LURID4: Longitudinal study of the World Wide Web: Detecting LLM output

1. Details

Summary

Longitudinal study of the Web provides evidence for the change from human-authored pages (the early Web) to pages more and more frequently assembled from material extracted from databases (Web 2.0) to the present and the advent of pages generated by Large Language Model (LLM) systems such as ChatGPT. The proposed work will extend our work on Common Crawl (CC) data in the LURID and LURID2 projects (EPSRC Tier 2, 2021-9 thru 2022-9 and 2022-12 thru 2023-04) and LURID3 (UKRI APP39557 2024-10 thru 2025-06).

In particular, we will be able to use the multiple CC datasets for which we have augmented indices containing LastModified dates to look back in time to the early Web and assemble a baseline collection of high-probability human-authored material.

The utility of the Web for the training of LLMs is at risk because of the growing presence of LLM output. We will use the human-authored baseline to evaluate existing LLM-output detection tools, and, in combination with other cues derived from our earlier studies, to explore approaches to 'cleaning' recent Common Crawl datasets to reduce the extent of LLM-generated 'pollution' they contain. The results will provide the basis for a subsequent responsive-mode application.

Start date

January 2026

Duration

12 months

---

2. Core team

Name	Role	Organisation
Henry S Thompson	Project lead	University of Edinburgh

3. Keywords

Application question

Provide up to 3 keywords that describes your application.

This is for administrative purposes to help the initial application processing, and for identifying introducers for your application.

LargeLanguageModels WebData Robustness

4. HPC service requested

Application question

Please name which HPC service are you applying to.

You may also name a reserve choice of service, providing you are willing to take up the offer of this service, should your first choice be oversubscribed and you are awarded time on your reserve choice.

The services available for this opportunity are:

- ARCHER2 (EPSRC remit only)
- Isambard 3 (UKRI remit)
- Other Tier-2 HPC services (UKRI-remit): Baskerville, Bede, Cirrus, CSD3, Jade, NI-HPC, Sulis.

Cirrus then Cirrus2

---

5. HPC resources requested

Application question

Please enter the quantity and units, for example compute units (CU), CPUh, GPUh, or Node-Hours (NH), of HPC resources that you are requesting.

N/A

1.5 million CPU hours

Dataset storage: 1,000TB disk storage, already in place

Working storage: Cirrus standard provision

---

6. High End Computing (HEC) consortia

Application question

If you are a member of a currently funded HEC consortium, and are applying for ARCHER2 compute, please enter the name of the consortium and briefly explain why you are not applying for time through this consortium.

N/A

N/A

---

7. Vision and Approach

Application question

What are you hoping to achieve with and how will you deliver your proposed work?

For the Vision, explain how your proposed work:

- is of excellent quality and importance within or beyond the field(s) or area(s)
- has the potential to advance current understanding, generates new knowledge, thinking or discovery within or beyond the field or area
- is timely given current trends, context and needs
- impacts world-leading research, society, the economy or the environment

Within the Vision section, we also expect you to identify the potential direct or indirect benefits and who the beneficiaries might be. In addition, include how it would contribute to computational science, for example through generating new codes, developing existing codes, increasing computational efficiency, or opening up HPC for new scientific areas and industrial sectors.

For the Approach, explain how you have designed your work so that it:

- is effective and appropriate to achieve your objectives
- is feasible, and comprehensively identifies any risks to delivery and how they will be managed
- if applicable, uses a clear and transparent methodology
- if applicable, summarises the previous work and describes how this will be built upon and progressed
- will maximise translation of outputs into outcomes and impacts
- describes how your, and if applicable your team's, research environment (in terms of the place, and relevance to the project) will contribute to the success of the work

Within the Approach section we also expect you to:

- include details of the codes and methods to be used and clarification of how they are appropriate for the work being proposed

References may be included within this section.

See attached.

0.07MB

8. Applicant and team capability to deliver

Application question

Why are you the right individual or team to successfully deliver the proposed work?

Evidence of how you, and if relevant your team, have:

- the relevant experience (appropriate to career stage) to deliver the proposed work
- the right balance of skills and expertise to cover the proposed work
- the appropriate leadership and management skills to deliver the work and your approach to develop others
- contributed to developing a positive research environment and wider community

Within this section you can also demonstrate elements of your responses in visual form if relevant.

When including images you must:

- provide a descriptive caption or legend for each image immediately underneath it in the text box (this must be outside the image and counts towards your word limit)
- insert each new image on a new line.
- use files smaller than 5MB and in JPEG, JPG, JPE, JFI, JIF, JFIF, PNG, GIF, BMP or WEBP format.

Images should only be used to convey important visual information that cannot easily be put into words. The following are not permitted, and your application may be rejected if you include:

- sentences or paragraphs of text
- tables
- excessive quantities of images

A few words are permitted where the image would lack clarity without the contextual words, such as a diagram, where text labels are required for an axis or graph column.

The word count for this section is 1,650 words: 1,150 words to be used for R4RI modules (including references) and, if necessary, a further 500 words for Additions.

Use the Résumé for Research and Innovation (R4RI) format to showcase the range of relevant skills you and, if relevant, your team (project and project co-leads, researchers, technicians, specialists, partners and so on) have and how this will help deliver the proposed work. You can include individuals' specific achievements but only choose past contributions that best evidence their ability to deliver this work.

Complete this section using the R4RI module headings listed. Use each heading once and include a response for the whole team, see the UKRI guidance on R4RI (https://www.ukri.org/apply-for-funding/develop-your-application/resume-for-research-and-innovation-r4ri-guidance/). You should consider how to balance your answer, and emphasise where appropriate the key skills each team member brings:

- contributions to the generation of new ideas, tools, methodologies, or knowledge
- the development of others and maintenance of effective working relationships
- contributions to the wider research and innovation community
- contributions to broader research or innovation users and audiences and towards wider societal benefit

Additions

Provide any further details relevant to your application. This section is optional and can be up to 500 words. You should not use it to describe additional skills, experiences, or outputs, but you can use it to describe any factors that provide context for the rest of your R4RI (for example, details of career breaks if you wish to disclose them).

Complete this as a narrative. Do not format it like a CV.

References may be included within this section.

The roles in funding applications policy (https://www.ukri.org/publications/roles-in-funding-applications/) has descriptions of the different project roles.

Module 1 – Contributions to the generation of new ideas, tools, methodologies or knowledge

I am Professor of Web Informatics, based in the Institute for Language, Cognition and Computation in the School of Informatics at the University of Edinburgh. I have been carrying out research and supervising student projects in the area of Web Architecture since 2004.

Selected relevant publications:

1. Thompson, H. S. (2010b) "What's a URI and why should I care?". Ariadne **65**. Online journal, available at http://www.ariadne.ac.uk/issue65/thompson/ (http://www.ariadne.ac.uk/issue65/thompson/).

2. Thompson, H. S., J. Rees and J. Tennison (2013) "URIs in data: for entities, or for descriptions of entities—A critical analysis". In *Proceedings of WebSci '13*, ACM, 479–482. Available online at https://doi.org/10.1145/2464464.2532514 (https://doi.org/10.1145/2464464.2532514)

3. Thompson, H. S. and C. Lilley, eds. (2014) *XML Media Types*. Internet Engineering Task Force (IETF). IETF Standards Track RFC 7303, available online at https://tools.ietf.org/html/rfc7303 (https://tools.ietf.org/html/rfc7303).

4. Thompson, H. S. and J. Tong (2018) "Can Common Crawl reliably track persistent identifier (PID) use over time?". In *Proceedings of The Web Conference 2018*, ACM, 1749–1755. Available online at https://doi.org/10.1145/3184558.3191636 (https://doi.org/10.1145/3184558.3191636)

5. Thompson, H. S. and Norman Walsh, eds. (2008) *Associating Resources with Namespaces*. W3C TAG Finding, World Wide Web Consortium, Cambridge. Available online at http://www.w3.org/2001/

tag/doc/nsDocuments/ (http://www.w3.org/2001/tag/doc/nsDocuments/)

6. Thompson, H. S. et al. (2022-01-10). Thread on common-crawl Google Group: https://groups.google.com/g/common-crawl/c/AmsXrCNVBzo/m/us4j_7tpAAAJ (https://groups.google.com/g/common-crawl/c/AmsXrCNVBzo/m/us4j_7tpAAAJ)

7. Thompson, H. S. (2024) "Improved methodology for longitudinal Web analytics using Common Crawl". *Proceedings of ACM Web Science Conference (WEBSCI '24), May 21–24, 2024, Stuttgart, Germany*. Available online at https://dl.acm.org/doi/10.1145/3614419.3644018 (https://dl.acm.org/doi/10.1145/3614419.3644018).

8. Thompson, H. S. (2025) "Common Crawl and Generative AI: An introduction, and thoughts about detection", presented at *Beyond Deepfakes: advancing watermarking in Generative AI* workshop. Available online at https://www.cogsci.ed.ac.uk/~ht/CC_risk_LLMs.pdf (https://www.cogsci.ed.ac.uk/~ht/CC_risk_LLMs.pdf)

Module 2 – The development of others and maintenance of effective working relationships

My empirical investigation of Web Architecture has provided the context for the dissertations of six generations of MSc students whom I have trained in the use of HPC techniques. Several of them have gone on to further research in this area. The proposed work will incorporate dissertation projects for five students in 2026, training them in using HPC tooling for very-large-scale data analysis.

Module 3 – Contributions to the wider research and innovation community

My previous HPC projects **LURID2** and **LURID3** developed and exploited an approach to expanding the possible scope of research into longitudinal trends on the Web. By adding Last-Modified timestamps to the CC datasets' indices we enabled efficient fine-grained longitudinal exploration of Web history.

This allowed us to produce year-, month- and day-level analysis of how URI lengths have changed over time. This inspired the approach we will explore in **LURID4** whereby we can instrument how the process of creating of Web pages itself has changed and is changing.

One augmented index is available online at https://catalogue.eidf.ac.uk/dataset/eidf125-common-crawl-url-index-for-august-2019-with-last-modified-timestamps (https://catalogue.eidf.ac.uk/dataset/eidf125-common-crawl-url-index-for-august-2019-with-last-modified-timestamps) and two others will be published there later this year.

Module 4 – Contributions to broader research/innovation-users and audiences and towards wider societal benefit

I was a member of the SGML Working Group of the World Wide Web Consortium (W3C) which designed XML, a major contributor to the core concepts of XSLT and W3C XML Schema and a member of the XML Core and XML Processing Model Working Groups at the W3C. I have been the University of Edinburgh's representative on the Advisory Council (AC) of the W3C since we joined in 1998. I was elected by the AC to four successive terms (9 years) on the W3C Technical Architecture Group (TAG). The TAG is chaired by Sir Tim Berners-Lee, and is responsible for the stewardship of the architecture of the Web.

I regularly review draft standards (RFCs) for the Internet Engineering Task Force (IETF) in the areas of Web protocols and data formats.

I have been a contributor to the Edinburgh Science Festival, Cafe Scientifique, a British Science Association Talk/Film series, the Albertus Institute and other public understanding of science fora.

I have spoken in several contexts about the need to address the existential risk that will arise from training LLMs on data which includes their own outputs.

I am committed to the proposition that our professional responsibilities and the success of our intellectual pursuits both depend on keeping our humanity clearly in view.

9. Computing resources and justification

Application question

What computing resources will you need to deliver your proposed work?

Provide evidence of:

- full justification of the resources requested and why you are requesting the specific system
- justification of why the computing time requested is essential for the successful completion of the project
- assurance that the resources requested are appropriate and justified

Assessors are not looking for a line-by-line breakdown of all project resources. Overall, they want you to demonstrate how the resources you anticipate needing for your proposed work:

- are comprehensive, appropriate, and justified
- represent the optimal use of resources to achieve the intended outcomes
- maximise potential outcomes and impacts

Based on experience in our previous three HPC projects, we estimate CPU requirements as follows

Purpose	Hours per task	Number of tasks	Total
Training MSc students in the use of the scheduling and data-parallelism tools	500	5	2,500
MSc pilot project	25,000	5	125,000
Producing an augmented index for one CC dataset	25,000	7	175,000
Post-process refetched PDFs	50,000	1	50,000
MSc dissertation project	150,000	5	750,000

Exploratory data analysis of authorship evidence	250,000	1	250,000
Contingencies	150,000	1	147,500
Total			**1,500,000**

10. Project partners

No project partners.

11. Project partners: letters (or emails) of support

Application question

Upload a single PDF containing the letters or emails of support from each partner you named in the Project Partner section. These should be uploaded in English or Welsh only.

Enter the words ‘attachment supplied’ in the text box, or if you do not have any project partners enter ‘N/A’. Each letter or email you provide should:

- confirm the partner’s commitment to the project
- clearly explain the value, relevance, and possible benefits of the work to them
- describe any additional value that they bring to the project
- the page limit is two sides of A4 per partner

Ensure you have prior agreement from project partners so that, if you are offered funding, they will support your project as indicated in the project partners’ section.

For audit purposes, UKRI requires formal collaboration agreements to be put in place if an award is made.

Do not provide letters of support from host and project co-leads' research organisations.

N/A

12. Ethics and responsible research and innovation (RRI)

Application question

What are the ethical or RRI implications and issues relating to the proposed work? If you do not think that the proposed work raises any ethical or RRI issues, explain why.

Demonstrate that you have identified and evaluated:

- the relevant ethical or responsible research and innovation considerations
- how you will manage these considerations

Within this section you can also demonstrate elements of your responses in visual form, such as images, if relevant.

When including images you must:

- provide a descriptive caption or legend for each image immediately underneath it in the text box (this must be outside the image and counts towards your word limit)
- insert each new image on a new line.
- use files smaller than 5MB and in JPEG, JPG, JPE, JFI, JIF, JFIF, PNG, GIF, BMP or WEBP format.

Images should only be used to convey important visual information that cannot easily be put into words. The following are not permitted, and your application may be rejected if you include:

- sentences or paragraphs of text

- tables
- excessive quantities of images

A few words are permitted where the image would lack clarity without the contextual words, such as a diagram, where text labels are required for an axis or graph column.

All our research to date, as well as the proposed work, conforms to the Common Crawl (CC) Terms of Use. We have sought and received from CC confirmation that our publication on the Web as part of LURID2 of a 220GB augmented index to the August 2019 CC dataset is allowed, and this permission would also apply to the publication of similar augmented indices for the additional CC datasets produced since then and in the proposed work.

13. Genetic and biological risk

Application question

Does your proposed research involve any genetic or biological risk?

In respect of animals, plants or microbes, are you proposing to:

- use genetic modification as an experimental tool, like studying gene function in a genetically modified organism
- release genetically modified organisms
- ultimately develop commercial and industrial genetically modified outcomes

If yes, provide the name of any required approving body and state if approval is already in place. If it is not, provide an indicative timeframe for obtaining the required approval.

Identify the organism or organisms as a plant, animal or microbe and specify the species and which of the three categories the research relates to.

Identify the genetic and biological risks resulting from the proposed research, their implications, and any mitigation you plan on taking.

Assessors will want to know you have considered the risks and their implications to justify that any identified risks do not outweigh any benefits of the proposed research.

Enter N/A in the text box if not applicable to your proposed work.

N/A

14. Research involving the use of animals

Application question

Does your proposed research involve the use of vertebrate animals or other organisms covered by the Animals Scientific Procedures Act?

If you are proposing research that requires using animals, download and complete the Research involving the use of animals template (DOCX, 52.5KB) (https://ukri-tfs-prod-assets.s3.eu-west-2.amazonaws.com/Research+involving+the+use+of+animals+template.docx) which contains all the questions relating to research using vertebrate animals or other Animals (Scientific Procedures) Act 1986 regulated organisms.

Enter N/A in the text box if not applicable to your proposed work.

N/A

15. Conducting research with animals overseas

Application question

Will any of the proposed animal research be conducted overseas?

If you are proposing to conduct overseas research, it must be conducted in accordance with welfare standards consistent with those in the UK, as in Responsibility in the use of animals in bioscience research, page 14 (https://www.nc3rs.org.uk/3rs-resources/responsibility-use-animals-bioscience-research). Ensure all named applicants in the UK and overseas are aware of this requirement.

If your application proposes animal research to be conducted overseas, you must provide a statement in the text box. Depending on the species involved, you may also need to upload a completed template for each species listed.

Statement

Provide a statement to confirm that:

- all named applicants are aware of the requirements and have agreed to abide by them
- this overseas research will be conducted in accordance with welfare standards consistent with the principles of UK legislation
- the expectation set out in Responsibility in the use of animals in bioscience research will be applied and maintained
- appropriate national and institutional approvals are in place.

Templates

Overseas studies proposing to use non-human primates, cats, dogs, equines or pigs will be assessed during NC3Rs review (https://www.nc3rs.org.uk/who-we-are-and-what-we-do#services) of research applications. Provide the required information by completing the template from the question ‘Research involving the use of animals’.

For studies involving other species, such as:

- rodents
- rabbits
- sheep
- goats
- pigs
- cattle

- xenopus laevis and xenopus tropicalis
- zebrafish

Select, download, and complete the relevant Word checklist or checklists by exploring NC3Rs checklist for the use of animals overseas (https://gbr01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fnc3rs.org.uk%2Fchecklists-use-animals-overseas&data=05%7C02%7CKirsty.Cook%40ukri.org%7C7776d059f1d0490639c708dd07cb999e%7C8bb7e08edaa44a8e927efca38db04b7e%7C0%7C0%7C638675293218338887%7CUnknown%7CTWFpbGZsb3d8eyJFbXB0eU1hcGkiOnRydWUsIlYiOiIwLjAuMDAwMCIsIlAiOiJXaW4zMiIsIkFOIjoiTWFpbCIsIldUIjoyfQ%3D%3D%7C0%7C%7C%7C&sdata=hLATRpVO7ftZ4ltg6A1pu6T4eaUn85sC2QHy5xKEFTU%3D&reserved=0).

Enter N/A in the text box if not applicable to your proposed work.

N/A

16. Research involving human participation

Application question

Will the project involve the use of human subjects or their personal information?

If you are proposing research that requires the involvement of human subjects, provide the name of any required approving body and whether approval is already in place.

Justify the number and the diversity of the participants involved, as well as any procedures.

Provide details of any areas of substantial or moderate severity of impact.

Enter N/A in the text box if not applicable to your proposed work.

N/A

17. Research involving human tissues or biological samples

Application question

Does your proposed research involve the use of human tissues, or biological samples?

If you are proposing work that involves human tissues or biological samples, provide the name of any required approving body and whether approval is already in place.

Justify the use of human tissue or biological samples specifying the nature and quantity of the material to be used and its source.

Enter N/A in the text box if not applicable to your proposed work.

N/A