Vo Declaration
Exhibit H

EXHIBIT F

Observations on LibGen-SciMag

The quality of the data looks very good overall: the quality is high and the documents are long so this should be great data to learn from, in particular, for highly specialized knowledge!

There are some potential areas of improvement that I observed (in no particular order):

Commented [1]: cc: [redacted]@meta.com Laurens had some interesting observations about scimag

Commented [2]: Thanks for the comments [redacted]@meta.com!
1 total reaction
Laurens van der Maaten reacted with □ at 2023-10-17 18:59 PM

Commented [3]: Excited to see all the amazing work you and the team have doing on LibGen, [redacted]@meta.com! The data looks really good :)

- We can try and remove more copyright headers and document identifiers. For example: `0925-5273/96/$15.00 Copyright © 1996 Elsevier Science B.V. All rights reserved _SSDI_ 0925-5273(95)00148-4` or `DOI: 10.1007/s11669-008-9251-x`. There also occasionally seems to be some other header / footer stuff, for example:
  - *"Received May 26, 1998*
    Revision received September 1, 1999
    Accepted November 23, 1999"

    "Taylor & Francis makes every effort to ensure the accuracy of all the information (the \"Content\") contained in the publications on our platform. However, Taylor & Francis, our agents, and our licensors make no representations or warranties whatsoever as to the accuracy, completeness, or suitability for any purpose of the Content. Any opinions and views expressed in this publication are the opinions and views of the authors, and are not the views of or endorsed by Taylor & Francis. The accuracy of the Content should not be relied upon and should be independently verified with primary sources of information. Taylor and Francis shall not be liable for any losses, actions, claims, proceedings, demands, costs, expenses, damages, and other liabilities whatsoever or howsoever caused arising directly or indirectly in connection with, in relation to or arising out of the use of the Content."

Commented [4]: yes, agree, it shouldn't be there. Currently we are removing all the copyright paragraphs from beginning and the end of the paper. But we can also try to remove any line containing "ISBN", "Copyright", "©", "All rights reserved", ...

Commented [5]: □

I don't know how easy or hard this is but it seems to me like it should be doable for most papers and give us a nice little boost.

- I think encoding of references may be suboptimal. Right now, the text contains things like `Haptic and visual perception modalities complement each other [2, 3]`. And then references 2 and 3 are at the end. (The exact reference style varies between documents.) In the Galactica project, we did quite some ablations on different ways to encode references and we found that cross-referencing titles works better; see Section 3.1.2 of the [HYPERLINK "https://arxiv.org/pdf/2211.09085.pdf" \h]. Marcin did most of this work and may have more details.

Commented [6]: Discussed with [redacted]@meta.com that in Galactica the main focus for this format of the references ([START_REF]) was to have the ability to predict the paper based on the context. Not sure we need this ability for our general purpose model.

I was thinking of removing the references completely, but then you could end up with incomplete sentences like this:
"As [2] suggests, ..." -> "As suggests". So I decided to leave the references as is for now.

If we have cycles to ablate more that, would be interesting to see the outcomes.

Commented [7]: Got it, okay. If we've thought about this and made an intentional decision to keep it as-is, that sounds good to me. It would be interesting to do an ablation on this at some point but I agree it is probably not a p0.

- To what extent should PII be removed from this dataset? There are author names and email addresses in this data as well as street addresses of research institutes.
  - Related, some papers (*e.g.*, `10.1002-macp.200790021.mmd`) have a huge number of authors. I wonder if we should keep those massive author indexes or

whether it is better to try and remove them?

Commented [8]: we didn't remove the authors in Galactica project, but yes we can experiment on that as well

Commented [9]: Probably also not a p0. Just something I was wondering about.

- The conversion to Markdown only seems partial. Like it worked for headers, lists, and italicized text. But tables were not converted to Markdown; they are still in LaTeX format. Similarly, math hasn't been converted to Markdown either (admittedly, Markdown math may not be rich enough for the equations we are trying to capture).

Commented [10]: I think LaTeX is better for math. But yeah, we can change it if needed

Commented [11]: How did we decide what stuff we convert to Markdown and what we keep in LaTeX?

And perhaps more importantly: Are we doing this consistently across all data sources? That kind-of seems important (but I may be wrong!).

- There is something funny happening in the parsing of the footnote in `10.1016-S1048-6666%2806%2980019-4.mmd`: that footnote appears to be replicated many times in the parsed document.

- Depending on the layout of the paper, figure captions sometimes appear in the middle of the text. For example in `10.1016-S1048-6666%2806%2980019-4.mmd` and `10.1016-s1044-0283%2802%2900050-9.mmd`, we have things like:
  - *Early attention is paid to flexion contractures. Gentle passive extension of the contracted joint may begin as early as 1 week by careful flexion of the*

    Fig 5: Stage II proximal and distal Inclions.

    Fig 8: Postoperative controlled mobilization.

    Fig 6: Stage II distal juncture.

    Fig 7: Stage II proximal juncture using Pulvertaft weave.

    adjacent joints. Gentle active motion is allowed at 3 to 4 weeks. The button is removed at 4 weeks. Additional protection can be provided for several weeks with a wrist cuff and rubber band. Blocking exercises are used at 6 weeks.

  - *The Italian CEF also had a standard*

    Figure 2: WEBS, CEFS, and efficient portfolios (postdevaluation).

    deviation at 28.84%.

- I suspect page breaks are sometimes encoded as \n's even when the text continues on the next page. For example, in `10.1001-archsurg.1995.01430090038015.mmd` we have:
  - *For example, the UK WEBS has a higher return (28.42% vs.*

    13.76%) and a lower standard deviation

- Should we keep Acknowledgements sections or can we remove them?

- There are some documents that appear to have OCR errors. For example, in `10.1001-jama.1920.02620500046022.mmd` we have:
    - **LOUISJANA**
    - **MISSISSIPIP**
    - **WORTH CAROLINA**
    - **PennsylvanIXANIA**

- The text in the JSON contains escape characters for the escape character. So it'll contain things like `K. L. Casner (\\(\\subseteq\\))`. This is probably okay as I imagine the first \ will be removed during JSON decoding.

Commented [12]: it's usually some good text there, but I don't have strong opinion on that. Unless there is PII data there

Commented [13]: It's usually a list of funding agencies, project numbers, and people who proofread the paper, right? Doesn't seem super-valuable but I may be wrong..?