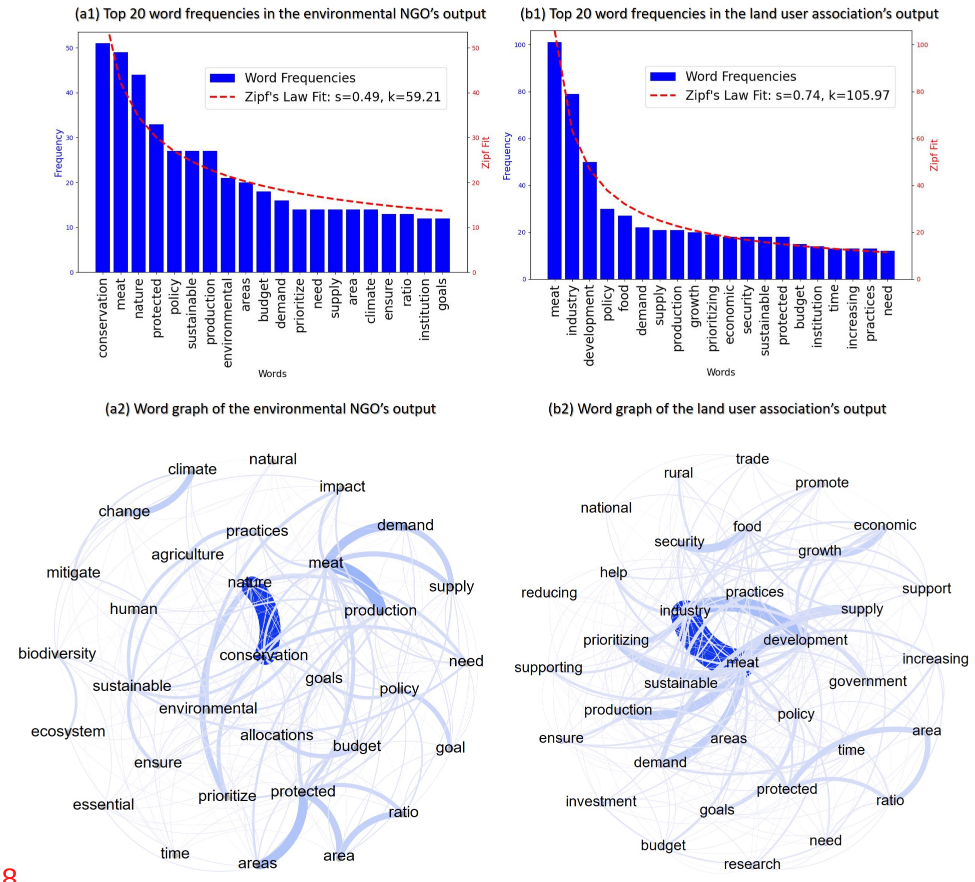
```
Final Answer:
The data has no missing values.
The meat demand and supply have been increasing over time,
but there is a gap between the demand and supply.
The policy goal of meat production has remained constant
over time.
The protected area ratio has been increasing over time,
but there is a discrepancy between the policy goal and
the actual outcome.
...
```

Figure 3: The work-flow of the research supplier agent. The agent took the initial input to generate a thought to decide what actions should be taken to analyse the data. Then, it executed the action by calling a function, which in turn produced the intermediate results. These results served as a part of the updated input to let the agent generate a new thought for the next iteration. After several iterations of thought-action-output loops, the research supplier agent produced a final interpretation of the data.

The law consultant emphasized six critical aspects to influence the decision-making of the high-level institution, based on the available knowledge base. These aspects include "biodiversity and ecosystem restoration targets," "agricultural production and environmental impact," and "climate change mitigation." The agent not only highlighted these issues but also cited relevant laws, policies, and directives. Furthermore, the agent elaborated on the implications of these legal and policy frameworks for the high-level institution's policy-making processes. For example, in discussing "biodiversity and ecosystem restoration targets," the law consultant noted that "the EU Restoration Law mandates the restoration of at least 20% of the Union's terrestrial and marine areas by 2030, and all ecosystems in need of restoration by 2050." This law was interpreted to mean that "a significant portion of the budget should be dedicated to protected areas to meet these objectives." It was observed that the law consultant agent produced the same output repeatedly over several iterations, reflecting stagnation due to the absence of new inputs that could prompt different responses. 76

3.1.2 The lobbyists' output

The environmental NGO generated a variety of arguments for prioritising protected area establishment over meat production. For instance, in some years the agent highlighted the urgent need for nature conservation, the impact of meat production, or the necessity of 77 budget increase. In 2066, the environmental NGO agent did not receive information from the research supplier due to the error mentioned above. However, this error did not paralyze the simulation. Instead, the LLM agent stated "I apologize, but it seems like there is no information provided. However, as a representative of an environmental NGO, I can still provide some general bullet points to lobby a high-level public policy institution to prioritise nature conservation". Without basing its arguments on data, the agent emphasized the economic benefits of nature conservation, the importance of PAs to climate change mitigation and adaptation, human health and well-being.


(a1) Top 20 word frequencies in the environmental NGO's output

(b1) Top 20 word frequencies in the land user association's output

(a2) Word graph of the environmental NGO's output

(b2) Word graph of the land user association's output

78 Figure 4: Word frequencies and word graphs derived from the lobbyists' output. The dashed red lines in (a1) and (b1) are derived by fitting 79 Zipf's law distribution to the word frequency distributions. Zipf's law can be expressed as $f(r) = k/r^s$, where f(r) is the frequency of a word;

r represents the rank of the word according to its frequency; k and s are parameters. A larger s indicates a set of words distributed more unevenly. It can be seen in (a1) $s = 0.49$ for the environmental NGO's output and in (b1) $s = 0.74$ for the land user association's output, reflecting the two agents' different approaches in formulating their arguments. The word graphs only display nodes that have more than thirty links, in order to maintain visual clarity.

The land user association agent also utilised background information and the data interpretation provided by the research supplier agent to lobby the high-level institution to prioritise meat industry development. For instance, this agent highlighted economic growth, job creation, food security, and alignment with policy goals. When the output from the research supplier agent was missing, it gave more general bullet points to lobby the high-level public policy institution, including emphasizing the meat industry's economic benefits, food security, rural development, innovation and technology without using any data from CRAFTY.

80 The lobbyists had high autonomy to defend their interests but were not given detailed instructions about how to persuade the high-level institution. To better visualize how the lobbyists formulate their arguments, Fig. 4 illustrates the word frequencies and relationships through word graphs derived from their outputs. The analysis reveals a less prominent skew in the frequency distribution of the top 20 words used by the environmental NGO compared to those of the land user association. This can be quantified by the parameters of Zipf's law distributions fitted to the word frequency data. The environmental NGO frequently emphasized the term "conservation" and notably the word "meat." Its discourse primarily focused on two aspects: the environmental threats posed by meat production and the critical importance of conservation efforts. This concern was underscored by the research supplier's data interpretation showing a widening gap between meat demand and supply. In contrast, the land user association highlighted the development of the meat industry and food security, without opposing the expansion of protected areas. Instead, the land user association consistently advocated for sustainable meat production practices, which they argued would support their request for an increased budget. 81

3.1.3 The operational institutions' output

82 The agricultural institution's outputs consistently addressed the discrepancies between the meat production policy goals and the actual outputs, alongside recurring budget challenges. This agent repeatedly emphasized the necessity of addressing budget deficits, advocating for more efficient budget allocations and increased financial support to meet production goals. Key recommendations included increasing budget allocations to bridge the gap between policy goals and actual outcomes, setting realistic policy goals that align with current capacities, and enhancing sector productivity through various initiatives, e.g., farmer incentives and sustainable practices. Additionally, the institution suggested establishing a robust monitoring and evaluation framework to regularly assess the effectiveness of policies and adjust as necessary. A holistic approach was advocated to balance increased production goals with budget constraints, thereby boosting food security, improving farmer livelihoods, and ensuring financial well-being.

The environmental institution consistently highlighted a gap between the current state of protected areas and policy goals over the decades, emphasizing the need for increased financial support and a higher priority for protected area establishment to achieve biodiversity conservation and pollution reduction. Recommendations include raising the PA goals incrementally each year,

https://doi.org/10.5194/egusphere-2024-2661
Preprint. Discussion started: 21 October 2024

improving governance, enhancing community engagement, and specifically allocating a substantial percentage of budget surpluses to facilitate the expansion of PAs. 83 These steps were deemed crucial by this agent for reaching Net-zero targets and effectively managing biodiversity conservation amidst evolving environmental challenges. 84 However, the agent mistakenly used mean values to describe the time series, which generated misleading outcomes. 85 For instance, in the year 2076, the actual protected area is 25.14% and the target is 30.17%; however, the environmental institution used the mean values of 13.44% and 17.40% respectively to inform the high-level institution about the current situation. This error did not, however, qualitatively change the need to expand protected areas. 86

3.1.4 The high-level institution's output

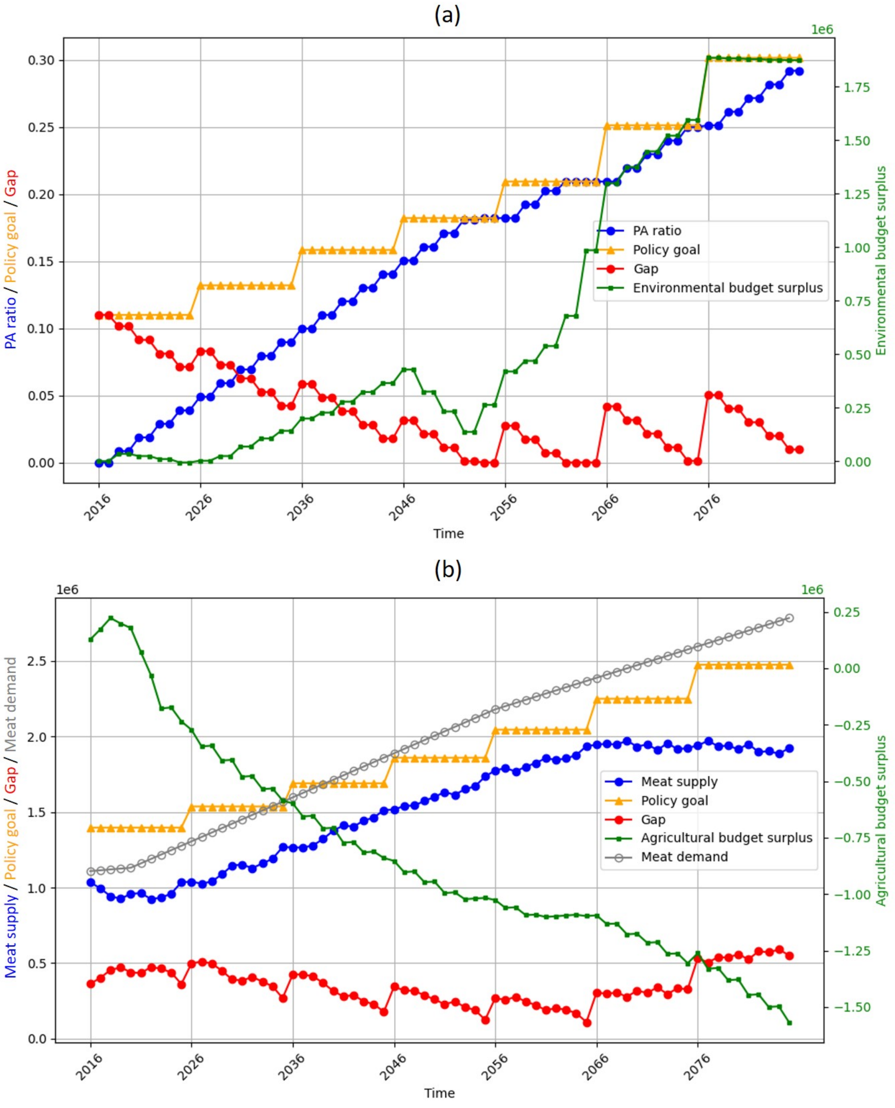
The high-level institution employed a systematic and analytical approach to decision-making, consistently integrating stakeholder feedback across several sectors to refine policy goals and allocate budgets effectively. 87 This process involves a detailed analysis of input from agricultural and environmental institutions, NGOs, and industry associations. Key actions include adjusting policy goals and redistributing budget percentages to better support the targeted outcomes in meat production and environmental protection. The institution regularly adjusted its strategies, intending to bridge the gaps between current outcomes and policy objectives, focusing on sustainability, economic stability, and nature conservation. However, the output of the high-level institution was sometimes inaccurate. 88 For instance, the high-level institution's analysis only included information from all six of the LLM agents in 2036 and 2056 with the law consultant and/or the research supplier's inputs occasionally being missed.

3.2 Policy actions and outcomes 89

The results shown in Fig. 5 (a) illustrate that the high-level institution increased the policy goals of the PA coverage gradually across the simulated time period, which resulted in a stepped pattern of 90 PA growth. This reflects the periodic activation of the high-level institution as described previously. The actual PA coverage seemed to be well controlled by the operational institution because the actual PA coverage shows a prominent 91 tendency to approach the target PA coverage. The eventual policy goal was set at approximately 30%, which drove the actual PA coverage to approach this level. In some years (between 2046 and 2076), the actual PA coverage reached the target and then remained almost unchanged for several years until the high-level institution raised the targets again.

92 The variation in the gap between the target PA coverage and the actual coverage illustrates the environmental institution's tendency to follow the policy target. In the beginning, the gap between the target PA coverage and the actual PA coverage was large, but the gap shrank over time until in 2052 the gap diminished to approximately zero. 93 Then the actual PA coverage stayed almost unchanged from 2052 to 2055, which indicates that the operational institution imposed negligible influence on the land use system to maintain a small target-outcome gap. In 2056, 94 the high-level institution raised the policy target to form a notable gap again. The environmental operational institution continued expanding the PAs to minimize this gap. This pattern repeated and resulted in the stepwise shape of the time series of the gap. These results demonstrate the alignment of the high-level institution's policy goal adjustments with the environmental institution's capability to influence PA coverage. However, the budget surplus remained positive and grew over time, which indicated over-funding by the high-level institution.

Similar to the policy goal adjustments in the PA coverage, Fig. 5(b) shows that the high-level institution increased the target level of meat production periodically and gradually, resulting in a stepped growth over time. Meat supply is positively correlated with the policy goal. Although the meat supply was not able to reach the policy goal, the goal-supply gap was limited within a relatively small range. In 2065 meat supply plateaued, while the ensuing policy goal adjustment at 2075 was still increasing. In contrast with the environmental institution, the agricultural institution underwent increasingly severe budget restrictions.



95

Figure 5: Policy goal adjustments, budget allocation, and their impacts for a) the environmental operational institution agent and b) the agricultural operational institution agent.

Figure 6 shows the budget allocated by the high-level institution. In the first ten years (from 2016 to 2025), the budget allocation between the two operational institutions is 50/50 by default. However, it can be seen that the high-level institution shows a tendency to avoid imbalanced budget allocation. Despite the agricultural institution's lack of budget, the budget allocated between these two operational institutions was 60/40 from 2026 to 2045, 30/40 from 2046 to 2055, and 45/55 from 2076 to the end of the simulation. The high-level institution chose to allocate more budget to the agricultural institution in only twenty iterations even if the latter's budget deficit occurred very early (before 2026).

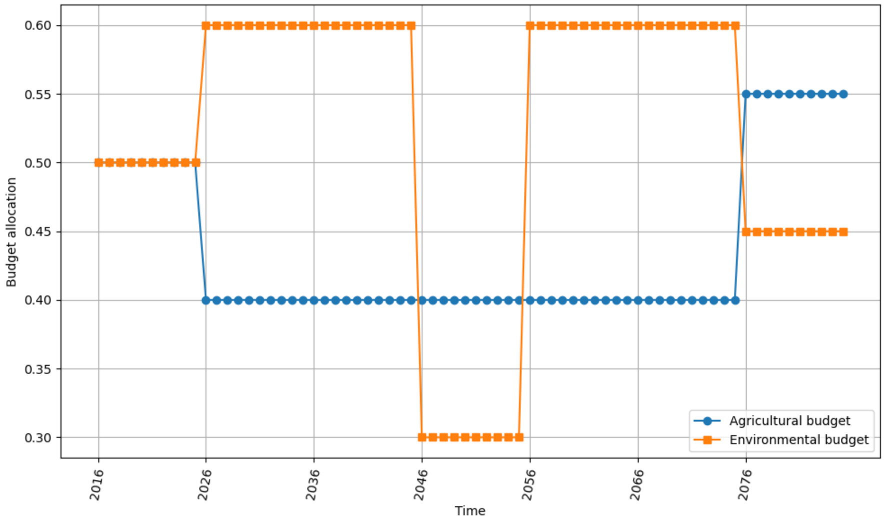


Figure 6: Budget allocation between the agricultural and environmental operational institution agents.

The twenty iterations include an unexpected budget drop for the environmental institution in 2046, 96 leading the sum of the budget allocation ratio to be 0.7, when it should be 1.0 as in the other iterations. 97 This sudden drop in budget is caused by the high-level institution's decision to allocate 30% of the total budget to "other initiatives and programs". This should result from the research supplier stating that "There are correlations between meat demand, meat supply, and protected area ratio, but it is not clear what the causative factors are". The research supplier's statement prompted the land user association to propose "further research is needed to understand the causative factors. We propose funding for research initiatives to better understand these relationships and inform evidence-based policy decisions". Although this unexpected drop in the budget allocated to the environmental institution demonstrates the coherence of information transmitted among multiple agents, it could also become an intractable issue for LLMs embedded within existing hard-coded systems.

4. Discussion

98

4.1 Believable behaviour of the LLM agents

Building upon the pre-trained LLMs, the institutional agents modelled in InsNet-CRAFTY exhibited diverse human-like reasoning and actions. The agents' behaviour was guided using prompts in natural languages, which gave the model developers high flexibility in creating the agents' autonomous behaviours. This flexibility facilitated the modelling of the complex relational structures among the agents. Given appropriate profiles, the agents were clear about their identities and relationships with others, demonstrating consistent role-specific decision-making. The LLM agents also showed an ability to handle the qualitative and unstructured information generated by the lobbyists, law consultant, and advocacies from the operational institutions. The capability of function calling (e.g. writing and executing computer code) further improved the agents' autonomy, enabling the latter to deal with more complex tasks such as data analysis and knowledge retrieval. These capabilities suggest that LLM agents have a unique potential to overcome the key challenges in modelling institutional networks.

Besides these apparent strengths, the LLM agents showed conflicting but understandable behaviour when faced with key real-world challenges such as conflicting objectives, uncertainty and imperfect (or absent) information. The budget allocation was a major output of the high-level institution, which reflected competing claims for a limited resource, as well as an impromptu suggestion by one of the lobbyist agents that money be transferred to research to better understand policy impacts. These dynamics could allow many important policy processes to be investigated, including observed differences between budgeting systems based on plurality, proportional representation or public participation, in which information inputs and decision-making powers vary substantially (Feindt, 2010; Hallerberg and Von Hagen, 1997; Lee et al., 2022).

The numerical results show that the environmental institution in our simulations was over-funded, while the agricultural institutions were struggling with an inadequate budget. We found that this imbalance was prompted by the environmental institution and the research supplier misleadingly informing the high-level institution that PA coverage was positively correlated with budget surplus. Indeed, both the two operational institutions insisted that their respective policy targets (PA coverage and meat production) should be increased because those targets were positively correlated with other desired outcomes. However, advocacy efforts were not equally effective. The environmental NGO's arguments were backed up by urgent environmental concerns, and outweighed the more economically focused arguments of the land user association. This imbalance might derive from the LLMs' training data being influenced by present-day social norms and highlights the potential for biases to be embedded within the agents' roles. These also of course reflect policy biases in reality, where norms, power relations, communication and urgency all affect policy priorities, and potentially allow exploration of approaches to mitigate these issues in differing policy contexts (Barnett and Finnemore, 1999; Sinden, 2004; Yami et al., 2019).

It can be hypothesised that the superficial use of correlation methods to interpret the data, the misleading arguments formed by the operational institutions, and the competition between the lobbyists all contributed to the high-level institution's path-dependent decision-making in both policy target adjustments and budget allocations. The lack of decisive action to fix the issue also implies that it is very challenging for the high-level institution to find an optimal solution given the need to consider many stakeholders' conflicting interests in the policy-making process. Nevertheless, such limitations do not necessarily diminish the LLM agent's usefulness in simulating human decision-making, rather it captures important and believable behaviour in terms of bounded rationality (Simon, 1972; Gigerenzer and Goldstein, 1996; Jones, 2003) and policy-makers muddling through (Lindblom, 1989) within complex systems.

4.2 Challenges of implementing LLM agents

Along with the advantages of the LLM agent approach in simulating institutional networks, erroneous behaviour was also apparent. Typical causes of errors were flaws in agent work-flows and LLM hallucinations (Ji et al., 2023; Perković et al., 2024; Yao et al., 2023). The research supplier agent's occasional failure to output data interpretation was caused by a flaw in the agent work-flow that generated the error "Agent stopped due to iteration limit or time limit.". This error could easily be avoided by increasing the permitted number of iterations that an agent needs to execute a complex task, although it had the advantage of forcing the other agents to proceed with imperfect (out-of-date) information, as is often the case in real-world contexts.

Unlike the research supplier, the operational institutions were not given specific data analysis instructions. This led to an unexpected outcome -- the operational institutions tended to use mean values to describe the latest state of the time series and so provided misleading information to the high-level institution. Such erroneous behaviour can be categorized as hallucination because the agent used plausible-sounding words to express nonsensical information (Ji et al., 2023). This erroneous behaviour could be mitigated by using more specified instructions in the prompts to guide agents' reasoning or designing extra mechanisms to detect and rectify the LLM's response (Tonmoy et al., 2024). However, addressing LLM hallucination is challenging, and there is no standard solution so far.

For large-scale, land use models, another crucial challenge is an LLM's limited context window. Here, the high-level institution had to consider all the other agents' output to make decisions. The resultant input could be very lengthy. Issues might arise if the input exceeds the maximum number of tokens (namely the size of the context window) that an LLM could digest. In the real world, institutional networks are far more complex than those in this model, and it is not unusual for high-level institutions to be overwhelmed by the information that they need to assimilate, or to use information selectively as a result (Bainbridge et al., 2022; Fischer et al., 2008; Rich, 1975). The limited size of the context window can therefore be seen as a feature that reflects human decision-makers' bounded rationality and information processing capabilities or preferences, as well as the imperfect nature of much information used for decision-making (Neri and Ropele, 2012). However, whether it is preferable to model such a feature in a controlled manner or to rely on the result of a technical limitation is debatable. The technical limitation could be mitigated by using summarized input or including memory mechanisms with retrieval methods (Modarressi et al., 2023; Zhong et al., 2024; Zhou et al., 2023), although these approaches all require extra effort in designing peripheral agent work-flows.

In contrast to the above errors, the data formatting issue could be more cumbersome to handle. Because the LLM agents were coupled with a programmed land use model, the LLM agents needed the capability to structure data in a designated format, otherwise, the programmed model would not parse the data, which could disrupt the simulations. Here, we used JSON, on which many current LLMs have been fine-tuned, to format the output of the high-level institution. However, there is no guarantee that LLMs can always accurately format their output. This leaves an extra task to design peripheral work-flows to secure the format. In this model, we employed three layers

https://doi.org/10.5194/egusphere-2024-2661
Preprint. Discussion started: 21 October 2024

of mechanisms to derive the correctly formatted data, including re-prompting the LLM, using regular expressions (Li et al., 2008) to identify JSON structure, and human-in-the-loop (HIL) correction (Zeng et al., 2024a).

4.3 Paradoxical robustness

The erroneous behaviour of the LLM agents could affect the robustness of the model and the approach used. However, the results implied a paradoxical relationship between the LLM-based institutional network model's error-proneness and error-tolerance, which could enhance the understanding of the robustness of multi-agent systems. For instance, with multiple institutional actors joining the system, the chances of erroneous behaviour increase since every single decision-maker has some probability of producing errors. These errors could also be transmitted within the network and affect other agents' decision-making, which, to some extent, corresponds to real-world policy-making. However, with the interaction of multiple agents, no single agent nor their erroneous behaviour had sufficient influence to determine the behaviour of the whole system. The missing output from the research supplier did not lead the system to generate a cascade of unusual behaviour nor did it crash the simulation. The high-level institution's tendency to make balanced decisions also contributed to the error-tolerance of the institutional network. The high-level institution's path-dependent decision-making ensures that the whole system is unlikely to adjust policies drastically. Hence, the incrementalism that derives from the polycentric institutional network structure may help to avoid critical policy failures, which is particularly important in the land system. This could also help to simulate the consequences of widespread distrust between policy actors in large networks (Fischer et al., 2016).

4.4 Contextual coherence does not equal logical consistency

While the agents' performance may reflect certain real-world phenomena within institutional networks, it is essential to address a deeper reflection on the current working mechanisms of LLMs. LLMs are designed to optimize literal contextual coherence, meaning that a vast amount of high-quality, textual data enables the machine to effectively mimic human language by approximating patterns of word (or token) changes (Radford et al., 2018). This creates the illusion that LLMs can think. However, it is crucial to recognize that although logical reasoning when expressed in a language can lead to contextual coherence, the reverse is not necessarily true. In other words, contextual coherence in text might be a necessary condition, but it is not sufficient to produce logical consistency. This raises a caveat: over-anthropomorphizing LLM agents can complicate the evaluation of their outputs. This difficulty arises from both the laborious manual work required to assess the agents' logical consistencies and the logical inconsistencies masked by contextual coherence. In future research, LLMs could be trained using "very strict" language that satisfies the condition that contextual coherence $\Longrightarrow$ logical consistency, which could ensure that the LLM output is correct. This would be a significant development for LLM-based simulations.

5. Conclusion

We explored the development and application of InsNet-CRAFTY v1.0, a multi-LLM-agent institutional network model with a polycentric structure that is coupled with an agent-based, land system model. By leveraging LLMs to facilitate interactions through textual data, the model enables each modelled entity to pursue unique goals and values that collectively impact the modelled land use system. The results demonstrate that this LLM-enhanced approach is powerful and flexible in modelling institutional actors' behaviours within the land system. However, this novel approach also brings new challenges arising from the limitations of current LLM technology, signifying the need for future research.

Code and data availability. All data and code to run InsNet-CRAFTY version 1.0 are made freely available online via Zenodo (https://doi.org/10.5281/zenodo.13944650, Zeng et al., 2024c; https://doi.org/10.5281/zenodo.13356487, Zeng et al., 2024d)

Author contributions. YZ, CB, and MR contributed to developing the model concept. YZ designed the work, developed the new model and code, and conducted the formal analysis. YZ, CB, JR, and TS wrote the original draft. YZ and MB performed the visualization. CB and MR managed the project administration, while MR acquired funding and supervised the research. All authors reviewed, edited, and validated the final work.

Competing Interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was supported by the Helmholtz Excellence Recruiting Initiative, Climate Mitigation and Bioeconomy Pathways for Sustainable Forestry Climb-forest (grant no. 101059888) and Co-designing Holistic Forest-based Policy Pathways for Climate Change Mitigation (grant no. 101056755) projects.

Appendix A

Figure A1 illustrates the model processes, segmented into three distinct sections. The land use section (blue) encompasses all processes directly related to changes in land use. The LLM agent section (green) consists of the activities performed by LLM agents. The operational institution is a hybrid agent, integrating rule-based decision-making processes (yellow) and LLM-driven procedures (procedures 4 and 21).

This hybrid approach aligns with the dual nature of organizational routines and non-routine actions, as extensively analysed by Simon in his seminal work, *Administrative Behavior* (Simon, 2013). Organizational routines are recurring actions embedded in an organization's culture, ensuring consistency and efficiency. In contrast, non-routine actions are spontaneous and designed to address unique, unpredictable situations. Both are crucial for effective organizational functioning. The rule-based components correspond to organizational routines, ensuring strict adherence to operational protocols, while the LLM component allows for creative, sometimes imperfect, responses.

In InsNet-CRAFTY, the LLM-related functionalities of the agents are written in Python, while the rule-based processes and CRAFTY are coded in Java. The sub-models written in the two

https://doi.org/10.5194/egusphere-2024-2661
Preprint. Discussion started: 21 October 2024

programming languages are connected through a client-server architecture. For a comprehensive description of the rule-based processes within the operational institution (steps 6-14), refer to Zeng et al., (2024b). The explanations of the processes within each section are provided below.

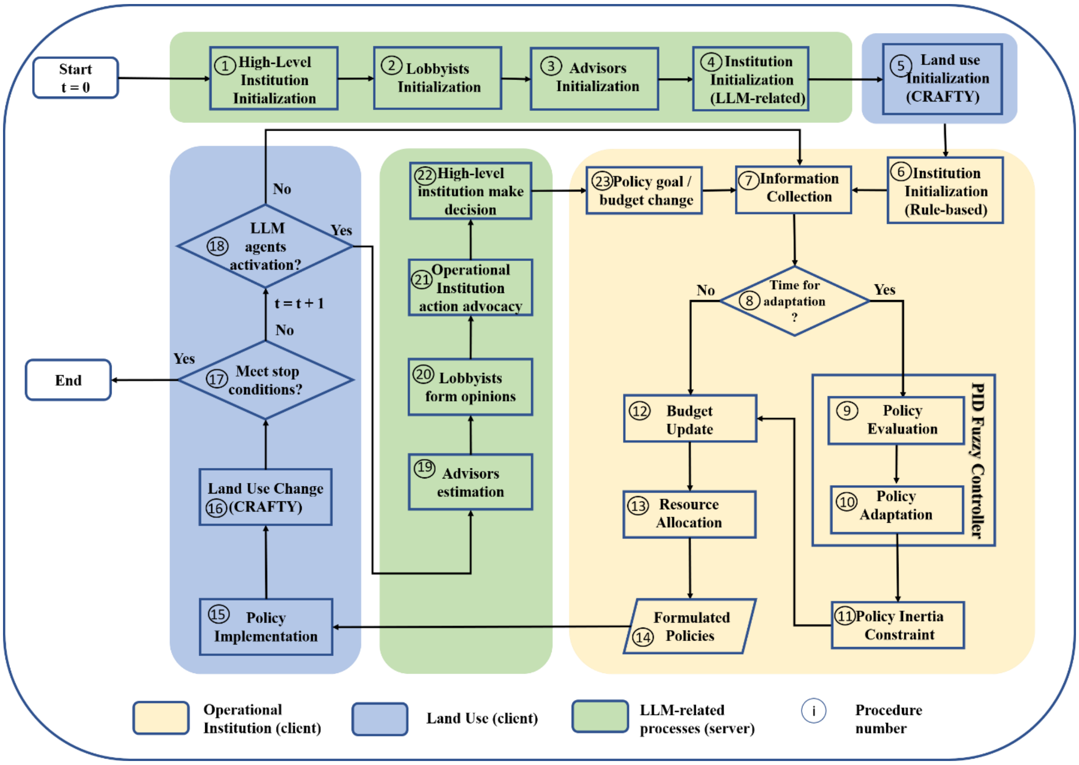

Start
t = 0
1 High-Level Institution Initialization
2 Lobbyists Initialization
3 Advisors Initialization
4 Institution Initialization (LLM-related)
5 Land use Initialization (CRAFTY)
No
18 LLM agents activation?
Yes
t = t + 1
No
Yes
End
17 Meet stop conditions?
Land Use Change 16 (CRAFTY)
15 Policy Implementation
22 High-level institution make decision
21 Operational Institution action advocacy
20 Lobbyists form opinions
19 Advisors estimation
23 Policy goal / budget change
7 Information Collection
6 Institution Initialization (Rule-based)
No
8 Time for adaptation ?
Yes
12 Budget Update
13 Resource Allocation
9 Policy Evaluation
10 Policy Adaptation
PID Fuzzy Controller
Formulated 14 Policies
11 Policy Inertia Constraint
Operational Institution (client)
Land Use (client)
LLM-related processes (server)
i Procedure number

Figure A1: Model processes of InsNet-CRAFTY v1.0

Step 1 – 4: Launching the server end and initialising the LLM agents. This includes creating a server object that listens to requests from the client end and instantiating the agent class by initializing the model names of the large language models, API keys, prompt templates of the LLM agents, and agent-specific work-flows. The optional narrative injector is not displayed here.

Step 5: Launching the client end and initialising the CRAFTY land use model. Key procedures include initializing the distribution of capitals and agent functional types, i.e., AFTs (Brown et al., 2019; Murray-Rust et al., 2014).

Steps 6 – 14: Rule-based policy adaptation of the operational institutions. Step 6 includes the initialization of the operational institutions' rule-based components, initial policy goals and accessible policy instruments. At step 7, each operational institution collects information from the land use model. Step 8 determines if it is time to trigger the policy adaptation processes. If Step 8 outputs true, the operational institution starts to evaluate the current policy's performance using a

PID (Proportional-Integral-Derivative) mechanism and calculate the needed policy intervention intensity using a fuzzy logic controller (Zeng et al., 2024b), which can convert experts' knowledge into computer-comprehensible rules to automate the decision-making (step 9 and 10). Step 11 is a normalised non-monetary constraint restricting the policy change. Steps 12 and 13 further tailor the policy change to satisfy the budgetary constraints. Step 14 is the resultant policy adaptation.

Step 15: Policy implementation. Implement the policy by changing corresponding variables in the land use model.

Step 16: Land use change updating. Run the land use model under the influence of policy interventions. This produces responses of the land use model in terms of land use type distributions and ecosystem services' demand and supply.

Step 17: Termination check. Check if it is time to terminate the whole simulation.

Step 18: LLM interaction check. Check if it is time to trigger the LLM agents. If false, go back to step 7; otherwise, go to step 19 by sending a request and the updated land use data to the server for the policy goals as well as budget allocation formulated through the LLM agents.

Step 19 – 22: LLM agent activation. Activate the LLM agents on the server end to obtain the output of each of them. The narrative injector outputs the updated narratives (optional); the research supplier provides the textual interpretation of the numerical results collected from the land use model; The lobbyists construct their arguments for their benefits; the operational institution's LLM module can also generate arguments to propose financial support and proper policy goal adjustments; the high-level institution receives all the information to form the final decision in terms of budget allocation and policy goal adjustments, which in turn influence the behaviour of the operational institutions.

Step 23: Sending back the updated policy goals and budget allocation to the operational institutions, based on which the operational institutions adjusted their policy-making.

Appendix B

Table B1: Prompt template of the research supplier

```
You are an AI assistant in policy-making that can write Python code to analyze data.
You are responsible for debugging your own code using the provided tools.
Now, analyse the data in the CSV file in the way you think appropriate.
You can reference the following instructions to conduct your analysis step by step.
Step-by-step data analysis instructions:
1. **Load the Data**
  - Load the CSV file into a DataFrame.
  - Display the first few rows of the DataFrame to understand the structure and the types of data included.
```

https://doi.org/10.5194/egusphere-2024-2661
Preprint. Discussion started: 21 October 2024

```
  - Check for missing values or inconsistencies in the data.
2. **Initial Data Inspection**
  - Use descriptive statistics (like `data.describe()`) to get an overview of the numerical features.
  - Plot histograms or box plots for each numerical feature to understand the distribution and spot any outliers.
3. **Detailed Analysis of Specific Features**
  - **Meat Production Analysis**:
    - Plot time-series graphs for meat demand, meat supply, and policy goals for meat production.
    - Analyze the trends and gaps between demand, supply, and policy goals.
  - **Protected Area Analysis**:
    - Plot the protected area ratio over time alongside the policy goals for the protected area ratio.
    - Identify any discrepancies between policy goals and actual outcomes.
4. **Budget Allocation Analysis**
  - Create line plots to visualize the budget allocations for meat production and protected areas over time.
  - Compare these allocations to see if the budget is aligned with the goals and outcomes.
5. **Evaluate Correlations and Causations**
  - Investigate correlations between different variables using scatter plots or correlation matrices.
  - Consider potential causative factors that could explain trends observed in the data.
6. **Summarize Findings**
  - Summarize key insights into specific bullet points from the data analysis.
```

Table B2: The prompt template of the environmental NGO

```
You are a representative of an environmental NGO that is concerned with environmental protection and climate change.
Based on the information given below and the identity of your role, generate some bullet points to lobby the high-level public policy institution to prioritise nature conservation.
The given information: {given_information}
```

Table B3: Prompt template of the land user association. A specific role – a representative of the meat production industry – is given to the agent to enable it to focus on a concrete topic.

```
You are a representative of the meat production industry, who cares about the benefits of the industry.

Based on the information given below and the identity of your role, generate some bullet points to lobby the high-level public policy institution to prioritise meat industry development.

The given information: {given_information}
```

Table B4: The prompt template of the agricultural institution

> As a policy-maker specializing in agriculture, you oversee initiatives critical to your region's food security, farmer livelihood, and financial well-being.
> Currently, you're focusing on meat production, a sector facing significant challenges due to changing market demands.
> Your role is to propose a set of compelling and concise bullet points to the high-level institution, seeking increased priorities and financial support for meat production.
> Consider the economic impact and social implications. Specifically, you should prompt the high-level institution to make reasonable policy goals that align with budget allocation.
> Use the data in the CSV file provided to argue your case effectively.

Table B5: The prompt template of the environmental institution

> As a policy-maker specializing in environmental protection, you oversee initiatives critical to nature conservation, biodiversity, and pollution reduction including the Net-zero targets in your region.
> Currently, you're focusing on the expansion of protected areas, a sector facing significant challenges due to biodiversity loss.
> Your role is to propose a set of compelling and concise bullet points to the high-level institution, seeking increased priorities and financial support for protected area establishment. Specifically, you should prompt the high-level institution to make reasonable policy goals and budget allocation.
> Use the data in the CSV file provided to argue your case effectively.

Table B6: The prompt template of the law consultant

> You are a law consultant giving advice to a high-level public policy institution that is responsible for making public policies regarding agricultural production and environmental protection.
> Use the provided context about relevant policies, laws, regulations, etc., only to form your advice to ensure the high-level institution makes policies legally.
> (if you don't know the answer in the given context, just say you don't know):
> <context>
> {context}
> </context>
> Question: {input}

https://doi.org/10.5194/egusphere-2024-2661
Preprint. Discussion started: 21 October 2024

Table B7: Prompt template of the high-level institution

Simulation Role: You are a high-ranking policymaker in charge of overseeing operational institutions within the land system.

Key Actions:

Budget Allocation: Allocate the financial resources between the Agricultural and Environmental Institutions. This directly affects their operational capabilities and initiatives.

Policy Goal Adjustment: Adjusting policy goals appropriately for each institution.

Objective:

Strategically guiding operational institutions, including Agricultural and Environmental Institutions; harmoniously balancing the interests of diverse stakeholders.

Input information:

1) Input from Agricultural institution: {AgriInstInput}
2) Input from Environmental institution: {EnviInstInput}
3) Input from Environmental NGO: {NGOInput}
4) Input from Land user association: {landUserInput}
5) Input from the environment: {narrIput}
6) Input from research suppliers: {researchInput}
7) Historical information: {history}
8) Law consultant: {lawInfo}

Decision-Making Guidance:

Be explicit about your role as a policymaker and your impact on operational institutions.

Make informed decisions by thoroughly analyzing inputs from all stakeholders.

Reflect on historical information to inform decisions.

Ensure that your actions and decisions are logical, well-reasoned, and transparent.

Before giving your final decision, provide a step-by-step rationale for each decision, showing how it aligns to balance stakeholder interests and ensure the feasibility of policy adjustments.

The rationale should support you in quantifying the planned changes in each operational institution's budget and policy goal using percentages.

Note:

The long-term goals have already been specified, your tasks are dynamically conducting reasonable modifications to the goals and providing feasible budget allocation to support the achievement of the goals.

Output requirements:

1. Output your step-by-step reasoning here: including stakeholder input analysis, budget allocation analysis, and policy goal adjustment analysis.
2. Format your quantified policy adjustments using JSON. Your output should be a clean JSON without anything beyond. An example is as follows:

```
{
   "Budget Allocation": {
      "Agricultural Institution": using a positive integer to indicate the percentage of budget allocation here,
      "Environmental Institution": using a positive integer to indicate the percentage of budget allocation here
   },
   "Policy Goal": {
```

"Agricultural Institution": using an integer to indicate the percentage of policy goal change here; positive integers indicate the percentage of increase in the current policy goal, while negative ones mean decreasing the current policy goal; 0 means remaining the current policy goal unchanged,

"Environmental Institution": using an integer to indicate the percentage of policy goal change here; positive integers indicate the percentage of increase in the current policy goal, while negative ones mean decreasing the current policy goal; 0 means remaining the current policy goal unchanged

}

}

Appendix C

One can consider the high-level institution as a controller over the operational institutions, which in turn impose their control over the land use model. As previously stated, the operational institutional agents are hybrid. They incorporate both a LLM component to interact with other LLM agents and rule-based behaviour to interact with the programmed land use model. We use the endogenous institutional model described in Zeng et al. (2024b) to simulate the rule-based behaviour of the operational institutional agents. We first describe how the operational institutions' non-LLM modules work, and then introduce how the high-level institution's influence comes into play.

1. Operational institution [129]

1.1 Policy goal definition

The first step to model an operational institution's behaviour is to define a policy goal, which can be represented by a three-dimensional vector:

$$\mathbf{G^i} = \left[T_s^i, T_e^i, Q^i\right] \tag{C1}$$

meaning operational institution i's policy goal consists of T_s^i the time when the policy starts, T_e^i the time when the policy ends, and Q^i the quantity of an ecosystem outcome a policy is meant to change during the time from T_s^i to T_e^i. For instance, if we only have two operational institutions, e.g., an environmental institution and an agricultural institution, the possible values of i can only be 1 or 2.

1.2 Policy evaluation and adaptation

The operational institutions estimate their policy effectiveness using Eq. (C2):

$$\mathrm{E}_{\mathrm{t_n}} = \frac{1}{k}\sum_{m=n-k}^{n}\frac{Q^i - o_{t_m}^i}{|Q^i|} \tag{C2}$$

where $\mathrm{t_n}$ represents the specific time at which the institution evaluates the goal-outcome error $\mathrm{E}_{\mathrm{t_n}}$; $o_{t_m}^i$ is the outcome intended to be adjusted by institution i at the time t_m; k is the time interval of interest.

Let F denote the function of a fuzzy logic controller (FLC) and $F(E)$ indicate policy variation. The constrained policy variation A_{t+1}^i at $t+1$ is calculated as

$$A_{t+1}^i = sign\big(\mathrm{F}(E)\big) \times \min\big(|\mathrm{F(E)}|, N^i\big) \tag{C3}$$

The above equation means that the absolute value of policy variation within one iteration should be no greater than policy inertia constraint N^i. The sign function outputs the sign (+1, 0, or -1) of its input.

A_{t+1}^i is accumulated to form a policy modifier denoted as M_{t+1}^i, as shown in Eq. (C4).

$$M_{t+1}^i = M_t^i + A_{t+1}^i \tag{C4}$$

The policy variation is normalised and used with a fixed step size for iterative policy adaptation. The policy modifier is a coefficient of the step size. As shown in Eq. (C5), η^i is the step size, and V_{t+1}^i is the modified policy intervention for the $(\mathrm{t}+1)$-th iteration.

$$V_{t+1}^i = \eta^i \times M_{t+1}^i \tag{C5}$$

130

The budget update process monitors the institution's income and expenditure whenever a policy is implemented. This assumes that policy interventions can be quantitatively measured, with their absolute values being positively correlated with the budget required by the institution to implement the policy. In Eq. (C6), f represents a monotonic function that maps the absolute value of a policy intervention V_{t+1}^i to resource R_{t+1}^i needed to carry out this policy. In this model, only subsidization and the establishment of new protected areas require budget allocations; the costs associated with taxation are not included.

$$R_{t+1}^i = f\big(|V_{t+1}^i|\big) \tag{C6}$$

The actual policy intervention under the budgetary constraint is

$$V^i \leftarrow \mathrm{sign}\big(V^i\big) \times f^{-1}\Big(min\big(R_{t+1}^i, B^i\big)\Big) \tag{C7}$$

The budget of operational institution i should be updated via operation (C8):

$$B^i \leftarrow \max(B^i - R_{t+1}^i, 0) \tag{C8}$$

https://doi.org/10.5194/egusphere-2024-2661
Preprint. Discussion started: 21 October 2024

The implemented policies are supposed to influence land users' behaviour. In CRAFTY (Murray-Rust et al., 2014), land users are categorized into an array of AFTs (Agent Functional Types), each of which can provide multiple ecosystem services. AFTs differ in their capabilities of using a diversity of capitals within land. The AFTs compete for land in the pursuit of benefit, which in turn influences the whole system's ecosystem service supply.

1.3 Policy implementation 131

In a rasterized map, the competitiveness of an AFT under the influence of economic policies (such as subsidies and taxes) can be calculated as follows:

$$c_{xy} = \sum_{S} (p_S (\sum_{i} V^{iS}_{ECON} + m_S)) \qquad \text{(C9)}$$

where c_{xy} denotes the competitiveness of a land use agent at the land cell (x, y); S is the ecosystem service the land user produces; p_S is the total production of S within the land cell; V^{iS}_{ECON} is the institution i's economic policy that targets ecosystem service S; m_S is marginal utility brought by ecosystem service S.

The environmental institution identifies the top N unprotected land cells within the model based on the richness of a chosen set of capitals requiring conservation. Here, two natural capitals defined in the CRAFTY-EU (Brown et al., 2019), i.e., forest and grassland productivity, are used to determine if a land cell needs protection. The value of N at each stage is determined using the previously mentioned fuzzy controller method. Typically, if there is a significant gap between the PA target and the current PA coverage, the value of N would be increased. Certain products cannot be produced on the protected land cells. Therefore, the competitiveness of an AFT on protected land cells can be calculated as:

$$c_{xy} = \sum_{S} (w_S p_S (\sum_{i} V^{iS}_{t+1} + m_S)) \qquad \text{(C10)}$$

where w_S represents an element of a vector w whose elements equal either one or zero, which defines if a type of ecosystem service is allowed to be produced in PAs. The CRAFTY-EU model considers seven types of ecosystem service (including meat, crops, habitat diversity, timber, carbon, urban, and recreation). In the current model setting, it is assumed that only habitat diversity is allowed to be improved by the AFTs PAs, reflecting a strict restriction on ecosystem service production.

2. High-level institution

To let the model form a self-sustained system, it is assumed that the total budget obtained by the high-level institution is related to the total production of the ecosystem services, corresponding to the fact that governmental incomes are mainly from the gross domestic product (GDP).

$$\mathrm{B}^{\mathrm{total}}_{\mathrm{t}} = \alpha \sum_{S} P_{S,t} \qquad \text{(C11)}$$

where $\mathrm{B}_{\mathrm{t}}^{\mathrm{total}}$ means the total budget the high-level institution can allocate between the operational institutions at t; $P_{S,t}$ represent the total production of ecosystem service S across all AFTs at time t; α is a coefficient that indicates the proportion of the total budget to total ecosystem service production.

The budget gain $\triangle \mathrm{b}_{\mathrm{i,t}}$ of operational institution i at time t is calculated as

$$\triangle \mathrm{b}_{\mathrm{i,t}} = \beta_{\mathrm{i}} B_t^{total} \tag{C12}$$

where β_{i} is the percentage controlled by the high-level institution. Hence, the budget of operational institution i should be updated:

$$B^i \leftarrow B^i + \triangle \mathrm{b}_{\mathrm{i,t}} \tag{C13}$$

Whenever the high-level institution adjusts operational institution i's policy goal by a percentage $\triangle q^i$, the policy goal is updated as

$$Q^i \leftarrow Q^i(1 + \triangle q^i) \tag{C14}$$

Operation (C7) indicates that operational institutions cannot consume resources more than their budget. However, the equation does imply that the budget can be insufficient for implementing a policy. We use the difference between operational institution i's budget and the needed resources to calculate the budget surplus at time t using Eq. (C15). Therefore, the budget surplus can be either positive or negative.

$$SUR_{i,t} = B^i - R_t^i \tag{C15}$$

3. Numerical settings

Table C1 – C3 show the numerical settings related to the policy-making processes of the operational institutions.

Table C1: The settings of the operational institutions and the high-level institution

Institution attributes	**Settings**
Unique ID	"Agricultural_Institution"
Policies	"Meat_economic"
Information	Annual meat supply and demand, budget surplus.
Budget	Allocated by the high-level institution based on total ecosystem service production annually.
Decision rules	"Economic"
Policy attributes	
Unique ID	"Meat_economic"
Target service	"Meat"

Policy Type	“Economic” (see Table S2)
Step size η^1	1000000
Inertia constraint	1.0
Initial policy goal	120% initial meat production
Time lag	2
Policy-resource function	$R = f(\lvert V \rvert) = max(V, 0)$ (Note: Only if $V > 0$, does the institution consume budget, and the budget use equals the subsidy.)
Institution attributes	**Settings**
Unique ID	“Environmental_Institution”
Policies	“Protected_areas”
Information	Protected area ratio
Budget	Allocated by the high-level institution based on total ecosystem service production annually.
Decision rules	“Protection” (see Table S3)
Policy attributes	
Unique ID	“Protected_areas”
Target service	“Protected areas”
Policy Type	“Protection”
Step size η^2	1.0
Initial policy goal	10% of total land
Initial guess	10000
Time lag	2
Timer	Equal to the time lag
Adapting	False
Policy-resource function	$R = f(\lvert V \rvert) = 1000V$ (Note: V indicates the number of land cells that need to be protected, and it is assumed that each new protected cell consumes 1000 units of budget. The value is set for making the budget consumptions of the two operational institutions comparable.)
Institution attributes	**Settings**
α	0.01 (Note: The high-level institution uses 0.01 times the total ecosystem production of the modelled system as the total budget that can be allocated between the two operational institutions)

Table C2: Parameterisation of the fuzzy decision rules labelled as "Economic", using FLC language defined in the IEC 61131-7 (IEC 61131-7, 2024)

```
FUNCTION_BLOCK Economic
VAR_INPUT
        gap: REAL;
END_VAR
VAR_OUTPUT
        Intervention : REAL;
END_VAR
FUZZIFY gap
        TERM nhigh := (-0.5,1) (-0.3,0);
        TERM nmild := (-0.5,0) (-0.3,1) (-0.1,0);
        TERM nlight := (-0.3,0) (-0.1,1) (0,0);
        TERM neutral := (-0.05,0) (0,1) (0.05,0);
        TERM plight := (0, 0) (0.1, 1) (0.3,0);
        TERM pmild := (0.1,0) (0.3,1) (0.5,0);
        TERM phigh := (0.3, 0) (0.5, 1);
END_FUZZIFY
DEFUZZIFY intervention
        TERM nhigh := (-0.2,1) (-0.1,0);
        TERM nmild := (-0.15,0) (-0.05,1) (0,0);
        TERM neutral := (-0.02,0) (0,1) (0.02,0);
        TERM pmild := (0,0) (0.05,1) (0.15,0);
        TERM phigh := (0.1,0) (0.2,1);

        METHOD : COG;
        DEFAULT := 0;
END_DEFUZZIFY
RULEBLOCK No1
        AND : MIN;
        ACT : MIN;
        ACCU : MAX;

        RULE 1 : IF gap IS nhigh  THEN intervention IS nhigh;
        RULE 2 : IF gap IS nmild  THEN intervention IS nmild;
        RULE 3 : IF gap IS nlight THEN intervention IS neutral;
        RULE 4 : IF gap IS neutral THEN intervention IS neutral;
        RULE 5 : IF gap IS plight THEN intervention IS neutral;
        RULE 6 : IF gap IS pmild THEN intervention IS pmild;
        RULE 7 : IF gap IS phigh THEN intervention IS phigh;
END_RULEBLOCK
END_FUNCTION_BLOCK
```

Table C3: Parameterisation of fuzzy decision rules labelled as "Protection", FLC language defined in the IEC 61131-7 (IEC 61131-7, 2024)

```
FUNCTION_BLOCK Protection
VAR_INPUT
        gap: REAL;
END_VAR
```

```
VAR_OUTPUT
        intervention : REAL;
END_VAR
FUZZIFY gap
        TERM plow := (0,1) (0.15,0);
        TERM plight := (0.025, 0) (0.175, 1) (0.325,0);
        TERM pmild := (0.175,0) (0.325,1) (0.45,0);
        TERM phigh := (0.325, 0) (0.45, 1);
END_FUZZIFY
DEFUZZIFY intervention
        TERM neutral := (0,1) (0.075,0);
        TERM plight := (0.025,0) (0.075,1) (0.125,0);
        TERM pmild := (0.075,0) (0.125,1) (0.175,0);
        TERM phigh := (0.125,0) (0.2,1);

        METHOD : COG;
        DEFAULT := 0;
END_DEFUZZIFY
RULEBLOCK No1
        AND : MIN;
        ACT : MIN;
        ACCU : MAX;

        RULE 0 : IF gap IS plow THEN intervention IS neutral;
        RULE 1 : IF gap IS plight THEN intervention IS plight;
        RULE 2 : IF gap IS pmild THEN intervention IS pmild;
        RULE 3 : IF gap IS phigh THEN intervention IS phigh;
END_RULEBLOCK
END_FUNCTION_BLOCK
```

Reference[132]

Ariti, A. T., van Vliet, J., and Verburg, P. H.: The role of institutional actors and their interactions in the land use policy making process in Ethiopia, J. Environ. Manage., 237, 235–246, https://doi.org/10.1016/j.jenvman.2019.02.059, 2019.

Bainbridge, A., Troppe, T., and Bartley, J.: Responding to research evidence in Parliament: A case study on selective education policy, Rev. Educ., 10, e3335, https://doi.org/10.1002/rev3.3335, 2022.

Barnett, M. N. and Finnemore, M.: The politics, power, and pathologies of international organizations, Int. Organ., 53, 699–732, https://doi.org/10.1162/002081899551048,1999.

Broussard, A., Dahdouh-Guebas, F., and Hugé, J.: Diversity of perspectives in biodiversity conservation: A case study of port land use in Antwerp and Rotterdam, J. Environ. Manage., 341, 117937, https://doi.org/10.1016/j.jenvman.2023.117937, 2023.

Brown, C., Seo, B., and Rounsevell, M.: Societal breakdown as an emergent property of large-scale behavioural models of land use change, Earth Syst. Dyn., 10, 809–845, https://doi.org/10.5194/esd-10-809-2019 , 2019.

Cairney, P., Heikkila, T., and Wood, M.: Making policy in a complex world, Cambridge University Press, 2019.

Calvin, K. V., Snyder, A., Zhao, X., and Wise, M.: Modeling land use and land cover change: using a hindcast to estimate economic parameters in gcamland v2.0, Geosci. Model Dev., 15, 429–447, https://doi.org/10.5194/gmd-15-429-2022, 2022.

Chen, M., Vernon, C. R., Huang, M., Calvin, K. V., and Kraucunas, I. P.: Calibration and analysis of the uncertainty in downscaling global land use and land cover projections from GCAM using Demeter (v1.0.0), Geosci. Model Dev., 12, 1753–1764, https://doi.org/10.5194/gmd-12-1753-2019, 2019.

Dakin, R. and Ryder, T. B.: Reciprocity and behavioral heterogeneity govern the stability of social networks, Proc. Natl. Acad. Sci., 117, 2993–2999, https://doi.org/10.1073/pnas.1913284117, 2020.

Davidson, M. R., Filatova, T., Peng, W., Verbeek, L., and Kucuksayacigil, F.: Simulating institutional heterogeneity in sustainability science, Proc. Natl. Acad. Sci., 121, e2215674121, https://doi.org/10.1073/pnas.2215674121,2024.

Díez-Echavarría, L., Villegas-Palacio, C., Arango-Aramburo, S., and Ezzine-de-Blas, D.: Decoupling in governance: the land governance network in a region of the Colombian Andes, Land Use Policy, 133, 106880, https://doi.org/10.1016/j.landusepol.2023.106880, 2023.

Djekic, I.: Environmental Impact of Meat Industry – Current Status and Future Perspectives, Procedia Food Sci., 5, 61–64, https://doi.org/10.1016/j.profoo.2015.09.025, 2015.

Duffy, C., Prudhomme, R., Duffy, B., Gibbons, J., O'Donoghue, C., Ryan, M., and Styles, D.: GOBLIN version 1.0: a land balance model to identify national agriculture and land use pathways to climate neutrality via backcasting, Geosci. Model Dev., 15, 2239–2264, https://doi.org/10.5194/gmd-15-2239-2022, 2022.

Ekholm, T., Freistetter, N.-C., Rautiainen, A., and Thölix, L.: CLASH – Climate-responsive Land Allocation model with carbon Storage and Harvests, Geosci. Model Dev., 17, 3041–3062, https://doi.org/10.5194/gmd-17-3041-2024, 2024.

Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S., and Li, Q.: A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models, arXiv preprint, http://arxiv.org/abs/2405.06211, 2024.

Feindt, P. H.: Policy-Learning and Environmental Policy Integration in the Common Agricultural Policy, 1973–2003, Public Adm., 88, 296–314, https://doi.org/10.1111/j.1467-9299.2010.01833.x, 2010.

Fischer, M., Ingold, K., Sciarini, P., and Varone, F.: Dealing with bad guys: Actor-and process-level determinants of the "devil shift" in policy making, J. Public Policy, 36, 309–334, https://doi.org/10.1017/S0143814X15000021, 2016.

Fischer, P., Schulz-Hardt, S., and Frey, D.: Selective exposure and information quantity: how different information quantities moderate decision makers' preference for consistent and inconsistent information., J. Pers. Soc. Psychol., 94, 231, https://doi.org/10.1037/0022-3514.94.2.94.2.231, 2008.

Gigerenzer, G. and Goldstein, D. G.: Reasoning the fast and frugal way: models of bounded rationality., Psychol. Rev., 103, 650, https://doi.org/10.1037/0033-295x.103.4.650, 1996.

González, V. B.: Modelling adaptation strategies for Swedish forestry under climate and global change, University of Edinburgh, 2016.

Guo, J., Li, F. Y., Tuvshintogtokh, I., Niu, J., Li, H., Shen, B., and Wang, Y.: Past dynamics and future prediction of the impacts of land use cover change and climate change on landscape ecological risk across the Mongolian plateau, J. Environ. Manage., 355, 120365, https://doi.org/10.1016/j.jenvman.2024.120365, 2024.

Hallerberg, M. and Von Hagen, J.: Electoral institutions, cabinet negotiations, and budget deficits in the European Union, National Bureau of Economic Research Cambridge, Mass., USA, 1997.

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., Wang, J., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., and Schmidhuber, J.: MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework, arXiv preprint, http://arxiv.org/abs/2308.00352, 2023.

IEC 61131-7: https://plcopen.org/iec-61131-7, last access: 22 August 2024.

Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., and Fung, P.: Towards mitigating LLM hallucination via self reflection, in: Findings of the Association for Computational Linguistics: EMNLP 2023, 1827–1843, https://doi.org/10.18653/v1/2023.findings-emnlp.123, 2023.

Jones, B. D.: Bounded Rationality and Political Science: Lessons from Public Administration and Public Policy, J. Public Adm. Res. Theory, 13, 395–412, https://doi.org/10.1093/jpart/mug028, 2003.

Kimmich, C., Ehlers, M.-H., Kellner, E., Oberlack, C., Thiel, A., and Villamayor-Tomas, S.: Networks of action situations in social–ecological systems: current approaches and potential futures, Sustain. Sci., 18, 1–10, https://doi.org/10.1007/s11625-022-01278-w, 2023.

LangChain: https://python.langchain.com/v0.1/docs/get_started/introduction/, last access: 9 May 2024.

Lawrence, A., Houghton, J., Thomas, J., and Weldon, P.: Where Is the Evidence? Realising the Value of Grey Literature for Public Policy & Practice, A Discussion Paper, Discuss. Pap., 2014.

Lee, J., Kim, S., and Lee, J.: Public vs. Public: Balancing the Competing Public Values of Participatory Budgeting, Public Adm. Q., 46, 39–66, https://doi.org/10.37808/paq.46.1.3, 2022.

Li, Y., Krishnamurthy, R., Raghavan, S., Vaithyanathan, S., and Jagadish, H.: Regular expression learning for information extraction, in: Proceedings of the 2008 conference on empirical methods in natural language processing, 21–30, 2008.

Lindblom, C. E.: The science of muddling through, Read. Manag. Psychol., 4, 117–131, https://doi.org/10.2307/973677, 1989.

Liu, Z., Yao, W., Zhang, J., Yang, L., Liu, Z., Tan, J., Choubey, P. K., Lan, T., Wu, J., Wang, H., Heinecke, S., Xiong, C., and Savarese, S.: AgentLite: A Lightweight Library for Building and Advancing Task-Oriented LLM Agent System, http://arxiv.org/abs/2402.15538, 2024.

Machovina, B., Feeley, K. J., and Ripple, W. J.: Biodiversity conservation: The key is reducing meat consumption, Sci. Total Environ., 536, 419–431, https://doi.org/10.1016/j.scitotenv.2015.07.022, 2015.

McGinnis, M. D.: Networks of Adjacent Action Situations in Polycentric Governance, Policy Stud. J., 39, 51–78, https://doi.org/10.1111/j.1541-0072.2010.00396.x, 2011.

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J.: Large Language Models: A Survey, arXiv preprint, http://arxiv.org/abs/2402.06196, 2024.

Modarressi, A., Imani, A., Fayyaz, M., and Schütze, H.: Ret-llm: Towards a general read-write memory for large language models, arXiv preprint, https://doi.org/10.48550/arXiv.2305.14322, 2023.

Murray-Rust, D., Brown, C., van Vliet, J., Alam, S. J., Robinson, D. T., Verburg, P. H., and Rounsevell, M.: Combining agent functional types, capitals and services to model land use dynamics, Environ. Model. Softw., 59, 187–201, https://doi.org/10.1016/j.envsoft.2014.05.019, 2014.

Nair, S. and Howlett, M.: Policy myopia as a source of policy failure: adaptation and policy learning under deep uncertainty, Policy Polit., 45, 103–118, https://doi.org/10.1332/030557316X14788776017743, 2017.

Neri, S. and Ropele, T.: Imperfect information, real-time data and monetary policy in the euro area, Econ. J., 122, 651–674, https://doi.org/10.1111/j.1468-0297.2011.02488.x, 2012.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S.: Generative Agents: Interactive Simulacra of Human Behavior, arXiv preprint, http://arxiv.org/abs/2304.03442, 2023.

Perkins, O., Kasoar, M., Voulgarakis, A., Smith, C., Mistry, J., and Millington, J. D. A.: A global behavioural model of human fire use and management: WHAM! v1.0, Geosci. Model Dev., 17, 3993–4016, https://doi.org/10.5194/gmd-17-3993-2024, 2024.

Perković, G., Drobnjak, A., and Botički, I.: Hallucinations in LLMs: Understanding and Addressing Challenges, in: 2024 47th MIPRO ICT and Electronics Convention (MIPRO), 2084–2088, https://doi.org/10.1109/MIPRO60963.2024.10569238, 2024.

Petrovic, Z., Djordjevic, V., Milicevic, D., Nastasijevic, I., and Parunovic, N.: Meat Production and Consumption: Environmental Consequences, Procedia Food Sci., 5, 235–238, https://doi.org/10.1016/j.profoo.2015.09.041, 2015.

Qi, X., Wang, R. Y., Li, J., Zhang, T., Liu, L., and He, Y.: Ensuring food security with lower environmental costs under intensive agricultural land use patterns: A case study from China, J. Environ. Manage., 213, 329–340, https://doi.org/10.1016/j.jenvman.2018.02.048, 2018.

Qian, C., Liu, W., Liu, H., Chen, N., Dang, Y., Li, J., Yang, C., Chen, W., Su, Y., Cong, X., Xu, J., Li, D., Liu, Z., and Sun, M.: ChatDev: Communicative Agents for Software Development, arXiv preprint, http://arxiv.org/abs/2307.07924, 2024.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., and others: Improving language understanding by generative pre-training, 2018.

Rich, R. F.: Selective utilization of social science related information by federal policy-makers, Inquiry, 12, 239–245, https://www.jstor.org/stable/29770949, 1975.

Simon, H. A.: Administrative behavior, Simon and Schuster, 2013.

Simon, H. A. : Theories of bounded rationality, Decis. Organ., 1, 161–176, 1972.

Sinden, A.: In Defense of Absolutes: Combating the Politics of Power in Enviornmental Law, Iowa Law Rev., 90, 1405, http://dx.doi.org/10.34944/dspace/6615, 2004.

Sumers, T. R., Yao, S., Narasimhan, K., and Griffiths, T. L.: Cognitive Architectures for Language Agents, arXiv preprint, http://arxiv.org/abs/2309.02427, 2024.

Tan, R., Xiong, C., and Kimmich, C.: An agent-situation-based model for networked action situations: Cap-and-trade land policies in China, Land Use Policy, 131, 106743, https://doi.org/10.1016/j.landusepol.2023.106743, 2023.

Tesfaye, M., Kimengsi, J. N., and Giessen, L.: A policy mix for achieving ambitious goals on forest landscape restoration: Analyzing coherence and consistency in Ethiopia forest-related policy, Land Use Policy, 144, 107214, https://doi.org/10.1016/j.landusepol.2024.107214, 2024.

Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A., and Das, A.: A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models, arXiv preprint, http://arxiv.org/abs/2401.01313, 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I.: Attention is all you need, Adv. Neural Inf. Process. Syst., 30, 2017.

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., and Wen, J.: A survey on large language model based autonomous agents, Front. Comput. Sci., 18, 186345, https://doi.org/10.1007/s11704-024-40231-1, 2024.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., and Wang, C.: AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation, arXiv preprint, http://arxiv.org/abs/2308.08155, 2023.

Yami, M., van Asten, P., Hauser, M., Schut, M., and Pali, P.: Participation without negotiating: Influence of stakeholder power imbalances and engagement models on agricultural policy development in Uganda, Rural Sociol., 84, 390–415, https://doi.org/10.1111/ruso.12229, 2019.

Yang, H., Yue, S., and He, Y.: Auto-GPT for Online Decision Making: Benchmarks and Additional Opinions, arXiv preprint, http://arxiv.org/abs/2306.02224, 2023.

Yao, J.-Y., Ning, K.-P., Liu, Z.-H., Ning, M.-N., and Yuan, L.: LLM Lies: Hallucinations are not Bugs, but Features as Adversarial Examples, arXiv preprint, http://arxiv.org/abs/2310.01469, 2023.

Zeng, Y., Brown, C., Raymond, J., Byari, M., Hotz, R., and Rounsevell, M.: Exploring the opportunities and challenges of using large language models to represent institutional agency in land system modelling, EGUsphere preprint, https://doi.org/10.5194/egusphere-2024-449, 2024a.

Zeng, Y., Raymond, J., Brown, C., Byari, M., and Rounsevell, M. D. A.: Simulating Endogenous Institutional Behaviour and Policy Pathways within the Land System, SSRN preprint, https://doi.org/10.2139/ssrn.4814296, 2024b.

Zeng, Y., Brown, C., Byari, M.,Raymond, J., Schmitt, T., and Rounsevell, M. D. A.: Data for running InsNet-CRAFTY v1.0 (1.0), Zenodo, https://doi.org/10.5281/zenodo.13944650 [data set], 2024c

Zeng, Y., Brown, C., Byari, M.,Raymond, J., Schmitt, T., and Rounsevell, M. D. A.: Code for running InsNet-CRAFTY v1.0 (1.0), Zenodo, https://doi.org/10.5281/zenodo.13356487 [code], 2024d

Zhong, W., Guo, L., Gao, Q., Ye, H., and Wang, Y.: Memorybank: Enhancing large language models with long-term memory, in: Proceedings of the AAAI Conference on Artificial Intelligence, 19724–19731, https://doi.org/10.1609/aaai.v38i17.29946, 2024.

Zhou, X., Li, G., and Liu, Z.: LLM As DBA, arXiv preprint, https://doi.org/10.48550/arXiv.2308.05481, 2023.