Paper Extraction and Anonymization Process

Can AI Solve the Peer Review Crisis? A Large-Scale, Cross-Model Experiment of LLMs' Performance and Biases in Evaluating over 1,000 Economics Papers

Pat Pataranutaporn
patpat@media.media.edu

Nattavudh Powdthavee
nick.powdthavee[at]ntu.edu.s

Chayapatr Archiwaranguprok
pub@from.pub

Pattie Maes
pattie@media.mit.edu

1 Paper Processing Pipeline

1.1 Data Collection and Organization

The process scans the `./Journals` directory containing the downloaded papers organized by journal. Each paper receives a unique identifier based on the journal ranking and collection position.

1.2 PDF Text Extraction

The library PyMuPDF (imported as `pymupdf`) converts PDF documents into plain text. Each page is processed individually and stored as a list of page content strings.

1.3 Metadata Generation with AI

We use OpenAI's `gpt-4o-mini` to generate structured metadata for each paper, extracting:

```
Title: The complete title of the academic paper
Author(s): All author names, separated by semicolons
Affiliation(s): Institutional affiliations of the authors
Publication: Complete publication details including journal name, volume, issue, page numbers, and DOI
Funding: Any mentioned funding sources or grant numbers
```

This metadata provides context for the anonymization process.

2 Anonymization Process

2.1 Duplicate Detection and Removal

We identify repetitive elements that could act as identifiers, such as journal names, author information, or download watermarks. The algorithm detects text segments that appear multiple times across pages (segments with a minimum length of 10 characters that occur 3+ times). These segments are evaluated by an LLM to determine if they contain identifying information.

2.2 Text Anonymization

Each page is processed by an LLM to target and remove:

- Author names, emails, and institutional affiliations
- Institutional information identifying authors
- Journal submission details (names, DOIs, volume numbers)

- Acknowledgment sections with specific people/organizations
- Funding details, grant numbers, ethics approval numbers
- Headers/footers with identifying information
- Conference/submission details identifying authors
- Classification codes and indexing terms

Anonymization runs iteratively until no identifiers are detected (maximum 5 attempts) while preserving citations, including self-citations in the third person.

2.3 Page Number Detection and Removal

The system detects and removes page numbers from the processed text. This process functions through several sequential steps:

1. Examination of each page's beginning segments to identify potential page number patterns via regex matching
2. Identification of pages where the first segment contains only numeric values
3. Cross-page analysis of these numeric values to detect consistent sequencing, even with occasional missing numbers
4. Pattern validation through interval and offset assessment, confirming whether the identified numbers form predictable sequences
5. Selective removal of segments confirmed as true page numbers based on their consistent progression across multiple pages

2.4 Publisher-Specific Artifacts Removal

We then remove publisher-specific formatting elements:

- Elsevier formatting (`"A R T I C L E I N F O"`, `"A B S T R A C T"`)
- Wiley formatting (`"O R I G I N A L A R T I C L E"`, `"K E Y W O R D S"`)
- JEL Classification codes and standardized elements

2.5 Parallel Processing

Python's `ThreadPoolExecutor` enables concurrent execution, accelerating processing with configurable worker limits to manage system resources.

3 Output

The system generates the following files:

1. Anonymized version: `./output/id.txt`
2. Original extracted text: `./output/id-original.txt`
3. Tracking dataframe with metadata and word counts

4 Appendix: Prompts

4.1 Metadata Extraction

```
You are a specialized metadata extraction assistant for academic research papers. Your task is to analyze the provided text and extract the following specific information:

1. Title: Identify the title of the paper.

2. Author(s): Extract the full names of all authors mentioned in the paper.
   - If multiple authors are present, separate them with semicolons.

3. Affiliation(s): Identify the affiliations of all authors.
   - If multiple affiliations are present, separate them with semicolons.
   - If an author has multiple affiliations, repeat the name and affiliations for each affiliation.

4. Publication: Identify the complete publication details including:
   - Journal/conference name
   - Volume and issue numbers
   - Page numbers
   - Publication year
   - DOI (Digital Object Identifier)
   - Any other relevant publication identifiers (e.g., ISSN, ISBN)
   Format these details in a standard academic citation style.

5. Funding: Extract any mentioned funding sources, grant numbers, or acknowledgments
   of financial support.

For each field, provide only the extracted information without additional commentary.
If a particular field cannot be determined from the provided text, respond with "Not specified".
```

4.2 Duplicate Flag Detection

```
Based on this metadata: {text}. Check if this excerpt is

1. Non-anonymized (Is the author's name or affiliations directly related to the author's institution but not mentioned in a third-party manner)
2. Related to journal or publication (title of the paper, journal name, conference name, or publication/download details)

Retain the original order of the list (includes every item) and original text without modification
```

4.3 Anonymization Prompt

```
# Academic Paper Anonymization

Identify ONLY identifying metadata that needs removal while preserving ALL research content.

## WHAT TO REMOVE (ONLY):
- Author names, emails, and institutional affiliations of the paper's authors ONLY (IMPORTANT: Only remove identifying information of the paper's own authors when presented as authors or in acknowledgments. DO NOT remove any names in citations or references, including self-citations where authors refer to their previous work in third person. Example: Remove "Smith: Harvard University (email: smith@harvard.edu)" but keep "as Smith (2018) demonstrated")
- Institutional information that directly identifies authors
- Journal submission details (names, DOIs, volume numbers)
- Acknowledgment sections mentioning specific people/organizations
- Funding details, grant numbers, ethics approval numbers
- Headers/footers with identifying information
- Conference/submission details that would identify authors
- Classification codes (e.g., JEL, MSC, ACM codes) and indexing terms/keywords (first page only)

## OUTPUT FORMAT:
Return a JSON object containing only the text segments to remove:
{"to_remove": ["exact text string 1", "exact text string 2", ...]}

## Important Rules:
1. Be precise - extract only the exact text to remove, not the surrounding content.
2. Never modify research content - only identify text for removal.
3. When in doubt, err on the side of preserving content.
4. Include formatting/whitespace in your extracted segments exactly as they appear.
5. The text given is a page extracted from a PDF, expect some information to be missing or weirdly formatted.
6. Do not anonymize the reference section
7. If the text is already anonymized, i.e., contains no identifying information, return ** BLANK ARRAY** in the "removed" field.

**REMEMBER TO DO NOT REMOVE ANY RESEARCH CONTENT THAT IS NOT IDENTIFYING INFORMATION.**

## PAPER METADATA (FOR CONTEXT): {metadata}

TEXT TO ANALYZE:
---
```