1. The issue of the results not being generalizable is acknowledged by the authors. This is, undoubtedly, one of the most significant limitations I perceive in this work. In their attempt to justify this, the authors state: "We have added the following text to reflect these points. 335-342: Nevertheless, although the training set is unknown, there is little reason to believe that this particular topic would be better represented than any other topic of socio-ecological importance."  
     
   From my perspective, this statement is not accurate, and could imply a lack of scientific rigor, being inherently biased and subjective. This does not align with the standards expected by this publication or even with the quality of the rest of the paper. Scientifically, it is untenable to generalize across an entire domain based on a singular example. As far as we know, the authors do not have access to the ChatGPT training dataset and, therefore, cannot assert whether their case study is "better" or "worse" represented than others. Proposing the hypothesis that performance in other case studies would be similar as an (unproven) truth is inappropriate. I kindly urge the removal of this assertion and a clear indication that a single example is insufficient for generalization. In my view, this is a threat to validity that the authors must acknowledge in the manuscript.

We have removed the sentence the reviewer takes exception to. The passage now reads:

*We note that an important limitation of our study is that we have included only a single case study on the topic of CBFM in the Pacific.  ~~Nevertheless, although the training data set is unknown, there is little reason to believe that this particular topic would be better represented than any other topic of socio-ecological importance. Thus~~ Nevertheless, its performance in this case study* ***suggests*** *that it might perform similarly when faced with other topics in this field. Because ChatGPT's responses cannot be anticipated ahead of time due to randomness, uncertainties in the training data set, and uncertainties in the underlying infrastructure, its usefulness can only be explored empirically. Future case studies will be necessary before we can be confident that its performance is consistent across domains and research questions.*

2. The authors should consider using an Open Science repository instead of directly linking to GitHub in the manuscript. I recommended utilizing Zenodo (as they themselves have cited: 10.5281/zenodo.10595355) to make the code, data, and other replication materials available. The use of GitHub does not ensure that the material's version matches what has been reviewed and accepted in this article. A GitHub repository could be altered post-acceptance, whereas a Zenodo entry cannot. The replication package must faithfully represent what has been accepted through the peer review process. It is unclear why Zenodo is mentioned in the response to reviewers and the editor but is not utilized in the manuscript, which still contains links to GitHub.

We have replaced all of the links to github with links to zenodo.

3. It would be beneficial if the authors included a more explicit clarification regarding the treatment of randomness within the manuscript itself. It appears that randomness could pose a threat to validity in this work, and it would be prudent for readers to be informed about this aspect and how the authors have considered it. Understanding the constraints governing the experimental setup is crucial.

It is unclear what the reviewer here is referring to or how randomness threatens the validity of the work.

Supplement F has a lengthy discussion of how we introduce randomness to avoid consistently incorrect responses from the LLM. We have declined to include this in the main text as we feel this long aside is auxiliary to the main message. However, if the editors feel that Supplement F is better suited to the Methods section then we would be happy to move it there.