Summarizing Qualitative Academic Research: Final Report
Mark Cohen

Introduction

Tools for research and writing represent an important domain of development for Large Language Models (LLMs) and AI in general. Tools like Elicit seek to provide support for researchers sifting through and deriving insights from the literature in their fields. These efforts have tended to focus on finds in science, technology, engineering, and medicine (STEM), seeking to pull out features like research design and key findings for comparison.

The characteristics of texts in the more qualitative segments of the social sciences and the humanities represent a distinct set of problems. One is the formal diversity of research writing in these fields. In contrast, the prevailing style of the STEM fields imposes an endogenous structure on texts — e.g. in the typical sequence of literature view, followed by hypotheses, data and research design, results, and discussion. The stylistic patterns of qualitative research tend to be more fluid. Moreover, articles in the humanities and qualitative social sciences can often be much longer than STEM texts: the submission guidelines of the journal *Nature*, for example, suggests that the typical article should be well under 5,000 words, whereas the *American Journal of Sociology* notes only that "referees may request more time to read papers over 18,000 words."

This project represents an experiment in using an LLM to condense qualitative research writing. This is facilitated by the fact that it is standard practice for academic writers to provide short (100-300 word) summaries of their own work in the form of an "abstract" that is typically appended to the start of the article.

This is, then, a guided summarization task, which is a standard problem in natural language processing (NLP). The specific subject-matter introduces two special challenges. The first is style: capturing the specific rhetorical and linguistic character of academic writing. The second is length: reducing very long original texts to paragraph-length summaries.

The generation of abstracts is largely a task of convenience, since the targets for generation are readily available alongside each published piece of research. Academics are unlikely to be interested in automating writing abstracts for their articles; it is not a major time bottleneck. Nonetheless, this "task of convenience" can serve as an experimental first step towards the potentially more useful application of automatically generating drafts of literature reviews, which consist of summary surveys of existing research on the topic of a research article. I will return in the conclusion to what would be involved in pursuing this further development.

One set of users for whom automated abstract generation could be directly useful, though, is students, especially those trying to write in a language other than their first. A major barrier of entry for non-native speakers to the humanities and social sciences in the English-speaking

world, for example, is the mastery of these fields' distinctive style in talking about one's research. It is not something that is formally taught but instead is expected to be picked up by emulation of examples. For such students, AI summarization cannot legitimately replace the need for them to learn this skill—if for no other reason than that most members of the field would object to this—but it could serve as a learning aid.

Data

The data for this project consists of paired abstracts and full text articles. The source for this data was the CORE database (Knoth et al 2023). CORE collects and indexes scholarly works submitted by institutional partners, in many cases university repositories. Public or institutional policy often requires scholars to provide open-access versions of published research work, and CORE thus offers a centralized API for searching full text and metadata for these works.

The texts archived by CORE are quite diverse in form, including not just research articles but also working papers, dissertations, and even internal institutional documents like press releases. For this project, the sample was restricted to published research articles for two reasons. First, published articles are more likely to have the desired format: a several-thousand word text summarized by a few-hundred word abstract. Longer works like dissertations are outside the desired scope of this project. Second, the publication process involves editing and polishing both texts and abstracts: the coherence of the text and the accuracy of the abstract are likely to be higher. While the CORE metadata do not characterize works in this way, the combination of word lengths and the presence of a "publication date" serves as an effective filter.

One further problem was restricting the sample to the subject areas of interest. The CORE metadata does not consistently categorize works by discipline. To overcome this, subject areas for each work were drawn from a second data source, linked by the unique DOI[1] given to the vast majority of research output in recent decades. DOIs are issued by registering authorities, the largest for academic journals being Crossref, which makes its full database available for download (Crossref 2023). The Crossref metadata includes much more extensive records on the subjects of research outputs. The subject selected were:

- General Social Sciences
- Literature and Literary Theory
- History
- Sociology and Political Science
- Cultural Studies
- Philosophy
- Arts and Humanities
- Gender Studies

[1] Domain Object Identifier. See doi.org.

- Urban Studies
- Political Science and International Relations

One final issue was that quite commonly, the abstract is included at the beginning of the full text, which is the typical format for how articles appear in print journals. Since the object here is to train a model to summarize text, not learn to extract the summary already included at the beginning, it was necessary to search for the text of the abstract and remove it if present.

Combining the subject information from CORE and Crossref, and filtering out records without a full abstract (using 90 words as a cutoff), the resulting dataset was 9,570 pairs of full text and abstract. This was divided into a roughly training set of around 7,500 records and a test set of just under 2,000 records. The full texts were truncated and padded to 16,384 tokens each after encoding.

It is worth noting that the original pool of data from CORE, having filtered only for publication and the presence of a DOI, consisted of three quarters of a million works. While a substantial number of exclusions were for technical reasons (e.g. lacking a full abstract), this also reflects the quantitative balance between STEM fields on the one side and the more qualitative social sciences and humanities on the others.

Modeling Approach

A key limitation of the attention mechanism of transformer models is that the resource requirements scale quadratically with the number of tokens. Devising alternative structures that provide comparable functionality but with better scalability is an ongoing problem.

Guo et al (2022) developed a variation of the T5 model architecture implementing two alternative attention mechanisms. One, “local” attention represents context as a moving average of neighboring tokens (127 on either side, in their released pretrained model). The other, “transient global” attention represents the overall context of the entire text through a set of “global tokens” calculated as the normalized sums of fixed-size blocks (16 in the pretrained model) in the entire input. The local attention mechanism, for a fixed local window size, scales linearly in input length, while transient global attention scales on the order of the square of the length over the block size.

This project employed the “base” (~250,00 parameters) checkpoints published by Guo et al. The models with local-only and local and transient global attention were separately fine tuned on a Google Compute Engine virtual machine using a single NVIDIA 24GB L4 GPU. The initial checkpoints were downloaded and training handled through the Hugging Face Transformers API[2] and PyTorch backend. Each model was trained on around 50,000 iterations.

[2] See https://huggingface.co/docs/transformers/index

Results

This section compares the performance of the two fine tuned models with both the baseline checkpoints and two other publicly available models fine tuned on a research article dataset derived from the arXiv archive, a repository for open-access research. Compared to CORE, arXiv, first, depends on individual uploads rather than systematic institutional contributions and, second, is focused on "the fields of physics, mathematics, computer science, quantitative biology, quantitative finance, statistics, electrical engineering and systems science, and economics." So, models trained on arXiv examples are fine tuned to STEM rather than the qualitative research that is the focus of this project.

Each model was evaluated by generating predicted summaries for the test data set and calculating average ROUGE scores.

Model	Rouge1	Rouge2	RougeL	RougeL Sum
BigBirdPegasus[3]	22.1	2.6	15.7	15.7
LSG[4]	33.4	14.1	21.5	23.4
Baseline: Local	1.1	0.0	0.8	0.8
Baseline: Transient Global	28.9	6.8	14.8	14.8
Tuned: Local	40.1	19.0	28.0	28.0
Tuned: Transient Global	49.1	26.0	33.4	33.4

We can see some of the characteristic tendencies of the models by comparing what each generated for one example in the test set. The actual abstract, from an article in political philosophy (Maffettone 2015), is as follows:

> In this article I address two objections to Rawls' account of international toleration. The first claims that the idea of a decent people does not cohere with Rawls' understanding of reasonable pluralism and sanctions the oppressive use of state power. The second argues that liberal peoples would agree to a more expansive set of principles in the first original position of Law of Peoples. Contra the first I argue that it does not properly distinguish between the use of state power aimed at curtailing difference and the oppressive use of state power. Contra the second I argue that transposing a liberal egalitarian set of

[3] A block attention model fine-tuned on the arXiv dataset for 4,096 tokens (Zaheer et al 2021). Note, then, that generating summaries on 16,384 token texts is well outside its designed applicability. See https://huggingface.co/google/bigbird-pegasus-large-arxiv

[4] LSG stands for "local, sparse, and global" attention mechanisms. This mechanism was proposed by Condevaux and Harispe (2022) who published a model fine tuned on arXiv papers with a maximum length of 16,384 tokens. See https://huggingface.co/ccdv/lsg-bart-base-16384-arxiv

> principles in Law of Peoples would entail the unnecessary duplication of entitlements within different levels of governance and affect liberal peoples' self-determination. The article also highlights how these criticisms are premised on the assumption that all societies should be liberal and that the correct view of global justice is a cosmopolitan one.

First we can compare the BigBirdPegasus and LSG models.

BigBirdPegasus	LSG
in this note we note that the claim of a conflict of the spirit of the statement of the title in the introduction to the paper is in error and that the result of the test is not final	in the first part of the paper, I have presented a critique of the Rawls' account of international toleration against two important objections . the first objection claims that the very idea of a decent people is incoherent with Rawls' understanding of reasonable pluralism and sanctions the oppressive use of state power by the government of a decent hierarchical people . the second objection addresses the core of this worry . according to many liberal cosmopolitans , liberal peoples should agree to a different and more expansive set of principles in the first international original position of LP: representatives of liberal peoples agreeing on a conception of

The former clearly struggles with the length of the text, while the latter, though clearly somewhat better, is not capturing the structure of the article's abstract and inverts the author's intent. The article is not *making* two objections, it is *responding* to two objections for Rawls' theory.

Next we can look at the baseline and fine tuned versions of the local attention model:

Baseline	Fine tuned
rarely stationary Should interiorul AreaDeutsch fingerprintengagement Zeug pharmacy hasslewak Basket supporterspreferably compliance FOR pantry3, regiunizugnnerhalb accidentsresembleRetrouvez becomes …	This paper argues that the alleged inconsistency between rawls' domestic and international theories of justice is premised on eliding the distinction between the oppressive use of state power and the use of state power aimed at curtailing difference. Drawing on the

[continues for about 300 words]	alleged inconsistency between Rawls' domestic and international theories of justice, we argue that the alleged inconsistency between rawls' domestic and international theories of justice is premised on eliding the distinction between the oppressive use of state power and the use of state power aimed at curtailing difference.

The baseline model here is basically non-functional prior to fine-tuning. Note that the "This paper…" construction begins the vast majority of predictions generated by this model. This is capturing a genuine tic of academic writing, though this particular example uses a variant. Aside from this we, see the strength of the model in its reproduction of one of the core points identified in the actual abstract: that certain objections to Rawls fail to properly recognize "the distinction between the oppressive use of state power and the use of state power aimed at curtailing difference." However, rather than summarizing the response to the other objection, the model just repeats the same point again. This kind of repetitiveness is a recurrent weakness of the local attention model.

Finally we can look at the transient global attention model.

Baseline	Fine-tuned
While positive appraisals of Rawls' account of international toleration do appear in the literature (see Porter 2012; Dogan 2010; Jenkins 2010; Reidy 2010; Avila 2007; Freeman 2007a; Mertens 2005; Hayfa 2004), this paper concentrates on two objections that have catered comparatively less attention and that, if successful, would prove fatal to the Rawlsian project. According to many liberal cosmopolitans, liberal peoples should agree to a different and more 2 expansive set of principles in the first international original position of LP: representatives of liberal peoples agreeing on a conception of international justice would include a more substantive set of human rights and some form of egalitarian distributive principle (see Pogge 2006; 2004; 2001; 1994; Beitz 2000). According to many liberal cosmopolitans, liberal peoples should agree to a different and more 2 expansive set of principles in the first international original position of LP: representatives of liberal peoples agreeing on a conception of international justice would include a more substantive set of human rights and some form of egalitarian distributive principle (see Pogge 2006; 2004; 2001; 1994; Beitz	This paper argues that the Rawlsian account of international toleration is inconsistent from a liberal cosmopolitan standpoint. The first part of the paper addresses the apparent tension between decent hierarchical peoples and Rawls' understanding of reasonable pluralism. The second part of the paper addresses the core of this worry. According to many liberal cosmopolitans, liberal peoples should agree to a different and more expansive set of principles in the first international original position of LP: representatives of liberal peoples agreeing on a conception of international justice would include a more substantive set of human rights and some form of egalitarian distributive principle (see Pogge 2006; 2004; 2001; 1994; Beitz 2000). This would automatically exclude decent peoples from

2000). One of the crucial aspects of Rawls' political turn, and the main message that the idea of the burdens of judgment conveys, is that when citizens' basic deliberative and expressive liberties are protected, even conscientious uses of human reason, unimpeded by simple bias and self-interest, will not deliver agreement on a single vision of the many non-political values that constitute the core of what Rawls calls a comprehensive doctrine (see Quong 2011; Dreben 2003). Yet, looking at (A) more closely, the emphasis on 'domestic' and 'international' should signal that there is a clear difference between the first and second parts of the proposition: when we transpose aspects of liberal democratic institutions into an international law of peoples, those aspects are transformed from a matter of domestic policy into a matter of international concern (see Beitz 2009a), and in so doing, the amount of control that liberal peoples have over domestic political affairs is curtailed (see Buchanan 2013).9 Would that be problematic from the standpoint of the first international original position in LP? And what happens when the two come into conflict	the scope of toleration as they would clearly be unable to respect, for the right reasons, a law of peoples which would require all societies to be internally liberal.

The baseline here is much more functional than the baseline local attention model. Nonetheless, the generated text does not function well as an abstract. The fine-tuned summary here is far smoother and more sophisticated in its language than any of the other models. However, it also includes a crucial error in the first sentence: the point of the paper is to argue that the cosmopolitan's criticisms of Rawls are misguided. Indeed, the rest of the predicted abstract captures key elements of this argument. To speculate on the underlying cause here, it is a feature of the discursive style of the humanities to spend a lot of space presenting and analyzing the arguments that are the target of criticism. This is what the model is capturing: the paper first lays out the opposing position at length (summarized in the first sentence of the generated text) and *then* explains its shortcomings (summarized in the rest of the predicted abstract).

Conclusion

This project demonstrates the performance that can be achieved by fine-tuning a not-so-large language model, on several thousand examples, for a few dozen hours, on a single GPU. The NLP task tackled by this project was summarization of very long texts. The best model, using a transient global attention mechanism, generated abstracts for research articles that matched the sophistication and subtlety of language of the discursive domain of research articles in the qualitative social sciences and humanities.

Even if the results were not always perfect — as in the example above — this represents a first step towards potentially automating time-consuming steps in the research process such as

literature reviews. The researching and writing literature reviews are tasks often assigned by academics to student assistants. Students volunteer to do this as a way to get "research experience," but in fact they are assigned such busywork rather than being mentored in question formulation and research methods. Automating this could thus be valuable for both researchers and students. The challenge of automating the writing of even a rough draft of a literature review is that the works surveyed need to be not just summarized individually but summarized *in relation to* one another. That is, of course, a further challenge not tackled in this project. Nonetheless, summarizing individual pieces — in particular showing that this can be done with a relatively small model with only moderate training time and resources — is a starting point.

References

Condevaux, C. and Harispe, S. (2022). LSG Attention: Extrapolation of pretrained Transformers to long sequences. https://doi.org/10.48550/arXiv.2210.15497

Crossref. (2023). April 2023 Public Data File from Crossref. http://dx.doi.org/10.13003/8wx5k

Guo, M., Ainslie, J., Uthus, D., Ontañón, S., Ni, J., Sung, Y.-H., & Yang, Y. (2022). LongT5: Efficient Text-To-Text Transformer for Long Sequences. *Findings of the Association for Computational Linguistics: NAACL 2022*, 724–736. https://aclanthology.org/2022.findings-naacl.55

Knoth, P., Herrmannova, D., Cancellieri, M. et al. (2023). CORE: A Global Aggregation Service for Open Access Papers. *Nature Scientific Data* 10, 366.. https://doi.org/10.1038/s41597-023-02208-w

Lin, C.Y. (2004). ROUGE: a Package for Automatic Evaluation of Summaries. *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, 74-81.

Maffettone, P. (2015). Toleration, decency and self-determination in The Law of Peoples. *Philosophy and Social Criticism*, 41(6), 537-556. https://doi.org/10.1177/0191453714567736

Zaheer, M. et al. (2021). Big Bird: Transformers for Longer Sequences. https://doi.org/10.48550/arXiv.2007.14062