LLMs for Citation Management

Aerith Netzer

Digital Publishing and Repository Librarian

Northwestern University Libraries

aerith.netzer@northwestern.edu

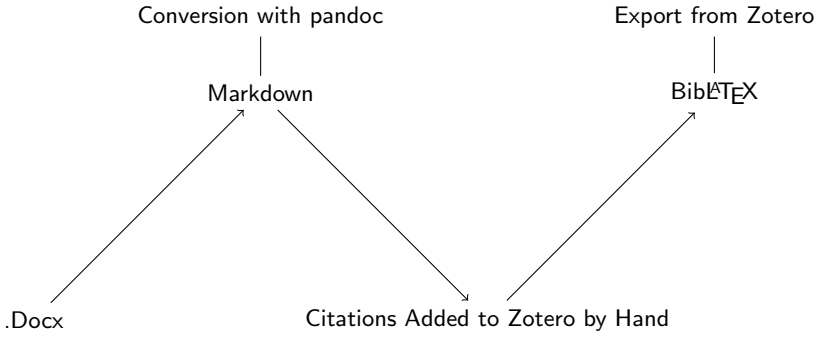
January 17, 2024

This Presentation's LaTeX File is Available on GitHub

- `https://www.github.com/aerithnetzer/OpenAI_Presentation`

The Code Demonstrated in this Presentation is Available on GitHub

- `https://github.com/aerithnetzer/py-bats`



The Current State is Not Scalable

- The current process is time-consuming
- The current process is error-prone
- The current process is not scalable

The Problem of Plaintext Citations

- Plaintext citations are not machine-readable
- Many different citation styles cause confusion
- Citations are often not linked to the cited work

Workaround: OpenAI GPT-3.5

- Parse citations from article submission
- Create a .txt file with the citations in plaintext
- Call the OpenAI API to generate a BibLaTeXfile

The Optimal Solution: BibLaTeX Evangelising

- Create resources that help create .bib files
- Communicate the importance of BibLaTeX to researchers
- Create a .tex file with the citations in BibLaTeX

Demo

Data Creation → LLM Evaluation → LLM Fine-Tuning

Create a synthetic dataset

Out-of-the-box

Customized for citations

Evaluating Large Language Models for Transforming Plaintext Citaitons to Machine-Readable Formats

- Highlight the importance of BibLaTeX for citations and parsing by Google Scholar
- Create a synthetic dataset of plaintext citations and their BibLaTeX equivalents across many citation styles
- Evaluate the effectiveness and economic efficiency of large language models out-of-the-box
- Fine-tune a large language model to create a citation parser and measure its effectiveness and economic efficiency

Implementing Error Correction in Crossref API

- Relatively often, HTML code is returned when BibLaTeXis requested
- Reduces effectiveness of the citation parser
- Error correction could be implemented to decrease confusion