Using ChatGPT with Confidence for Biodiversity-Related Information Tasks

Michael J Elliott‡, José AB Fortes‡

‡ University of Florida, Gainesville, United States of America

Corresponding author: Michael J Elliott (mielliott@ufl.edu)

Abstract

Recent advancements in conversational Artificial Intelligence (AI), such as OpenAI's Chat Generative Pre-Trained Transformer (ChatGPT), present the possibility of using large language models (LLMs) as tools for retrieving, analyzing, and transforming scientific information. We have found that ChatGPT (GPT 3.5) can provide accurate biodiversity knowledge in response to questions about species descriptions, occurrences, and taxonomy, as well as structure information according to data sharing standards such as Darwin Core. A rigorous evaluation of ChatGPT's capabilities in biodiversity-related tasks may help to inform viable use cases for today's LLMs in research and information workflows. In this work, we test the extent of ChatGPT's biodiversity knowledge, characterize its mistakes, and suggest how LLM-based systems might be designed to complete knowledge-based tasks with confidence.

To test ChatGPT's biodiversity knowledge, we compiled a question-and-answer test set derived from Darwin Core records available in Integrated Digitized Biocollections (iDigBio). Each question focuses on one or more Darwin Core terms to test the model's ability to recall species occurrence information and its understanding of the standard. The test set covers a range of locations, taxonomic groups, and both common and rare species (defined by the number of records in iDigBio). The results of the tests will be presented. We also tested ChatGPT on generative tasks, such as creating species occurrence maps. A visual comparison of the maps with iDigBio data shows that for some species, ChatGPT can generate fairly accurate representationsof their geographic ranges (Fig. 1).

ChatGPT's incorrect responses in our tests show several patterns of mistakes. First, responses can be self-conflicting. For example, when asked "Does *Acer saccharum* naturally occur in Benton, Oregon?", ChatGPT responded "YES, *Acer saccharum* DOES NOT naturally occur in Benton, Oregon". ChatGPT can also be misled by semantics in species names. For *Rafinesquia neomexicana*, the word "neomexicana" leads ChatGPT to believe that the species primarily occurs in New Mexico, USA. ChatGPT may also confuse species, such as when attempting to describe a lesser-known species (e.g., a rare bee) within the same genus as a better-known species. Other causes

of mistakes include hallucination (Ji et al. 2023), memorization (Chang and Bergen 2023), and user deception (Li et al. 2023).

Some mistakes may be avoided by prompt engineering, e.g., few-shot prompting (Chang and Bergen 2023) and chain-of-thought prompting (Wei et al. 2022). These techniques assist Large Language Models (LLMs) by clarifying expectations or by guiding recollection. However, such methods cannot help when LLMs lack required knowledge. In these cases, alternative approaches are needed.

A desired reliability can be theoretically guaranteed if responses that contain mistakes are discarded or corrected. This requires either detecting or predicting mistakes. Sometimes mistakes can be ruled out by verifying responses with a trusted source. For example, a trusted specimen record might be found that corroborates the response. The difficulty, however, is finding such records programmatically; e.g., using iDigBio and Global Biodiversity Information Facility's (GBIF) search Application Programming Interfaces (APIs) requires specifying indexed terms that might not appear in an LLM's response. This presents a secondary problem for which LLMs may be well suited. Note that with presence-only data, it can be difficult to disprove presence claims or prove absence claims.

Besides verification, mistakes may be predicted using probabilistic methods. Formulating mistake probabilities often relies on heuristics. For example, variability in a model's responses to a repeated query can be a sign of hallucination (Manakul et al. 2023). In practice, both probabilistic and verification methods may be needed to reach a desired reliability. LLM outputs that can be verified may be directly accepted (or discarded), while others are judged by estimating mistake probabilities. We will consider a set of heuristics and verification methods, and report empirical assessments of their impact on ChatGPT's reliability.

Keywords

LLMs, NLP, AI, verification, uncertainty quantification,

Presenting author

Michael J Elliott

Presented at

TDWG 2023

Funding program

The research reported in this work was funded by grants from the National Science Foundation (DBI 2027654) and the AT&T Foundation.

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Chang TA, Bergen BK (2023) Language Model Behavior: A Comprehensive Survey. arXiv https://doi.org/10.48550/arXiv.2303.11504
- Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, Ishii E, Bang YJ, Madotto A, Fung P (2023) Survey of Hallucination in Natural Language Generation. ACM Computing Surveys 55 (12): 1-38. https://doi.org/10.1145/3571730
- Li K, Patel O, Viégas F, Pfister H, Wattenberg M (2023) Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. arXiv https://doi.org/10.48550/arxiv.2306.03341
- Manakul P, Liusie A, Gales MF (2023) SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. arXiv https://doi.org/10.48550/arxiv.2303.08896
- Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le QV, Zhou D (2022) Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv https://doi.org/10.48550/arxiv.2201.11903

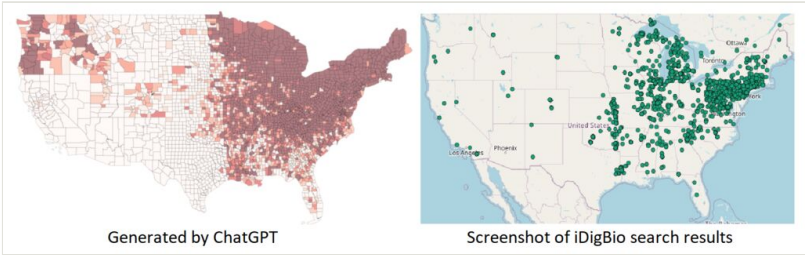

Figure 1.

Occurrence maps for *Acer saccharum*. **Left** - Generated generated by GPT 3.5. **Right** - A screenshot of https://www.idigbio.org/portal/search.