A publishing infrastructure for AI-assisted academic authoring

This manuscript (permalink) was automatically generated from greenelab/manubot-gpt-manuscript@1a77711 on December 25, 2022.

Authors

- **Milton Pividori**
  0000-0002-3035-4403 · miltondp · miltondp
  Department of Genetics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA

- **Casey S. Greene** ✉
  0000-0001-8713-9213 · cgreene · GreeneScientist
  Center for Health AI, University of Colorado School of Medicine, Aurora, CO 80045, USA; Department of Biomedical Informatics, University of Colorado School of Medicine, Aurora, CO 80045, USA

✉ — Correspondence possible via GitHub Issues or email to Casey S. Greene <casey.s.greene@cuanschutz.edu>.

Abstract

Academics often communicate through scholarly manuscripts. These manuscripts describe new advances, summarize existing literature, or argue for changes in the status quo. Writing and revising manuscripts can be a time-consuming process. Large language models are bringing new capabilities to many areas of knowledge work. We integrated the use of large language models into the Manubot publishing ecosystem. Users of Manubot can run a workflow, which will trigger a series of queries to OpenAI's language models, produce revisions, and create a timestamped set of suggested revisions. Given the amount of time that researchers put into crafting prose, we expect this advance to radically transform the type of knowledge work that academics perform.

Introduction

The manuscript pre-dates the invention of printing by thousands of years, but the practice of producing exclusively scientific journals only started roughly 350 years ago [1]. The implementation of external peer review varies by journal but for many is less than 100 years old [2]. To date, most manuscripts have been written by humans or teams of humans working together to describe scholarly advances.

Modern scholarly manuscripts often describe new advances, summarize existing literature, or argue for changes in the status quo. However, writing and revising can be a time-consuming process. Academics can sometimes be long-winded in getting to key points, making writing more impenetrable to their audience [3].

Modern computing capabilities and the widespread availability of text, images, and other data on the internet has laid the foundation for artificial intelligence (AI) models with many parameters. Large language models, in particular, are opening the floodgates to new technologies with the capability to transform how society operates [4]. The GPT-3 model, with its 175 billion parameters, has demonstrated strong performance on many tasks [5].

We developed a software publishing platform that imagines a future where authors co-write their manuscripts with the support of large language models. We used, as a base, the Manubot platform for scholarly publishing [6]. Manubot was designed as an end-to-end publishing platform for scholarly writing for both individual and large-collaborative projects. It has been used for collaborations of approximately 50 authors writing hundreds of pages of text reviewing progress during the COVID19 pandemic [7]. We developed a new workflow that parses the manuscript, uses a large language model with section-specific custom prompts to revise the manuscript, and then creates a set of suggested changes to reach the revised state. Changes are presented to the user through the GitHub interface for author review and integration into the published document.

Methods

Results

Conclusions

References

1. **A history of scientific & technical periodicals: the origins and development of the scientific and technical press, 1665-1790**
   David A Kronick
   Scarecrow Press (1976)
   ISBN: 9780810808447

2. **The history of the peer-review process**
   Ray Spier
   Trends in Biotechnology (2002-08) https://doi.org/d26d8b
   DOI: 10.1016/s0167-7799(02)01985-6 · PMID: 12127284

3. **How to write a first-class paper**
   Virginia Gewin
   Nature (2018-02-28) https://doi.org/ggh63n
   DOI: 10.1038/d41586-018-02404-4

4. **Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models**
   Alex Tamkin, Miles Brundage, Jack Clark, Deep Ganguli
   arXiv (2021-02-05) https://arxiv.org/abs/2102.02503

5. **Language Models are Few-Shot Learners**
   Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, … Dario Amodei
   arXiv (2020-07-24) https://arxiv.org/abs/2005.14165

6. **Open collaborative writing with Manubot**
   Daniel S Himmelstein, Vincent Rubinetti, David R Slochower, Dongbo Hu, Venkat S Malladi, Casey S Greene, Anthony Gitter
   PLOS Computational Biology (2019-06-24) https://doi.org/c7np
   DOI: 10.1371/journal.pcbi.1007128 · PMID: 31233491 · PMCID: PMC6611653

7. **An Open-Publishing Response to the COVID-19 Infodemic.**
   Halie M Rando, Simina M Boca, Lucy D'Agostino McGowan, Daniel S Himmelstein, Michael P Robson, Vincent Rubinetti, Ryan Velazquez, Casey S Greene, Anthony Gitter
   ArXiv (2021-09-17) https://www.ncbi.nlm.nih.gov/pubmed/34545336
   PMID: 34545336 · PMCID: PMC8452106