Prompting large language models for quality ecological statistics

Christopher J. Brown[1,2] & Scott Spillias[2,3]

1. Institute for Marine and Antarctic Studies, University of Tasmania, Hobart, Tasmania
2. Centre for Marine Socio-Ecology, University of Tasmania, Hobart, Australia
3. CSIRO Environment, Hobart, Australia

Contact: c.j.brown@utas.edu.au

Abstract

Large language models (LLMs) are rapidly transforming scientific workflows, including statistical analyses in ecological sciences. While these AI tools offer impressive capabilities for code generation and analytical guidance, evaluations reveal significant limitations in their reasoning for standard statistical tests. Ecological statistics typically require special consideration due to spatial and temporal structuring, so LLM performance on these tasks is likely to be worse than for other disciplines. This perspective addresses the need for effective prompting guidelines to ensure quality statistical analyses when using LLMs. Drawing on empirical evaluations and practical experience, we provide a framework for ecological scientists to leverage these powerful tools while maintaining statistical rigor. Key recommendations include: separating workflows into components that align with LLM strengths and limitations; providing context through domain knowledge, data summaries, and research questions; combining context with structured prompting techniques like Chain of Thought reasoning; and maintaining human oversight of statistical decisions. By understanding LLM capabilities and employing these prompting strategies, researchers can harness these technologies to improve rather than compromise statistical quality in ecological research. Future research should focus on evaluations of LLMs for ecological statistics, development of specialized prompting strategies, and integration of LLMs with traditional statistical approaches.

Introduction

Large language models (LLMs) are rapidly transforming scientific workflows, with profound implications for statistical analysis in environmental sciences. Most researchers now incorporate LLMs into their workflows (Liao et al. 2024), with many specifically using them for statistical advice and code generation (Jansen et al. 2025; Zhu et al. 2024). The appeal is clear: LLMs almost instantly generate statistical code and analyses that would traditionally require extensive training and time to develop. For example, researchers can now produce a complete bioinformatics analysis analysis—including code and visualizations (Jansen et al. 2025) in under 15 minutes. LLMs can also interpret statistics and figures to produce written results.

The efficiency of relying on LLMs for statistics comes with significant risks. Recent evaluations reveal concerning limitations in LLMs' statistical reasoning abilities. One study found that accuracy of late 2024 LLMs for suggesting appropriate statistical tests was typically below 40% for anything beyond basic descriptive statistics (Zhu et al. 2024). Crucially, the quality of statistical advice from LLMs depends heavily on how questions are framed (Onan and Alhumyani 2024; Zhu et al. 2024; Jansen et al. 2025) — effective prompting strategies can almost double the accuracy of recommendations. Present evaluations of LLMs are focused on statistical analyses where samples are independent. It is likely that their performance is significantly poorer for the complex dependence structures and observation patterns that are common in ecology.

This perspective article addresses the urgent need for guidelines on using LLMs for statistical analysis in ecological research. As LLM adoption outpaces formal evaluation, we cannot wait for comprehensive peer-reviewed assessments before establishing best practices. Drawing on empirical evaluations, practical experience, and broader AI literature, we provide a framework for leveraging these powerful tools while maintaining statistical rigor. By understanding LLM capabilities and limitations and employing structured prompting strategies, researchers can harness these technologies to enhance rather than compromise statistical quality in environmental research. The guidelines presented here aim to help environmental scientists navigate this rapidly evolving landscape responsibly and effectively.

Challenges for statistical analysis quality in environmental sciences

Statistical analysis in ecological sciences faces numerous challenges that predate the emergence of LLMs but may be exacerbated by their use. Modern data analysis requires two interrelated skills: computer programming and statistical reasoning. There exists a substantial gap between specialists at the forefront of statistical computing and experts in specific ecological disciplines who use statistics irregularly (Gilbert et al. 2024). Environmental data often violate standard statistical assumptions, requiring specialized analytical approaches (Gilbert et al. 2024) that may not be well-represented in the text that LLMs are trained on. Ecological analyses may also require advanced computer programming skills where it is easy to make mistakes (e.g. Kendall et al. 2019).

Lack of statistical training among environmental scientists has long undermined research quality and application. Reproducibility is a wide-spread issue and ecology is no exception. P-hacking and other forms of bias caused by manipulating analyses after viewing results are already prevalent in ecology and evolution, often justified by researchers as necessary for career survival (Fraser 2018; Forstmeier, Wagenmakers, and Parker 2017).

Accidental statistical mistakes caused by inappropraite training or misguided conventions are also an issue. Common misapplications include inappropriate transformations of response variables (O'Hara and Kotze 2010), applying methods that assume independent samples to time-series analysis (Brown et al. 2011), using linear regression for zeroinflated data (Warton et al. 2016), conflating prediction with causality (Arif and MacNeil 2022), and inappropriate use of multi-model averages (Bolker 2024).

Problems can also arise from flawed implementation of ecological analysis (e.g. Kendall et al. 2019). These errors are not merely academic concerns—they can lead to misinformed policy actions with real consequences for conservation outcomes (Shoemaker and Loope 2025).

Applications and Risks of LLMs in Environmental Statistics

Large language models present both significant opportunities and challenges for statistical practice in environmental sciences. When used with appropriate guidance and oversight, these AI tools can enhance research workflows, but they also introduce risks that require careful consideration.

Risks and Limitations

LLMs present several specific risks for statistical practice that require careful mitigation strategies (Table 1).

LLMs may amplify existing problems with statistical quality. By dramatically accelerating the ability to try multiple analytical approaches, LLMs could enable unprecedented levels of p-hacking and selective reporting. Researchers can now explore tens or hundreds of alternatives for solving a statistical issue in minutes, creating far more opportunities to cherry-pick favorable results. Strong research reporting standards and ethics are ultimately needed to combat this issue.

Many of the statistical methods suggesting by current LLMs are plausible, syntactically correct, but logically flawed (Zhou et al. 2024; Jansen et al. 2025). They almost always provide an answer, typically with high apparent certainty, even when their suggestions are inappropriate or incorrect. When asked to self evaluated their answers, LLMs exhibit overconfidence in their statistical recommendations, and this trend is worsening as models are scaled-up to large parameter sets (Zhou et al. 2024). For example, they perpetuate common misunderstandings of confidence intervals and p-values (Ellis and Slade 2023) (as of Claude 4.0 this was still true). This characteristic is particularly problematic in environmental sciences, where data often have complex structures requiring specialized approaches. Current LLMs may not adequately recognize or account for these nuances of environmental data. Notably their conversation patterns differ remarkably from a human statistical consultant - the human will tend to ask more questions early on in a conversation than an AI assistant would, giving the human better understanding of study design and research context (Figure 1).

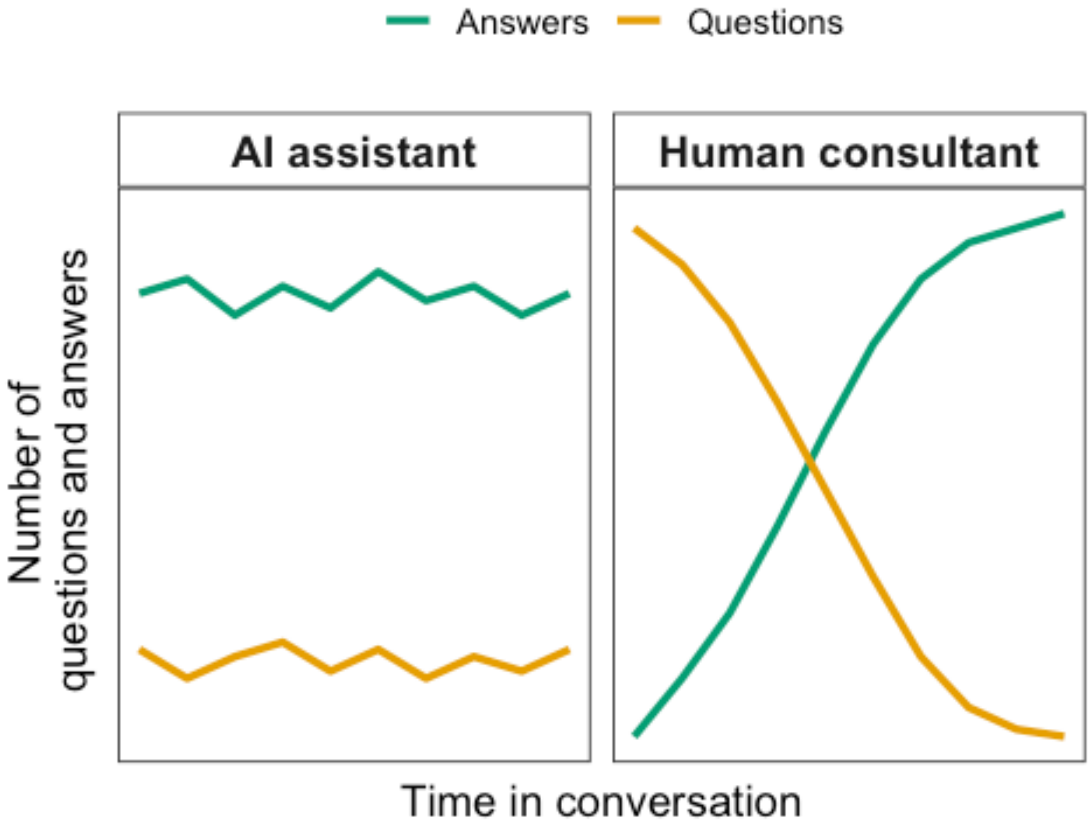

Figure 1 Comparison of how an experienced human statistical consultant would structure a conversation compared to a typical prompt chain with an AI assistant (figure 1). The human consultant will usually ask more questions than provide answers at the start of a conversation, then switch to providing more answers once they understand the context of the study. An AI assistant will tend to be constant in the number of questions it asks, unless explictly prompted to ask questions rather than provide answers. This means it provides answers without first gathering appropriate context.

The propensity to give believable answers with high confidence creates an illusion that LLMs true have statistical understanding (Messeri and Crockett 2024). Unlike traditional statistical software that implements specific algorithms, LLMs generate responses based on patterns learned from training data, they do not "understand" statistics and cannot reason in the way human experts do. They work by predicting statistically likely responses to text - therefore the challenge is to use prompts that shift the distribution of likely response to better overlap with accurate responses. This fundamental limitation means they may confidently suggest inappropriate methods (Zhou et al. 2024), apply internally consistent logic to the wrong question, fail to recognize violations of statistical assumptions, or generate plausible-sounding but incorrect interpretations.

Inexperienced users may be particularly vulnerable to the risk false statistical suggestions (Ellis and Slade 2023). Without sufficient statistical background to critically evaluate LLM suggestions, researchers might implement inappropriate analyses or misinterpret results. The apparent authority and confidence of LLM responses can create a false sense of security (Messeri and Crockett 2024), potentially leading to erroneous conclusions that influence scientific understanding and policy decisions.

Beyond creating errors in individual research projects, over-reliance on LLMs creates a risk of statistical deskilling in the research community. If researchers increasingly rely on LLMs for statistical decisions without developing their own understanding, the collective statistical literacy of the field could decline over time. This would create a dangerous dependency on tools that lack true statistical reasoning capabilities.

Table 1: Risks and opportunities when using LLMs for ecological statistics.

Risks	Action	Opportunities
AI accelerated p-hacking	Improving reporting standards, honest reporting	Use LLMs to improve reporting; understand sensitivity of results to different ecological model formulations
AI suggests logically flawed statistical or coding approaches	Verification with humans and/or literature	More efficient statistical computing workflows with lower risk of errors
AI overconfidence in its suggestions misleads users	Verification with humans and/or literature	Improve accessibility of statistical analyses to novice users
Deskilling of research community	Use LLMs as learning tool	Bespoke AI tutors; Improved access and relevance of learning resources

Opportunities for Enhanced Statistical Practice

All of the above risks can be turned into strengths with appropriate use and community standards for LLM use in statistics (Table 1).

AI accelerated p-hacking can become AI accelerated sensitivity analysis if different model options are appropriately reported. Using LLMs to rapidly explore alternative analytical approaches is appropriate if the alternatives are reported. This capability could support more robust sensitivity analyses, as researchers can efficiently implement various models to assess how analytical choices influence results. For instance, an ecologist studying species distributions could use an LLM to implement both frequentist and machine learning approaches to the same question, comparing outcomes without investing extensive time in coding each approach from scratch. This could then enable the researcher to report on how ecological model complexity relates to predictive power and overfitting (e.g. Fordham et al. 2018).

An ongoing challenge for statistical ecology is promoting the documentation and reporting of data and code (Culina 2020; Popovic et al. 2024; Jenkins et al. 2023). Better reporting standards are neccessary for transparency and replication of studies, as well as allowing combatting the bias towards publishing only significant results. LLMs can assist here by making the mundane task of documenting code more efficient, as well as refactoring bespoke code to meet disciplinary conventions. In our experience, confidence in how their code looks to experts is one of the main reasons researchers with lower levels of experience in statistical coding do not share their code publically (though we know of no specific surveys quantifying this effect). LLMs can help to clean up code and give researcher confidence about sharing their code. This advantage addresses a persistent challenge in the field, where code is typically not shared (Culina 2020; Popovic et al. 2024).

Ecological analysis, and data science more generally, is best done in multi-disciplinary teams (Gibert et al. 2018). LLMs can democratize access to statistical expertise, providing researchers who lack ready access to statistical collaborators with guidance on appropriate methods and implementation strategies. Likewise, they can assist with scientific coding (Jansen et al. 2025). This democratization is particularly valuable in resource-constrained settings or for early-career researchers still developing their statistical abilities. For example, researchers from institutions without dedicated statistical support can leverage LLMs to explore analytical options that might otherwise be inaccessible. There is still the risk that inexperienced researchers are misled by LLM overconfidence, and this needs to be addressed by treating all suggestions with skepticism and validating statistical logic with experts and/or the literature.

Thoughtful use of LLMs as tutors can address the risk of deskilling (Ellis and Slade 2023). For example, they can turn a tutorial on a statistical approach into a tutorial that is bespoke to a researcher's data. When used as interactive tutors rather than black-box solution providers, these models can enhance researchers' statistical understanding by explaining concepts, suggesting relevant literature, and demonstrating proper implementation techniques. Human-led coursework on ecological statistics is still essential (Touchon and McCoy 2016), as suggestions of AI tutors need still human validation.

Toward Effective Human-AI Statistical Partnerships

The challenge is to develop workflows that maximize LLMs' strengths while compensating for their weaknesses. This requires providing sufficient context about research questions, data characteristics, and analytical constraints to guide the model toward appropriate statistical recommendations. It also involves maintaining oversight of model outputs, particularly for decisions requiring deeper statistical understanding such as model formulation, assumption checking, and result interpretation.

The opportunity lies in developing a statistical workflow that combines human expertise with LLM capabilities. In this workflow, researchers maintain responsibility for statistical decisions while using LLMs to implement analyses efficiently, explore options, and enhance documentation. This human-AI partnership is a middle path between complete automation and traditional manual implementation—leveraging the efficiency and consistency of LLMs while preserving the critical judgment and domain expertise of human researchers. The key to this partnership is effective prompting—providing LLMs with the context, constraints, and guidance needed to generate high-quality statistical advice and code that advances rather than compromises statistical rigor in environmental research.

LLM Overview

To develop effective prompting strategies, it's essential to understand how LLMs function. At their core, LLMs are prediction engines that generate text one token at a time based on patterns learned during training. A token is roughly equivalent to part of a word, a word, or a common phrase.

Several key parameters influence LLM behavior (Boonstra 2024):

1. **Temperature**: Controls randomness in token prediction. Lower temperatures (closer to 0) make responses more deterministic and conservative, while higher temperatures (greater than 1.0) increase creativity but potentially reduce reliability. For statistical applications, lower temperatures typically produce more consistent and conventional recommendations.

2. **Context window**: The amount of text an LLM can consider when generating a response. Current LLMs have context windows typically in the range from 100,000 to 2,000,000 tokens. Larger context windows allow for including more detailed information about data, research questions, and statistical requirements.

3. **Model complexity**: Different models have varying capabilities based on their size, training data, and architecture. More complex models (e.g., Claude-4.0-Opus vs. Claude-4.0-Sonnet) generally provide more nuanced statistical guidance but at higher computational and financial cost.

4. **System prompt**: Sets the overall context and constraints for the LLM. This "behind-the-scenes" instruction shapes how the model responds to user queries and can significantly impact statistical advice quality.

5. **AI assistant, AI programmer pair**: Software that assists a user to interact with an LLM. Examples include Github Copilot and Claude Code. This software manages user interactions, including setting the system message (which may be proprietary information) and managing the context window.

6. **Tools and MCP**: Tools allow LLMs to perform tasks. Examples include running R scripts, searching the internet and downloading online data. A common standard for tool definition is the Model Context Protocol (MCP).

7. **Agents**: Agents are software systems that allow LLMs to iteratively develop their own task, with or without human supervision. For instance, an agent can have a tool

allowing it to run and read terminal commands. This lets the agent write R scripts, run them, check for errors, and correct iteratively. Agents are most commonly used within AI assistant software like Github Copilot, though there is some development of agents for bespoke statistical problems (Jansen et al. 2025).

Prompting Guidelines Best Practices

There are now many formal evaluations of LLMs for statistical advice. However, many of these studies are not replicable and do not follow statistical best practice. For instance, they do not provide the prompts they used, do not replicate prompts (LLM responses differ ever time) or use statistics that inflate estimates of effect size (Gallo et al. 2024). Here we summarize the handful of evaluations that provide sufficient information to assess the scientific credibility of their claims.

The key findings of these studies are that more accurate responses are obtained when:

- Role prompting is used, e.g. `You are an expert in the statistical analysis of ecological data` (Jansen et al. 2025)
- Examples and reference material are included (also called one-shot or few-shot prompting) (Zhu et al. 2024)
- Context about the data collection process is included (Zhu et al. 2024)
- The data are attached as part of the prompt (Jansen et al. 2025; Zhu et al. 2024)

Providing examples that pair types of statistical questions with appropriate solutions one of the most effective approaches to improve the precision of responses (Sivarajkumar et al. 2024; Zhu et al. 2024). This approach should be used wherever possible, however, accurate examples may not be readily available to the novice statistician. Reference material can also be provided in place of examples. For example, a user could attach a blog or package vignette that illustrates the application of an analysis to answering a research question.

A further tactic, ‘chain of thought’ reasoning, has mixed success. Chain of thought reasoning encourages the model to structure its prompt as in a step-by-step way and tends to improve the quality of reasoning (Wei et al. 2022). It can be as sample as adding to a prompt `Use chain of thought reasoning`. Its utility has mixed performance for statistical analyses (Jansen et al. 2025; Zhu et al. 2024). Chain of thought prompting is best combined with prompts that include the data and measurement context.

Prompts that say what to do, rather than what not to do, are generally also considered to be more effective (Boonstra 2024).

Recognize different steps in workflows

It is helpful to separate statistical workflows into distinct components that align with LLM strengths and limitations:

1. **Select statistical approach**: Determining appropriate statistical methods for research questions
2. **Plan implementation**: Designing the analytical workflow and code structure
3. **Write code**: Writing the actual code to implement analyses
4. **Guidance on Interpretation**: Understanding and reporting results

We deal with steps 1-3 here (Figure 2). The credibility of statistical interpretation needs further empirical evaluation, so we leave that for future studies.

LLMs perform differently across these components. They excel at code generation and implementation planning but are less reliable for selecting appropriate statistical approaches or interpreting complex results.

LLMs can be used across all of these steps, but we recommend that each step is treated separately. This encourages informed decision making and avoids making decisions on the fly. For instance, it is better to design the statistical analysis prior to setting an agent up to automate the implementation of that analysis.

The separation of workflow steps also helps prevent overreliance on LLMs for statistical decisions.

Below we will work through three examples that align to each of the steps above.

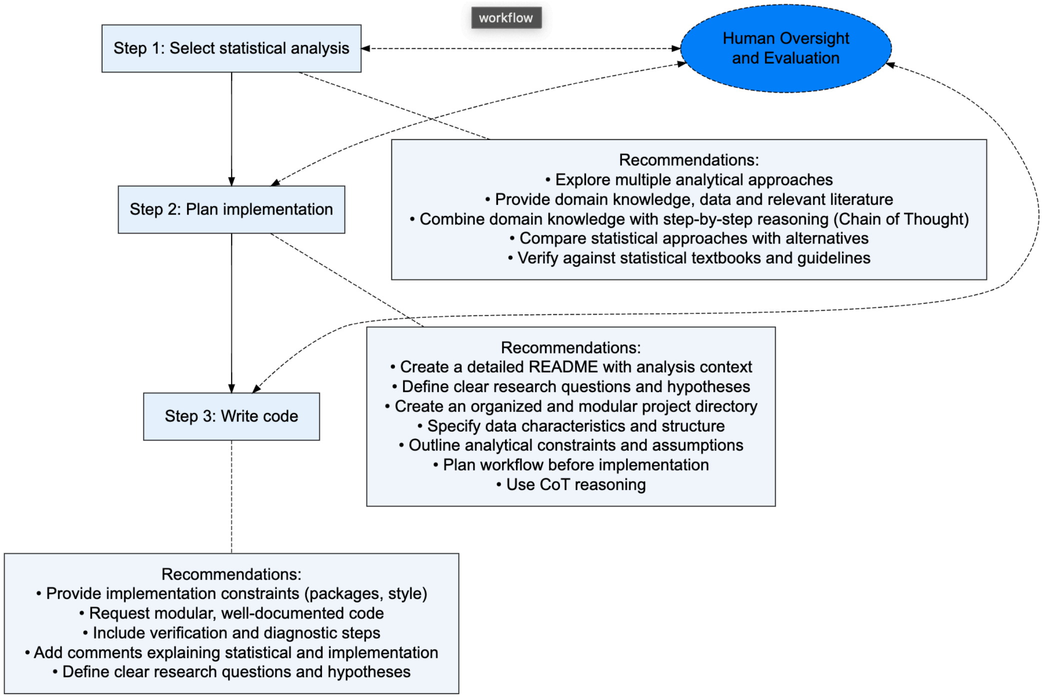
workflow
Step 1: Select statistical analysis
Human Oversight and Evaluation
Step 2: Plan implementation
Step 3: Write code
Recommendations:
• Explore multiple analytical approaches
• Provide domain knowledge, data and relevant literature
• Combine domain knowledge with step-by-step reasoning (Chain of Thought)
• Compare statistical approaches with alternatives
• Verify against statistical textbooks and guidelines
Recommendations:
• Create a detailed README with analysis context
• Define clear research questions and hypotheses
• Create an organized and modular project directory
• Specify data characteristics and structure
• Outline analytical constraints and assumptions
• Plan workflow before implementation
• Use CoT reasoning
Recommendations:
• Provide implementation constraints (packages, style)
• Request modular, well-documented code
• Include verification and diagnostic steps
• Add comments explaining statistical and implementation
• Define clear research questions and hypotheses

Figure 2 Recommended workflow for using LLMs in statistical analysis, showing the four key steps alongside specific recommendations for effectively leveraging LLMs at each stage while maintaining scientific rigor.

Example Step 1. Statistical approach selection

Selecting the appropriate statistical approach is a critical first step in any analysis workflow. Our example prompt demonstrates several key strategies that significantly improve LLM performance for this task:

```
You are an expert in ecological statistics with the R program.
I want to statistically test the dependence of fish abundance on coral cover.
I have observations of coral cover (continuous percentage) and fish abundance
(count of number of fish). Observations were made at 49 different locations.
Observations were made with standardized surveys, so the area surveyed at eac
h site was the same.
Sites are spatially clustered into different regions. Provide me with several
options for statistical approaches would be appropriate for answering my rese
arch question. Also include suggestions for verification of statistical assum
ptions and suggestions for visualizations.  Use chain of thought to reason ab
out each approach before providing a final summary.
I've attached the data [data] and a reference on analysis of count data with
ecology [reference].
```

This example prompt integrates multiple evidence-based strategies for obtaining quality statistical advice from LLMs:

1. Clear hypothesis and domain knowledge specification

The prompt clearly states the research question (testing dependence of fish abundance on coral cover) and provides context about the variables (coral cover as continuous percentage, fish abundance as count data). It also requests verification steps and visualizations. This specificity helps the LLM recognize that count-appropriate methods like Poisson GLM would be suitable. Without this context, LLMs often default to inappropriate methods like linear regression for count data.

2. Detailed experimental/observational design context

The prompt includes critical information about the sampling methodology (standardized surveys with equal area across sites) and spatial structure (sites clustered into regions). By explicitly stating that there is spatial structure the LLM is more likely to recommend methods that account for spatial structure in the data, such as mixed-effects models or spatial autocorrelation analyses. This contextual information is crucial for ecological data where spatial and temporal dependencies are common but often overlooked in standard statistical approaches.

3. Attaching data and references

A consistent finding across formal evaluations of LLMs for statistical advice (Jansen et al. 2025; Zhu et al. 2024) is that attaching your data dramatically improves response quality. In our example, we reference attached data and a relevant ecological statistics reference.

4. Chain of Thought reasoning with authoritative sources

The prompt explicitly requests Chain of Thought (CoT) reasoning (Wei et al. 2022): “Use chain of thought to reason about each approach before providing a final summary.” This prompting strategy encourages step-by-step analytical thinking, which is most effective when combined with authoritative references (Zhu et al. 2024).

5. Iterative refinement through follow-up prompts

The initial prompt can be enhanced through follow-up prompts. For instance, you could use a web search tool to gain further reference information:

```
Search the web to find robust recommendations for ecologists to analyze count
data before proceeding with your recommendations.
```

You can also request self-evaluation:

```
Evaluate the robustness of each suggestion on a 1-10 scale and explain
the strengths and limitations of each approach. Use chain of thought to caref
ully think about the appropriateness of each suggestion.
```

Similarly to chain of thought reasoning, self evaluation is most effective if it is paired a reputable reference, because it encourages the LLM to compare its suggestions against established statistical best practices.

6. Prompt bootstrapping

Users can employ “prompt bootstrapping” by starting with simpler prompts and then asking the LLM to help draft more complete prompts:

```
I'm trying to improve a prompt asking an assistant for statistical advice. He
re is the prompt [prompt]. What other information can I provide to improve th
e accuracy of the assistant's response?
```

This meta-prompting approach leverages the LLM’s own capabilities to identify what additional information would be helpful. The user can then edit and refine the suggested prompt to include all relevant context about their research question, data characteristics, and analytical constraints.

Step 2. Plan implementation

LLMs are most effectively if used via an AI assistant like GitHub Copilot. These assistants have access to context from your project files and can edit directly into documents and scripts.

We recommend users carefully organize their workflow and then create a project directory structure that facilitates workflow management. Being organized makes it easier for humans and AI assistants to understand the project.

Our recommendation is to start a new folder on your computer and insert a `readme.md` file into this folder. (we prefer markdown format for this file, see supplemental material). The readme file can be developed iteratively with an LLM’s help. In our readme file we include:

- Research context
- Research aims
- Analysis methodology
- Technology context including R package preferences
- Analysis steps
- Directory structure
- Data file locations and meta-data

The readme can be attached to every prompt to provide project context, as well as give the LLM a memory across different chat sessions.

Information from the first step (Plan analysis) can be entered into this readme file. Then we would use the LLM to assist in filling out the other sections. An initial prompt could look like:

```
Help me plan the steps to complete this analysis. This should include a serie
s of scripts that we will need to make for each step of the analysis. It shou
ld also include a plan for how to structure the project directory. Create mod
ular scripts for this analysis with separate files for data preparation, mode
l fitting, diagnostics, and visualization. Save data files for intermediate s
teps. Use chain of thought reasoning to think carefully abotu each step.
```

Scripts should be kept short and modular. When helping ecologists with code with often find that they write long scripts that span data wrangling to plotting and may not even evaluate in a top to bottom order! Keeping your project modular organized allows for agents to easily navigate your project and for you to more easily manage the size of prompt attachments to fit within the context window. It also lets you point precisely to files that may need attention. This structure not only improves immediate code quality but also enhances long-term project sustainability and knowledge transfer.

You can then create the project directory, or have an AI agent do this for you (`Create the directory structure in readme.md`).

For example, a typical project directory may look like this:

```
my-project/
├── README.md
├── .gitignore
├── Scripts/ # R code
│   ├── 01_data-prep.R
│   ├── 02_data-analysis.R
│   └── 03_plots.R
├── Shared/
│   ├── Outputs/
│   │   ├── Figures/
│   │   ├── data-prep/
│   │   └── model-objects/
│   ├── Data/
│   └── Manuscripts/
```

Step 3. Write code

If the data, methodology and plan are carefully documented in a readme file, starting the analysis is as simple as:

```
Start on the first step of this analysis [readme.md]. Review and update the r
eadme when you complete this step.
```

Starting a new chat for each discrete task will reduce syntax errors (Jansen et al. 2025) (but note that web platforms may retain a memory across prompts). Keep the readme.md as a memory across chat sessions.

Further precision can be gained by adding additional context to the prompt or readme.md file. For example, you can provide implementation constraints in the `Technology context` section:

```
Implement this analysis using the tidyverse ecosystem and INLA for Bayesian
modeling. Follow tidyverse style guidelines and prioritize code readability.
```

It is better to say what to do, rather than what not to do (Boonstra 2024).

The rate of syntax errors can also be reduced by more than three times by allowing an agent to iterate through running and debugging (Jansen et al. 2025). Code should be carefully reviewed by a human to ensure logical correctness, agents often produce code that is syntactically correct, but logically flawed (Jansen et al. 2025). The LLM can assist (but not replace) the user with a prompt such as:

```
Review my choice of analysis from the perspective of a peer-reviewer in an ec
ological journal.
```

Writing the code may not be the final step. Often when we analyse data new issues emerge, or the process of analysis helps us identify logical inconsistencies in a method. The steps above can be iterated, just be clear which step you are operating in.

Discussion and Conclusion

Large language models represent both opportunity and challenge for statistical practice in environmental sciences. When used thoughtfully with effective prompting strategies, they can enhance analytical workflows, improve code quality, and potentially address longstanding issues in statistical implementation. However, uncritical reliance on LLMs risks perpetuating or even amplifying existing problems in statistical practice.

The prompting guidelines presented in this perspective provide a framework for leveraging LLMs while maintaining statistical rigor. By separating workflows into components that align with LLM strengths and limitations, providing appropriate context and constraints, and maintaining human oversight of critical decisions, researchers can harness these tools while mitigating their risks. Our recommendations are common sense and align closely with best-practices for ecological statistics and open data in general (Jenkins et al. 2023; Popovic et al. 2024) - what is better for human reproducibility is also good for LLMs efficiency.

Several principles emerge from our perspective that can guide effective LLM use in ecological statistics. First and foremost, researchers must maintain critical thinking when using these tools. LLMs should complement rather than replace research expertise (Messeri and Crockett 2024). Statistical suggestions from LLMs require careful evaluation against domain knowledge and established statistical principles. This critical evaluation becomes particularly important when analyzing ecological data with complex dependencies that may not be adequately captured in standard statistical approaches.

Providing rich context dramatically improves LLM performance for statistical tasks. Our experience suggests that most researchers are under-utilizing the potential for detailed prompts (Boonstra 2024). Modern LLMs have context windows that allow hundreds of thousands of tokens—equivalent to several theses worth of text. This capacity enables researchers to include comprehensive information about research questions, data characteristics, and analytical constraints. Studies have consistently shown that LLM statistical guidance improves when provided with this detailed contextual information (Zhu et al. 2024; Jansen et al. 2025). Researchers should leverage this capability by developing more comprehensive prompts that include relevant background information, data summaries, and specific analytical requirements.

As LLMs become increasingly integrated into research practices, developing “LLM literacy” becomes an essential skill for environmental scientists. This literacy encompasses understanding how these models work, recognizing their limitations, and mastering effective interaction strategies (Messeri and Crockett 2024). Researchers need to develop the ability to craft effective prompts, critically evaluate model outputs, and understand when human expertise should take precedence over model suggestions. Educational institutions and professional organizations should consider incorporating LLM literacy into statistical training programs for environmental scientists.

The rapid evolution of LLM capabilities suggests that their role in statistical workflows will only increase. Current models already show impressive performance in code generation and implementation planning, and future models may address some of the limitations identified in statistical reasoning. However, the fundamental nature of LLMs as prediction engines rather than reasoning systems means that human oversight will remain essential for ensuring statistical quality.

Research Needs

Research is needed to inform the appropriate use of LLMs in ecological statistics. We identify several priorities. First, we need more formal evaluations of LLM statistical performance for ecological datasets and problems. Ecological data presents unique problems for statistical analysis (Gilbert et al. 2024) and it is not yet clear how reliable LLMs advice on these problems. Ensuring that researchers report use of generative AI is important both for transparency, but to enable evaluation of how the tools are influencing our field.

Second, we need to develop prompting templates that novice analysts and statistical coders can use to reliably develop their analyses (Jansen et al. 2025). These could include the recommendations above, as well as reference material that is attuned to specific types of ecological data.

Finally, LLMs have certain biases (Messeri and Crockett 2024; Ji et al. 2025), but it is not yet clear if there are important implications for ecological statistics. One bias of LLMs is towards overconfidence in their own answers being correct (Zhou et al. 2024). Researchers should therefore independently validate all analyses against reputable sources. Further research is needed to understand how LLM use may bias analyses in harmful ways.

We have argued that LLMs present both risks and opportunities to the quality of ecological statitisics. By addressing these research needs and adopting thoughtful prompting strategies, environmental scientists can apply large language models to enhance rather than compromise statistical quality. The future of environmental statistics likely lies not in choosing between human expertise and artificial intelligence, but in developing effective partnerships that leverage the unique strengths of each.

Acknowledgements

CJB was supported by a Future Fellowship (FT210100792) from the Australian Research Council.

References

Arif, Suchinta, and M. Aaron MacNeil. 2022. “Predictive Models Aren’t for Causal Inference.” *Ecology Letters* 25 (8): 1741–45. https://doi.org/https://doi.org/10.1111/ele.14033.

Bolker, Benjamin M. 2024. “Multimodel Approaches Are Not the Best Way to Understand Multifactorial Systems.” *Entropy* 26 (6). https://doi.org/10.3390/e26060506.

Boonstra, Lee. 2024. “Prompt Engineering.” Google. https://www.kaggle.com/whitepaper-prompt-engineering.

Brown, Christopher J., David S. Schoeman, William J. Sydeman, Keith Brander, Lauren B. Buckley, Michael Burrows, Carlos M. Duarte, et al. 2011. “Quantitative Approaches in Climate Change Ecology.” *Global Change Biology* 17 (12): 3697–3713. https://doi.org/https://doi.org/10.1111/j.1365-2486.2011.02531.x.

Culina, Ilona AND Evans, Antica AND van den Berg. 2020. “Low Availability of Code in Ecology: A Call for Urgent Action.” *PLOS Biology* 18 (7): 1–9. https://doi.org/10.1371/journal.pbio.3000763.

Ellis, Amanda R, and Emily Slade. 2023. “A New Era of Learning: Considerations for ChatGPT as a Tool to Enhance Statistics and Data Science Education.” *Journal of Statistics and Data Science Education* 31 (2): 128–33.

Fordham, Damien A, Cleo Bertelsmeier, Barry W Brook, Regan Early, Dora Neto, Stuart C Brown, Sébastien Ollier, and Miguel B Araújo. 2018. “How Complex Should Models Be? Comparing Correlative and Mechanistic Range Dynamics Models.” *Global Change Biology* 24 (3): 1357–70.

Forstmeier, Wolfgang, Eric-Jan Wagenmakers, and Timothy H. Parker. 2017. “Detecting and Avoiding Likely False-Positive Findings – a Practical Guide.” *Biological Reviews* 92 (4): 1941–68. https://doi.org/https://doi.org/10.1111/brv.12315.

Fraser, Tim AND Nakagawa, Hannah AND Parker. 2018. “Questionable Research Practices in Ecology and Evolution.” *PLOS ONE* 13 (7): 1–16. https://doi.org/10.1371/journal.pone.0200303.

Gallo, Robert J, Michael Baiocchi, Thomas R Savage, and Jonathan H Chen. 2024. “Establishing Best Practices in Large Language Model Research: An Application to Repeat Prompting.” *Journal of the American Medical Informatics Association* 32 (2): 386–90. https://doi.org/10.1093/jamia/ocae294.

Gibert, Karina, Jeffery S. Horsburgh, Ioannis N. Athanasiadis, and Geoff Holmes. 2018. “Environmental Data Science.” *Environmental Modelling & Software* 106: 4–12. https://doi.org/https://doi.org/10.1016/j.envsoft.2018.04.005.

Gilbert, Neil A., Bruna R. Amaral, Olivia M. Smith, Peter J. Williams, Sydney Ceyzyk, Samuel Ayebare, Kayla L. Davis, Wendy Leuenberger, Jeffrey W. Doser, and Elise F. Zipkin. 2024. “A Century of Statistical Ecology.” *Ecology* 105 (6): e4283. https://doi.org/https://doi.org/10.1002/ecy.4283.

Jansen, Jacqueline A, Artür Manukyan, Nour Al Khoury, and Altuna Akalin. 2025. “Leveraging Large Language Models for Data Analysis Automation.” *PloS One* 20 (2): e0317084.

Jenkins, Gareth B, Andrew P Beckerman, Céline Bellard, Ana Benítez-López, Aaron M Ellison, Christopher G Foote, Andrew L Hufton, et al. 2023. “Reproducibility in Ecology and Evolution: Minimum Standards for Data and Code.” *Ecology and Evolution* 13 (5): e9961.

Ji, Wenlong, Weizhe Yuan, Emily Getzen, Kyunghyun Cho, Michael I Jordan, Song Mei, Jason E Weston, Weijie J Su, Jing Xu, and Linjun Zhang. 2025. “An Overview of Large Language Models for Statisticians.” *arXiv Preprint arXiv:2502.17814*.

Kendall, Bruce E, Masami Fujiwara, Jasmin Diaz-Lopez, Sandra Schneider, Jakob Voigt, and Sören Wiesner. 2019. “Persistent Problems in the Construction of Matrix Population Models.” *Ecological Modelling* 406: 33–43.

Liao, Zhehui, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X Zhang. 2024. “LLMs as Research Tools: A Large Scale Survey of Researchers’ Usage and Perceptions.” *arXiv Preprint arXiv:2411.05025*.

Messeri, Lisa, and MJ Crockett. 2024. “Artificial Intelligence and Illusions of Understanding in Scientific Research.” *Nature* 627 (8002): 49–58.

O’Hara, Robert B., and D. Johan Kotze. 2010. “Do Not Log-Transform Count Data.” *Methods in Ecology and Evolution* 1 (2): 118–22. https://doi.org/https://doi.org/10.1111/j.2041-210X.2010.00021.x.

Onan, Aytuğ, and Hesham Alhumyani. 2024. “Assessing the Impact of Prompt Strategies on Text Summarization with Large Language Models.” In *International Conference on Computer Applications in Industry and Engineering*, 41–55. Springer.

Popovic, Gordana, Tanya Jane Mason, Szymon Marian Drobniak, Tiago André Marques, Joanne Potts, Rocío Joo, Res Altwegg, et al. 2024. “Four Principles for Improved Statistical Ecology.” *Methods in Ecology and Evolution* 15 (2): 266–81. https://doi.org/https://doi.org/10.1111/2041-210X.14270.

Shoemaker, Kevin T., and Kevin J. Loope. 2025. “We Need Better Ways to Re-Evaluate Conservation Policies When They’re Founded on Flawed Research.” *Proceedings of the National Academy of Sciences* 122 (19): e2426166122. https://doi.org/10.1073/pnas.2426166122.

Sivarajkumar, Sonish, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2024. “An Empirical Evaluation of Prompting Strategies for Large Language Models in Zero-Shot Clinical Natural Language Processing: Algorithm Development and Validation Study.” *JMIR Medical Informatics* 12: e55318.

Touchon, Justin C, and Michael W McCoy. 2016. “The Mismatch Between Current Statistical Practice and Doctoral Training in Ecology.” *Ecosphere* 7 (8): e01394.

Warton, David I., Mitchell Lyons, Jakub Stoklosa, and Anthony R. Ives. 2016. “Three Points to Consider When Choosing a LM or GLM Test for Count Data.” *Methods in Ecology and Evolution* 7 (8): 882–90. https://doi.org/https://doi.org/10.1111/2041-210X.12552.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” *Advances in Neural Information Processing Systems* 35: 24824–37.

Zhou, Lexin, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. 2024. “Larger and More Instructable Language Models Become Less Reliable.” *Nature* 634 (8032): 61–68.

Zhu, Yizhang, Shiyin Du, Boyan Li, Yuyu Luo, and Nan Tang. 2024. “Are Large Language Models Good Statisticians?” *arXiv Preprint arXiv:2406.07815*.