A publishing infrastructure for AI-assisted academic authoring

This manuscript (permalink) was automatically generated from greenelab/manubot-gpt-manuscript@2af1312 on December 27, 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

We implemented the AI-based revision infrastructure in Manubot [6]. Manubot takes Markdown as input and produces HTML, PDF, or other pandoc-supported formats as output. It includes a robust cite-by-persistent-identifier infrastructure. Its workflows are implemented in continuous integration software (Appveyor, GitHub Actions, etc). Most workflows run with each commit.

We used the OpenAI API for access to large language models, with a focus on the completion endpoints. This API incurs a cost with each run that depends on manuscript length. Because of this

cost, we implemented our workflow in GitHub actions, making it triggerable by the user. The user can select the model that they wish to use, allowing costs to be tuned. With the most complex model, `text-davinci-003`, the cost per run is under $0.50 for many manuscripts.

When the user triggers the action, the manuscript is parsed by section and then by paragraph, passed to the model along with a set of custom prompts, returned, reformatted, and output. Our workflow then uses the GitHub API to generate a new pull request, allowing the user to review and, if desired, modify the output before merging. This workflow allows text to be attributed either to the initial user or to the language model, which may be important in the event that future legal decisions alter the copyright landscape around the outputs of generative models.

Results

We used this infrastructure to revise an existing manuscript as well as to author a new one. We back-ported the changes in Manubot to a manuscript describing the Clustermatch Correlation Coefficient (CCC) [8]. The CCC was designed to capture both linear and non-linear relationships between variables. The CCC manuscript describes its use, in particular with gene expression data.

The abstract of the CCC manuscript before revision had a Flesh-Kincaid readability score of X and a grade level of Y. > PREVIOUS_VERSION

After suggested revisions, the readability score was X and the grade level was Y and read as follows: > NEW_VERSION

The full manuscript before AI-based revision is available at [link], and the revised version is available at [new_link]. We noticed that the model has difficulty with the Manubot citation style, which may lead to some references becoming incorrect. This pipeline is not fully automated: authors will need to review changes and verify the output.

We also used this framework in the context of authoring a new manuscript that described a publishing infrastructure that implemented large language models to suggest revisions. The abstract before revisions had a Flesh-Kincaid readability score of X and a grade level of Y and read as follows: > 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.

After suggested revisions, abstract had a Flesh-Kincaid readability score of X and a grade level of Y and read as follows: > NEW_VERSION

Conclusions

We implemented AI-based models into publishing infrastructure. While most manuscripts have been written by humans, the process is time consuming and academic writing can be difficult to parse. We sought to develop a technology that academics could use to make their writing more understandable without changing the fundamental meaning. This work lays the foundation for a future where

academic manuscripts are constructed by a process that incorporates both human and machine authors.

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

8. **An efficient not-only-linear correlation coefficient based on machine learning**
   Milton Pividori, Marylyn D Ritchie, Diego H Milone, Casey S Greene
   Cold Spring Harbor Laboratory (2022-06-17) https://doi.org/gqcvbw
   DOI: 10.1101/2022.06.15.496326