STATEMENT OF WORK

Enhancing Academic Paper Visibility with Intelligent Title Generation

Project lead: PROF. Pavlos Protopapas

Members of Title Tailors

1. Ashish Kumar (ashish28082002.ak@gmail.com)
2. Karthik Rathod (karthikmrathod1999@gmail.com)
3. Swarnava Bhattacharjee (swarnavab98@gmail.com)
4. Vaishnavi M R (vaishnavi.mr.144@gmail.com)

Background and motivation

Attracting readership is vital for research impact, yet intriguing title creation is challenging. With over sixteen million astronomy papers in NASA/ADS (Astrophysics Data System), enhanced visibility presents a major opportunity. Recent AI advancements have the potential for automatic title generation using natural language processing methods. However, these applications even if developed may lack customization, readership data integration and scalability issues. This project aims for developing a platform for researchers to obtain optimized titles for boosting engagement. Leveraging ADS database, the system aims to generate clickbait titles for each paper. These generated titles ought to be more captivating than conventional naming while preserving scientific relevance of the paper. Beyond benefiting individual researchers, the project has far-reaching implications. Enhanced visibility facilitates scientific collaboration and even drawing more attention of the public. Streamlining compelling title creation also allows scholars to focus even better on their disciplines rather than marketing tactics.

Problem statement

The main goal of this project is to develop an intelligent system for generating optimized titles that increase the discoverability and readership of papers in NASA/ADS database. The intended outcome is an easy-to-use platform that boosts the visibility of scholarly publications through data-driven title generation. This project aims to provide the facility to everyone interested, therefore ensuring scalability and adaptability is also required.

Key objectives

The objectives of the project are as follows:

1. To analyse the NASA/ADS Database and user statistics from the APIs.
2. To curate dataset for instruction fine-tuning open source LLM model/s (Astrollama, Mistral, etc).
3. To evaluate the model's robustness on available benchmarks for maximum efficacy.
4. To construct a scalable backend system for handling multiple user requests.
5. To utilise RAG techniques to reconstruct information for the model from the provided input.
6. To design a straightforward and engaging user interface.

Dataset

Data source: NASA/ADS astronomy paper database accessed via search API (https://ui.adsabs.harvard.edu/).

Data description: Contains metadata over sixteen million publications along with paper titles, abstracts, full texts, citations, usage metrics on astronomy.

Key attributes:

- Bibcodes – For extracting the relevant pdfs of the research papers.
- Contents - Texts to train large language models.
- Titles - To establish baseline patterns.
- Abstracts – Abstracts of existing papers.
- Reader counts - To gauge real-world impact.

Relevance to the project: The dataset we are collecting will be based on metrics (read count, citations, classic factor, etc) that we will use to filter out the papers ought to show higher interaction possibly having a significant contribution from its title itself.

Data quality concerns:

- API search limits may constrain to data volume that can be fetched.
- Text/metadata formats may require cleaning.
- Some search results may have missing attributes.

Scope

The scope of this research encompasses summarization of thousands of astronomy papers from the Astrophysics Data System (ADS) database using large language models to enable optimized title recommendations. This involves fine-tuning custom natural language processing models on a corpus of papers to boost engagement while preserving relevance. A scalable web application will be developed for seamless researcher title suggestions based on paper uploads. There will be a focus on CPU-based inference for computational efficiency and cost savings. Throughout, domain expert academics will inform iterative improvements to maximize title impact and discoverability. The priority is a user-friendly interface allowing astronomers to easily submit papers and obtain AI-optimized titles to increase reach. Overall, this connects customized natural language generation, web accessibility, and academic validation to promote astronomy work through enhanced discoverability powered by machine learning.

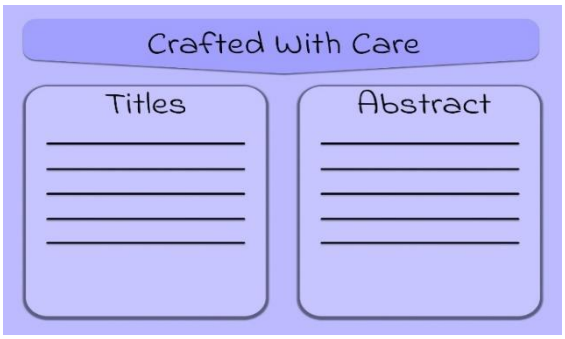
Fun factor

Implementing LLMOps to construct a fully functional web app is quite challenging yet intriguing and fun domain to explore.

Limitations and risks

Limitations of this research include expansion beyond the astronomy paper database from the Astrophysics Data System to incorporate other scientific corpora, which is not feasible currently. There will also be no manual review or editing of each auto-generated title to ensure complete accuracy. This work focuses narrowly on AI-optimization of titles and abstracts only. Finally, developing supplementary tools to enhance abstracts, key points, etc or scientifically evaluating paper quality and contributions is not possible given constraints. With limited computational resources inferencing an LLM might have substantial trade-off between performance and speed.

Mock design of application interface



The core project directives are enhancing title visibility for existing manuscripts through automated, personalized recommendations. While future work may expand across domains and functionalities, the current effort will prioritize title generation capabilities for a defined paper repository via an easy-to-use platform.

Milestones

- Dataset curation
- Fine tuning the LLMs
- Deployment
- Optimising for scalability
- Application interface development

References

1. https://huggingface.co/universeTBD/astrollama [AstroLLaMA]
2. https://blog.wordvice.com/choosing-research-paper-keywords/
3. https://www.nature.com/nature-index/news/five-features-highly-cited-scientific-article
4. https://arxiv.org/abs/2309.06126
5. https://arxiv.org/pdf/2312.00909.pdf