Thanks for this fascinating paper. In general, as noted in my community comment, I am very supportive of this work. However, I have major comments around the reproducibility, scalability and generalisability of the work presented which need to be addressed to more precisely frame the extent of the advances made. I also have minor comments on the text (which follow below).

**Major comments**

**Reproducibility**

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

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

- Overall, I think a supplementary information expanding lines 179-191 to be really clear on what you did would be very helpful and make the work more reproducible. E.g. which RCPs / SSPs did you use? I *think* from Figure 5 that you are showing means across the scenarios, but this isn't totally clear. It would also be very interesting to see how much variance there is by scenario.

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

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

**Scaling**

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

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

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

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

**Generalisability**

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

Lines 308-312

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

Lines 407-414

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

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

**Minor comments**

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

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

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

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

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

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