Bridging Scientific Publication Accessibility: LaTeX - Markup - PDF Alignment

Changxu Duan

Technische Universität Darmstadt

July 20, 2024

Outline

LLMs can Read Scientific Publication

Large Language Models (LLMs) like GPT can read and interpret scientific papers effectively due to several key capabilities:

- Advanced Text Understanding and Contextual Analysis
- Information Retrieval and Summarization Skills
- Multimodal Understanding: Table, Math, Figure
  - Text only input
  - Image only input
  - Markdown + LaTeX input

LLMs can Understand LaTeX Better

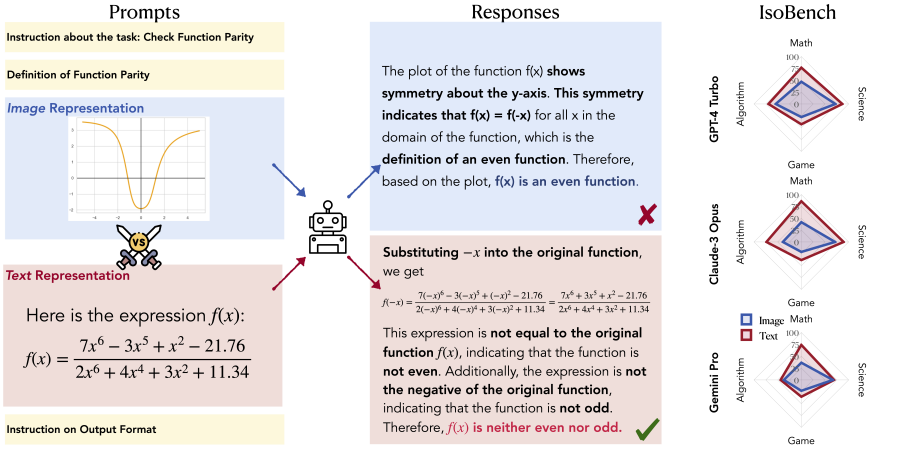


Figure: IsoBench[1] shows LLM can understand LaTeX better than image.

[1] Deqing Fu et al. *IsoBench: Benchmarking Multimodal Foundation Models on Isomorphic Representations*. 2024. arXiv: 2404.01266 [cs.AI]. URL: https://arxiv.org/abs/2404.01266.

LLMs Needs Accessibility

- **PDF with LaTeX Source**: Extract code for Math and Table directly from LaTeX
- **PDF without LaTeX Source**: Convert Math and Table from PDF files into LaTeX
  - PDF Screenshot to LaTeX + Markdown model available[2]
  - Challenges with Current Models:
    - Operates on a page-by-page basis without specific element positioning
    - Requires more precisely aligned data for improvements

[2] Lukas Blecher et al. *Nougat: Neural Optical Understanding for Academic Documents*. 2023. arXiv: 2308.13418 [cs.LG]. URL: https://arxiv.org/abs/2308.13418.

LLMs Needs Accessibility

What should fine-aligned data have?

- Reading order
- Tree structure of page elements
- Whether elements should be ignored (e.g. watermarks, headers and footers)
- The LaTeX code for each non-plaintext element on the PDF

Data Source

LaTeX source codes for scientific publications are downloaded from arXiv

- arXiv hosts a large number of scientific publications
- arXiv ensuring the LaTeX codes are functional and ~~error-free~~
- Documents on arXiv are well-structured

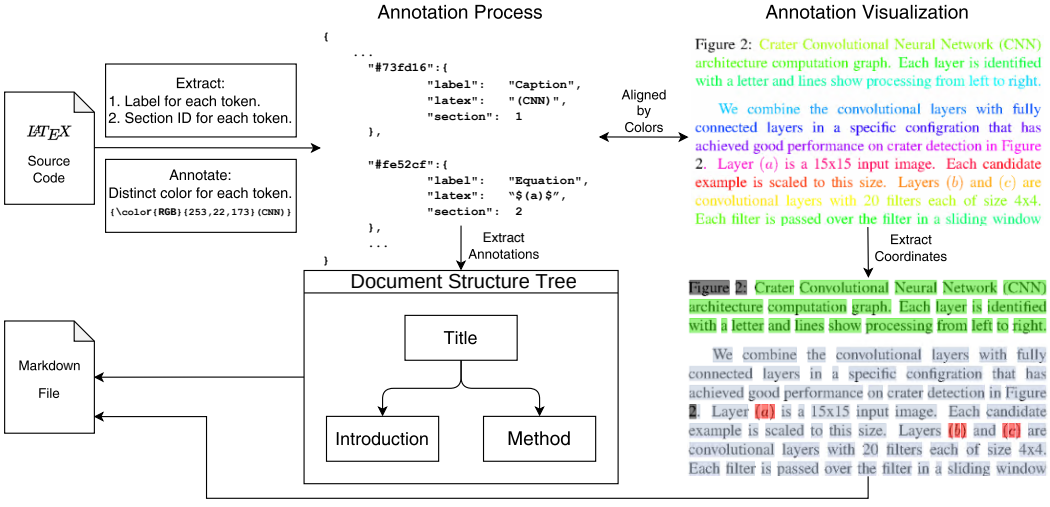
Preprocessing LaTeX Code

1. Fixing unclosed brackets
2. Fixing unmatched environments
3. de-macro[3]: expand macros defined in (re)newcommand or (re)newenvironment commands
4. Enhancing the command database of pylatexenc[4] with the IDE's[5] database.
5. Parsing the abstract syntax tree
6. Assigning each page element a unique color.

[3] https://ctan.org/pkg/de-macro
[4] https://github.com/phfaist/pylatexenc
[5] https://github.com/James-Yu/LaTeX-Workshop

Compiling LaTeX and Extracting Annotations

I use AutoTeX to compile the colored LaTeX source code and then assign each PDF page element its properties, including LaTeX code, reading order, and tree structure.



Standardization of Annotations

Math extracted directly from LaTeX code may contain user-defined commands, which can differ significantly from standard mathematical expressions.
To standardize and normalize the extracted math expressions, we utilized LaTeXML[6]:

1. Collecting all math LaTeX
2. Combining the preamble of the LaTeX source with the math expressions and assigning a unique section title to each math block.
3. Using LaTeXML to convert the assembled LaTeX code into HTML format
4. Extracting normalized math expressions from converted HTML

[6] https://math.nist.gov/~BMiller/LaTeXML/

Future Work

Coloring schemes don't always work:

- **Non-Colorable Elements**: Certain elements such as images, caption labels, and citation tags cannot be colored.
- **Pre-Colored Text**: Some text is already colored, and the additional coloring does not take effect.

Figure 3: A crater and non-crater candidate are processed by the first convolutional layer. Eight filters with interesting activation patterns are shown to the right of each candidate image in false color. Values are scaled

East (3_24+3_25)	79.77	86.09	89.51	90.29

Table 1: F1-score via 10-fold cross validation.

Future Work

- **Parser Limitations**: Despite significant efforts to improve our Python-based LaTeX parser, approximately 40% of papers still fail to be successfully parsed and annotated.
- **Planned Improvement**: tagpdf[7] package generates annotation from LaTeX level, which I plan to use instead.

[7] https://ctan.org/pkg/tagpdf

Thank you!