Project Title: BERT for Humanists

Institution: Cornell University

Project Directors: Matthew Wilkens, David Mimno, Melanie Walsh (University of Washington)

Grant Program: Digital Humanities Advancement Grants, Level III

Personnel

Project team

Matthew Wilkens, project director, Associate Professor, Department of Information Science, Cornell University

David Mimno, co-project director, Associate Professor, Department of Information Science, Cornell University

Melanie Walsh, co-project director, Assistant Teaching Professor, Information School, University of Washington

Rosamond Thalken, lead developer, PhD Candidate, Department of Information Science, Cornell University

Advisory board

David Bamman, Associate Professor, School of Information, UC Berkeley

Peter Bol, Professor, East Asian Languages and Civilizations, Harvard University

Rachel Buurma, Associate Professor, English, Swarthmore College

Matthew Gold, Associate Professor, English and Digital Humanities, The Graduate Center, CUNY

Alvin C. Grissom II, Assistant Professor, Computer Science, Haverford College

Lauren Klein, Associate Professor, English and Quantitative Theory and Methods, Emory University

Lucy Li, PhD student, School of Information, UC Berkeley

Allison Parrish, Assistant Arts Professor, New York University

Ted Underwood, Professor and Associate Dean, School of Information Sciences and English, University of Illinois, Urbana-Champaign

Hongsu Henry Wang, Research Fellow, Institute for Quantitative Social Science, Harvard University

Richard Wicentowski, Professor, Computer Science, Swarthmore College

Narrative

Enhancing the humanities

Beginning with BERT in 2018, large language models (LLMs) have revolutionized natural language processing (NLP). Such models are now the foundation for a wide range of familiar systems, from grammar checking and autocomplete to translation and speech-to-text. By combining huge numbers of parameters with vast text collections, pre-trained LLMs offer advanced general-purpose language understanding off-the-shelf. Rather than starting from scratch for each new language application, we can now start from a strong "foundation" model. Newer models such as LaMDA and GPT-3 have shown stunningly strong abilities.[1]

We seek a level III DHAG to support the well-established BERT for Humanists project. BERT for Humanists provides three key resources to inform, empower, and inspire humanities scholars to use LLMs in their disciplines in creative new ways:

1. **Education**. Current LLM documentation is geared towards machine learning (ML) researchers. It includes too much information about the inner workings of models and not enough information about how to apply them. We build and maintain tutorials and worked examples that show how to use large language models such as BERT and its newer descendants on real, concrete humanities problems. The materials include a limited, maintainable set of open-source software libraries that simplify advanced computational tasks for use with the unique collections that are key to humanistic research.
1. **Research.** Most current LLM applications focus on NLP benchmarks or commercial products. We conduct new research to define the humanities-relevant affordances of large language models. This research includes both the evaluation of novel LLM-based techniques on the complex, diverse documents of interest to humanists and the use of high-performing LLM methods to address existing humanities questions.
2. **Community.** We support a community of experienced and LLM-curious humanists who discuss their unique problem domains, successes, pain points, and desiderata with experienced and humanities-curious machine learning researchers for direct incorporation into tools, models, and methods.

Together, the three products of BERT for Humanists provide an intellectual framework for understanding and evaluating new computational language technologies, so that humanists may be positioned to make use of — and to critique, as appropriate — the advances that will inevitably supplant BERT (in particular) and large language models in general.

As powerful as NLP tools can be in the humanities, we have also seen clear examples of the problems that arise from a lack of good humanities-focused resources to interpret their outputs and to provide guidance to researchers. Lack of clarity about

[1] For GPT-3, see Tom B. Brown et al. "Language Models are Few-Shot Learners." arXiv:2005.14165 [cs.CL], https://doi.org/10.48550/arXiv.2005.14165. For LaMDA, see Romal Thoppilan et al. "LaMDA: Language Models for Dialog Applications." arXiv:2201.08239 [cs.CL], https://doi.org/10.48550/arXiv.2201.08239.

protocols and incomplete understanding of models have led to concerns about the validity and reproducibility of findings. The need to self-teach computational tools has also privileged researchers who have more computational experience or institutional support, leading to perceptions that such methods are accessible only to a small group.

These issues are especially acute in the case of large language models. Previously, NLP researchers tended to train a new model from scratch for each specific task (such as translation, named-entity identification, or sentiment analysis). BERT introduced a new paradigm, in which a single general-purpose model is pre-trained on a (very) large collection of text. Researchers then create a copy of the general-purpose model and "fine-tune" it through additional training on a smaller, special-purpose data set to perform individual NLP tasks. This approach has led to noticeable and sometimes dramatic improvements in a range of applications, but brings with it a number of challenges. The general-purpose base language models are sufficiently large and complicated that they can be produced only by a handful of corporate or governmental organizations, and their training data remain difficult to inspect in full. The process of fine-tuning is computationally expensive and requires considerable experience in machine learning and stochastic optimization.

We established BERT for Humanists to help humanists move quickly and confidently into LLM-based research, while remaining aware of its limitations. We have seen strong interest in the project (including hundreds of workshop participants and thousands of tutorial users in less than a year) and have registered a clear need for additional resources to remain on top of this rapidly evolving field.

Several specific developments will be relevant for humanists in the years ahead:

- Text-to-text generation. Earlier models were focused on producing short, simple outputs like category labels or missing-word predictions. Newer models such as GPT-3 and T5 are capable of producing long spans of fluent text.
- Multilingual models. Models such as XLM-RoBERTa and mT5 have been pre-trained on large text collections in many languages, not just English. In some cases, these models appear to be able to transfer behaviors learned in one language to other languages.
- Few-shot and zero-shot learning. One of the most challenging aspects of large language models is fine-tuning general models to perform specific tasks. There is increasing evidence that carefully specifying text input in the form of plain-text instructions can enable the original general-purpose model to carry out complex, specific tasks.

Currently, humanists must find, install, and use a different specialized software for each individual text analysis process. But with these new large language models, it may be possible for researchers with little to no experience with NLP to accomplish complicated tasks simply by specifying examples of desired input/output pairs. Consider an example of zero-shot learning. When we ask the T5-large model to generate continuation text that could follow the input “my house is full of …,” the top output is “books.” But when we ask it to generate text from the input “translate English to German: my house is full of …,” the top answer is “mein Haus ist voll von.” In other words, by changing the input text, we are able to convert a general-purpose language model into a translation model *without any further training*. But this example also demonstrates the risks of being too confident in large models: using the same input but replacing “German” with “Italian” *also* results in the output “mein Haus ist voll von.” While the potential of large language models is huge, the dangers for naïve use are also great.

Text generation models also have the potential to provide paradigmatic use cases for researchers who might not otherwise see the potential of NLP. To date, we have been most successful in reaching users already familiar with language technology. But it remains more difficult to convince less experienced audiences that the effort needed to absorb new tools will be individually beneficial. In the related case of image models, text-to-image generation apps like DALL-E have provided a clearly understandable view of what computer vision does well and what is still challenging. Similar ludic applications in NLP models will provide researchers with an analogous perspective on the capabilities of large language models.

The pace of developments in NLP is unlikely to slow. We have sufficient experience in the process of building connections between new technologies and the humanities that we can reliably identify and predict where many difficulties will arise. These include definitions and descriptions of tasks, data formats, and computational affordances, but also infrastructure and environment setup. Creating a more abstract framework for building resources to help humanists adapt to new technologies, and for technologists to recognize the compelling problems of humanists, has made this process more streamlined and less haphazard. Our team of humanists and computer scientists, housed under the interdisciplinary umbrella of Information Science, is ideally positioned to continue these advanced translational tasks.

Environmental scan

BERT-like methods have only just begun to be incorporated into DH scholarship, owing largely to the difficulties we have noted. Manjavacas and Fonteyn have released BERT models trained on historical English texts for use by the community of technically advanced users in DH and linguistics.[2] Sims et al. and Bamman et al. have used a BERT-based architecture to perform event detection, entity recognition, and coreference resolution in literary texts, and have built a high-level system, BookNLP, that performs those specific tasks for users who can run the software locally.[3] Our group at Cornell has used BERT models to study the evolving rhetoric of the Supreme Court, the social dynamics of online medical support communities, and the historical development of multilingual literary geography.

Outside the humanities, researchers have created language models trained on discipline-specific texts (BioBERT, SCIBERT) that are well suited to field-relevant tasks.[4]

[2] Manjavacas, Enrique and Lauren Fonteyn. "MacBERTh: Development and Evaluation of a Historically Pre-trained Language Model for English (1450-1950)." *Proceedings of the Workshop on Natural Language Processing for Digital Humanities* (NLP4DH), pp. 23–36. December 19, 2021. https://macberth.netlify.app/

[3] Bamman, David, Olivia Lewke, and Anya Mansoor, “An Annotated Dataset of Coreference in English Literature,” arXiv:1912.01140[cs], http://arxiv.org/abs/1912.01140. Sims, Matthew, Jong Ho Park, and David Bamman, “Literary Event Detection,” *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Florence, Italy: ACL, 2019), 3623–3634, doi:10.18653/v1/P19-1353, https://www.aclweb.org/anthology/P19-135311.

[4] For BioBERT, see Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining.” arXiv:1901.08746 [cs.CL], https://doi.org/10.48550/arXiv.1901.08746. For SciBERT, see Iz

The widely used software libraries from Hugging Face have helped to make BERT and its derivatives more readily available to advanced users in computer science and applied machine learning. But the very fact that there exists a popular set of programming libraries to make large language models easier for *computer scientists* to use suggests the need for additional resources geared toward humanists.

Existing efforts to make computational methods more accessible to humanists have had high impact. The Programming Historian and The Data-Sitters Club offer a range of tutorials, but focus on simpler and more entry-level tools and concepts compared to LLMs. Other comparable programs include ODH-funded workshops via the Institutes for Advanced Topics in the Digital Humanities program. These initiatives have been extraordinarily valuable in helping newcomers begin to explore computational methods.

One group of users we serve is more advanced than the ones targeted by such projects; they are familiar with the fundamentals of text analysis and can generally write their own code. But their work does not yet take advantage of the significant advances made possible by large language models. Existing tutorials created by and for machine-learning researchers are mostly inaccessible to these users, both because those tutorials assume background technical knowledge that humanists rarely possess and because their application domains are far removed from the historical, literary, and interpretive areas at the core of humanities research.

Another group of users we now seek to target is the larger pool of humanities scholars who are not programming literate, but who might benefit from the accessibility of text-to-text LLMs. These users are likely to have encountered the results of computational work and to pursue research agendas amenable to quantitative evidence, but lack the expertise or resources to build computational models from scratch. For these users, we aim both to explain the affordances of generative language models and to explore the space of problems that are (and are not) well suited to the low-code and no-code techniques these models enable.

History of the project

BERT for Humanists was established in 2020 with the support of an NEH ODH Level I DHAG. The project currently offers tutorials, recorded workshops, and other resources to a large and growing user base (hundreds of cumulative live participants and thousands of asynchronous users). It also guides new research with and about large language models. BERT for Humanists organizes and benefits from the active participation of a sizable advisory board that helps to shape our work.

To date, we have carried out three workshops at Cornell, the first two by invitation to a panel of experts and the third to the public. Two hundred researchers from around the world registered to participate in our public virtual workshop, and more than 90 attended the event. Beyond this workshop, we are pleased that the resources created by our project have also been shared, used, and discussed widely, such as in blog posts by DARIAH's OpenMethods platform and Princeton's Center for Digital Humanities. Testifying to its success as a teaching tool within the scholarly community, the BERT for Humanists Project was nominated for the "Best Digital Humanities Training Materials"

Beltagy, Kyle Lo, Arman Cohan. "SciBERT: A Pretrained Language Model for Scientific Text." arXiv:1903.10676 [cs.CL], https://doi.org/10.48550/arXiv.1903.10676.

award as part of the 2021 DH Awards. Our BERT for Humanists resources have also proved useful for undergraduate and graduate students, and they have been integrated into courses across the country, such as in Ted Underwood's 2021 "Data Science in the Humanities" course in the Information School at the University of Illinois Urbana-Champaign, and in Lauren's Klein's 2021 "Practical Approaches to Data Science with Text" course at Emory University. We presented many of the lessons learned from designing these materials at the 2021 Association for Computational Humanities (ACH).

After running our initial tutorials, we have received several invitations to run sequels and variations of the original tutorials, such as for the online school NLP+CSS in December 2021, for the International Conference of Web and Social Media (ICWSM) in June 2022, and for Bell Labs at Cambridge University in July 2022. Over the course of these workshops, we developed an extensive set of use cases with worked Python notebook examples. We also solidified connections among the members of our large advisory board and built an interdisciplinary community of researchers beyond the walls of our directly participating institutions.

Our tutorials use the Hugging Face *transformers* library. This library is a foundational resource for LLM-related work, lowering significantly the barriers to entry for researchers who are already familiar with common programming practices in natural language processing. Among our goals in these tutorials has been to help NLP-literate humanists understand how to leverage their existing knowledge for use with the LLMs that Hugging Face exposes.

LLM-based work in languages other than English is relatively easy at the technical level. Many languages now have pretrained models available through the Hugging Face repository. Our work includes an example in Spanish demonstrating a Spanish-language model running on Golden-Age sonnets (early 16th century to late 17th century), and we plan to add additional non-English and multilingual tutorials as an integral part of our future work.

Activities and project team

Our activities during the 36-month performance period will be divided into two parts, **research** and **translational education,** which we will develop in parallel. In earlier work we found that these were tightly linked: building tutorials required us to develop new applications, and helped focus our research on examples that were both immediately compelling and also readily explainable.

Our **research** will involve three case studies that highlight LLM abilities to identify subtle and previously unmeasurable patterns. In the first, we will leverage our existing expertise in literary geography and connections to linguistically diverse collaborators to study evolving engagement with the Arabic and Scandinavian regions in a multilingual corpus of fiction published between 1800 and 2010. This work requires identifying and disambiguating personal and geographic names across languages, a challenging problem requiring the linguistic and cultural knowledge embedded in LLMs.[5]

[5] See Elizabeth F. Evans and Matthew Wilkens, "Nation, Ethnicity, and the Geography of British Fiction, 1880-1940," *Journal of Cultural Analytics*, 2018, https://doi.org/10.22148/16.024 and Matthew Wilkens, "The Geographic Imagination of Civil War-Era American Fiction," American Literary History 25.4 (2013): 803–40.

Our second case study in rhetoric will evaluate whether we can identify and distinguish rhetorical devices in legal documents using a small number of examples in a few- and zero-shot setting, in comparison to using a larger labeled set in a fine-tuned setting. Our third case study will use LLMs to predict the information- and support-seeking goals of narratives in a large corpus of texts from online medical support communities, with the aim of better understanding the unmet needs of marginalized patients.

Together, these three projects address important questions in literature, history, medical humanities, gender studies, and postcolonial studies. We intend our results to communicate directly with scholars in these fields, whether or not they consider themselves digital humanists. If we are to make the case for the value of advanced LLM-based techniques, we must ultimately demonstrate that those techniques allow us to build knowledge that would be otherwise out of reach. These projects are attempts to do just that, each building on prior success by our group.

In addition to case studies, we will evaluate how well the emergent paradigm of few-shot and zero-shot learning with generative models such as GPT-3 and T5.[6] These methods have been shown to achieve good performance on difficult linguistic tasks such as answering reading-comprehension questions. Humanities applications could include named-entity extraction and disambiguation, and the inference of social relationships between characters. The ability to identify a complex concept simply by presenting a few examples of a phenomenon could revolutionize humanists ability to study texts at scale, but it remains unclear in the NLP literature exactly which problems can expect good performance with few- and zero-shot methods, and there is little existing work that addresses use cases relevant to humanists. We intend to produce both experiential knowledge and clearly formulated guidelines to help humanists decide between possible LLM approaches. To say, in effect: “For problems of type *x*, you can usually get away without fine tuning a model, but for problems of type *y*, you’ll likely need to fine tune or to train from scratch. And problems of type *z* remain largely beyond the current state of the art.”

In order to best serve the priorities of existing humanities research, we will convene regular online meetings of our large and diverse advisory board (comprising humanists, computer scientists, linguists, and social scientists) to discuss their needs as well as the affordances of new and emerging technologies. We anticipate holding one such meeting every six months, as we have done with significant success to date. These meetings surface new developments in large language models that may be relevant to humanities research, provide continuing insight for computer scientists into the complicated questions that interest humanists, and supply the literal and metaphorical spaces in which transdisciplinary understanding develops. The knowledge emerging from these meetings will continue to guide our development of translational learning resources (workshops and tutorials) and software packages tailored to the needs of DH scholars.

Our **translational education** will involve two classes of deliverables. We will develop detailed tutorials and hold workshop sessions open to the public (for existing tutorials and sample workshop agendas, see the appendix), as we have done in the past, in order to make the methods we explore as widely available as possible. Priority

[6] See also Francesco De Toni et al. “Entities, Dates, and Languages: Zero-Shot on Historical Texts with T0.” arXiv:2204.05211 [cs.CL], https://doi.org/10.48550/arXiv.2204.05211.

topics for development include separate few-shot and zero-shot learning examples, the first built around named entity recognition in historical documents, the second around literary character detection. Additional topics include a fully trained model to characterize relationships of relative status in narrative fiction and an evaluation of surprisal (or predictability) in longer texts generated by LLMs in response to user prompts. We will develop additional tutorials in response to user demand and to new developments in the rapidly evolving field.

Our workshops represent a bridge outward from the transdisciplinary community of our board to the wider body of computation-aware humanists. We will devote the entirety of one of these sessions – to be held in person at Cornell – to critiques and ethical considerations related to large language models, while incorporating explicit consideration of those issues into all of our work and authoring a white paper that synthesizes recent debates in the ethics of AI and large language models.[7] In these aspects of the project, we benefit in particular from the participation of faculty members with expertise in DH pedagogy in both large university and small college settings.

We will also develop and maintain a limited set of software libraries tailored to the specific pain points of humanities text analysis with LLMs. These include utilities to preprocess and divide long-format documents such as novels and first-person historical accounts into chunks suitable for analysis by pretrained LLMs, and software to connect code notebooks to cloud-based computational resources. Our team has experience producing and maintaining such friction-reducing packages while avoiding duplication of existing resources.[8] We will also work to make our tools and documentation complementary to official Hugging Face documentation, both to make those tools more accessible to humanists and to broaden the impact of our work.

All of our tutorials, notebooks, software, and workshop materials will continue to be made available for free via our existing website, http://www.bertforhumanists.org. Project co-directors **Matthew Wilkens** and **David Mimno** (both Information Science, Cornell University) will be responsible for supervision and coordination of all aspects of the project. They will be jointly responsible for mentoring students involved in project activities and for ensuring a respectful, mutually beneficial environment for all participants. Wilkens will be responsible for reporting and compliance certification. **Melanie Walsh** (Information School, University of Washington) will contribute expertise in literary text analysis and humanities-oriented computer science pedagogy, and will provide additional coordination of meetings, workshops, and online presence. **Rosamund Thalken** (Information Science, Cornell) will serve as lead software developer for the project.

[7] Examples of recent ethics work includes Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜.” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23. FAccT ’21. New York, NY, USA: ACM. https://doi.org/10.1145/3442188.3445922; Martin, Meredith. 2022. “AI off the Rails.” The Center for Digital Humanities at Princeton. April 15, 2022. https://cdh.princeton.edu/updates/2022/04/15/ai-off-the-rails-a-response-to-ai-and-the-everything-in-the-whole-wide-world-benchmark/; Irene Solaiman et al. “Release Strategies and the Social Impacts of Language Models.” arXiv:1908.09203 [cs.CL], https://doi.org/10.48550/arXiv.1908.09203; Laura Weidinger et al. “Ethical and social risks of harm from Language Models.” arXiv:2112.04359 [cs.CL], https://doi.org/10.48550/arXiv.2112.04359.

[8] See, for example, our Little MALLET Wrapper, as well as the glossaries and tutorials available on the project site.

Final products and dissemination

In keeping with the three core deliverables of the BERT for Humanists project, we anticipate three classes of final products:

1. Our **educational** outputs will take the form of tutorials, code notebooks, live and recorded workshop sessions, white papers, and similar training materials. These will be delivered both asynchronously and live online, on campus at Cornell, the University of Washington, and others, and at conference venues including DH, ACH, MLA, and CSCW. All materials are now and will remain open access and freely available to all users.
2. Our **research** outputs will be published in relevant journals and conferences. These include established humanities venues such as *PMLA* and *JAH*, newer DH publications including *Cultural Analytics* and the *Journal of Digital Arabic and Islamic Research*, and computer science, information science, and social science conferences where experts in computational methods can encounter distinctively important humanities results. We have enjoyed a strong record of publication success that we expect to continue through the performance period.
3. Our **community** outputs rely in part on our close working relationships with the advisory board and on the wide intellectual circles of the project team. We rely on the members of our advisory board to bring our shared work to colleagues at their institutions and in their subfields. In this way, we can extend our reach far beyond the professional networks of the core project members. We see the close relationships we have built with our advisory board members as a project deliverable in their own right. We have also reached thousands of online users from around the United States (and the world). The project's ability to put these users into active communication with one another is among our most significant results. Finally, we will deliver a concluding white paper that will serve as a playbook for evaluating and incorporating future developments in NLP and AI into humanities research. No one can anticipate exactly how language technologies will evolve in the years ahead, but we believe that the experiences and relationships between computing and the humanities that we build today will be critical to maintaining productive ties between the two fields in the future.

Work plan

Schedule

Our two major project tracks – research and education – are mutually informing and will proceed in parallel during the project's 36-month performance period. A tabular schedule of work follows the narrative below.

Research

We will produce three pieces of significant new humanities research using large language models. These address, respectively, 1.) the evolving rhetorical strategies of the Supreme Court of the United States, 2.) the explicit and implicit medical support needs of patient participants in online support communities, especially those dedicated to marginalized people, and 3.) the multilingual literary geography of the Nordic countries and of the Arab world.

We have significant existing expertise in these three areas and have already curated suitable corpora for our work. We will sequence our efforts in the order listed (Supreme Court, medical communities, literary geography) over the first 24 months of the performance period, which will allow us to explore progressively greater reliance on the capabilities of increasingly powerful pretrained language models as we move from fine tuning to few-shot to zero-shot methods. Rosamond Thalken will lead topics one and two. Matthew Wilkens will lead topic three. David Mimno and Melanie Walsh will contribute advice and expertise as needed.

As an allied aspect of our application-specific research, we will also study the affordances of low-code interactions with base language models such as T5 and LaMDA. Here, the goal is to understand the differences in model performance between the types of contemporary web- and media-based applications on which the models are conventionally evaluated in machine learning and NLP research and the historical and literary use cases that are central to the humanities. In addition to incorporating these methods into our topical research as appropriate, we will evaluate competing approaches to few- and zero-shot learning as applied to humanities work. These studies will include historical named entity recognition, character relationship extraction, cohesive narrative text generation. David Mimno will lead this aspect of the research with the input of the other team members and the advisory board. These efforts will begin immediately and will continue through the first 24 months of the performance period.

Education and dissemination

We want to help humanists explore the power – and the limits – of contemporary large language models as quickly as possible. To this end, we will focus our early attention on very recent developments on few- and zero-shot learning. We will develop two new tutorials on these topics during the first 12 months of the performance period. We will present these tutorials in conjunction with workshops for our advisory board in months 6 and 12 in order to gather their feedback and advice.

Between months 12 and 18, we will convene an in-person meeting of our advisory board and selected outside participants to help develop a white paper on the ethical implications of large language models. This is an area of intense interest within the computer science and NLP community that has to date proceeded without adequate humanities involvement. In consequence, the debate as it stands has been impoverished by a lack of philosophical and social sophistication, and has not widely penetrated the circle of humanists who rely on language models for their own work.

In the second half of year two, we will develop a tutorial and workshop that compares fully fine-tuned performance to low-code approaches on the complex task of character relationship detection and scoring. In this case, we seek to identify relationships of respect, obligation, servitude, friendship, and the like over the full character roster of literary texts, in order to build more critically robust character interaction networks. We anticipate that this task will be challenging but feasible and of direct relevance to scholars of literature.

We will devote the final 12 months of the project to two advisory board meetings and the development of two tutorials. We leave the subject matter of these final products somewhat open in order to accommodate the very rapidly evolving state of research in large language models. Our plans call tentatively for a consideration of the limits of narrative coherence in longer texts (beyond the scale of the paragraph or the page) generated by LLMs and for a summary playbook of advice to humanists on approaches that are best suited to different humanities research use cases.

Rosamond Thalken will serve as the lead developer of all tutorials. Melanie Walsh will lead development and presentation of the associated workshop sessions. David Mimno and Matthew Wilkens, in consultation with the advisory board, will provide additional input and guidance.

Our dissemination strategy, in addition to publicly available asynchronous tutorials and hybrid online/in-person workshops to be offered at steady intervals throughout the performance period, includes conference and journal publications of our research results. The formal publications will be concentrated in the second half of the performance period.

Timeline

Year	Term	Research	Education	Meetings
2023	Spring	Supreme Court (Thalken); Low-code approaches (Mimno)	Few-shot tutorial and workshop (Walsh, Thalken)	Advisory board (online, May)
	Fall	Medical communities (Thalken, Wilkens); Zero-shot learning and prompt engineering (Mimno)	Zero-shot tutorial (Walsh, Thalken)	Advisory board (online, Oct)
2024	Spring	Literary geography (Wilkens)	AI ethics white paper (Walsh)	Advisory board meeting with invited speakers (in-person, May)

	Fall	Literary geography (Wilkens); Character status modeling (All)	Comparative approaches to character status modeling (Mimno, Thalken)	Advisory board meeting (online, Nov)
2025	Spring	Text generation (Thalken, Mimno)	Evaluating narrative coherence in generated text (Walsh, Thalken)	Advisory board meeting (online, May)
	Fall	New problems in LLMs (All)	Playbook for AI and LLM use in the humanities (All)	Concluding advisory board meeting (online, Dec)

Risks

We identify three primary risks to the project:

Retention of effort from project participants. We have built a dedicated project team and a large, diverse, highly skilled advisory board to direct the project's development. Our work depends centrally on the efforts of the primary project members, for each of whom we have budgeted appropriate compensation. We do not rely on any single (uncompensated) board member's involvement, but we benefit from their expertise and would suffer if they were to leave the project. We have hedged against this risk by increasing the size of the board, by offering compelling content that members could not replicate elsewhere, and by giving members meaningful input into our shared work. We have not, to date, had problems with board attrition.

The research elements of the project may not achieve the results we expect. Like any research effort, our success is uncertain. Our primary hedges are the expertise of the project members, the existence of preliminary results supporting each of our hypotheses, and the fact that we have suitable corpora in hand. Even negative results in these cases would likely represent useful contributions to the field.

Changes in the culture or economics of language model development. Contemporary large language models require significant resources to train. For this reason, pre-training is usually performed by corporate or governmental entities (Google, the EU, etc.) The trained models are then usually released for free to the public. We use these publicly-available models in our work. If future models were to be withheld, our work would be impacted, though we could go far with the models currently in hand. We judge the likelihood of a significant change in model availability to be small, but we hold personal and professional relationships with model-building entities that would allow us to access new models, even if they were withheld from public release.

Data management plan

BERT for Humanists uses large amounts of (primarily textual) data, but it does not generate or store directly most of the data it uses. For this reason, the project relies on a range of specialized third-party sites for most of our data storage and access needs. The types of data used and generated by the project are summarized in the table below.

Data we consume

Data that we consume takes the form of digitized books, user-generated web-based content, government documents, pretrained language models, and publicly-available reference corpora for model training and evaluation. All of this data is provided by outside entities, from HathiTrust to Google to Reddit and many others. We store some of this data temporarily for local processing on Cornell's computing resources, but we do not archive or redistribute it, nor do we seek to do so.

Cornell's research computing cluster provides fully managed and secured file storage, CPU and GPU compute pools, and project management expertise. The costs of cluster use and support for the full performance period are included in our budget.

In every case, we provide unique identifiers, code, and detailed methods descriptions that allow other researchers to replicate our work, provided they have access to the same third-party sources we do. We seek to avoid the use of commercial or restricted-access data sources wherever possible. We do not anticipate any significant use of commercially or legally restricted data sets other than those provided by HathiTrust (which are available to all members of the academic HathiTrust consortium).

Data we produce

We generate several types of project data. Our educational materials include code notebooks and libraries that we author, video recordings of workshop sessions and live tutorials, blog posts, glossaries, and white papers. All of these materials are linked from our web site and are stored on freely-available platforms appropriate to their format (including GitHub, Google Colab, YouTube, and Wordpress). The core project members are responsible for producing and managing these materials.

Our generated research data takes the form of derived features and metadata about the documents with which we work. It also includes in-house training data for our models, as well as trained or fine-tuned models for specific use cases. Most of this data is made available on GitHub. Full models are published via Hugging Face and/or stored on Google Cloud resources if they exceed GitHub's file size limits. Each core project member is responsible for organizing and managing research data related to their work. Project directors Wilkens, Mimno, and Walsh, supported by Cornell's research computing staff, are responsible for training other members in appropriate data management practices.

We do not anticipate collecting, processing, or distributing individually identifiable information about human subjects or medical records. Any protected or individually identifiable data we might collect will be encrypted, anonymized in accordance with current best practices, and will not be redistributed to outside researchers.

Archiving

At the conclusion of the performance period, we will archive all of the data we generate via the Cornell institutional repository, eCommons. eCommons provides a no-cost, professionally managed, persistently available endpoint for long-term access to all of our funded products.

Dissemination

We disseminate summaries of our regular advisory board meetings as blog posts on our project website. We announce publicly all of our tutorials and workshops on our site, on social media, and via relevant email listservs, and we make portions of these tutorials and workshops available as recordings on YouTube. As noted above, we make all of our educational materials available for asynchronous use via several free, public platforms. We release relevant Python packages through *pip*, the standard package manager for Python. The Python packages will also be released under a GNU General Public License, which allows for the greatest amount of circulation and use of the materials.

Summary of managed data

Note that all data will be archived via Cornell's institutional repository at the conclusion of the project.

Data type	**Storage**	**Notes and limits**
Blog posts	Project website	
Video recordings	YouTube	For live recordings, presenter view only to preserve participant privacy
Python code	GitHub, pip, Hugging Face	
Tutorials and examples (Google Colab notebooks)	GitHub, project website	Live executable notebooks hosted on Google Colab
Research data: Literary texts, court decisions, and derived data	HathiTrust, GitHub, project website	Literary texts and extracted features supplied by HathiTrust. Derived features (e.g., named entities, metadata) and identifiers released on GitHub.
Research data: Medical communities derived data	GitHub	Code and scraping methods only, to protect participant privacy
Research data: Models and training data	Hugging Face, GitHub	
White papers	NEH, project website	

Sustainability plan

The resources needed to sustain BERT for Humanists after the performance period are modest. We seek to make a specific, punctual intervention at a moment of rapid transformation in natural language processing and in the humanities. The project is not, and does not seek to become, a commercial or institute-style home to manage longer-term developments at the intersection of these fields. We believe longer-term interventions are the proper domain of durable institutions such as academic departments, funding organizations like NEH, and commercial entities.

Our deliverables are educational materials linked to the current state of the art, research results that aim to intervene in contemporary debates and to provide models for future work, and a robust set of personal and institutional connections among project participants. The products will be sustained primarily by the larger community of scholars, rather than by the ongoing labor of the project team.

We have budgeted the project to cover the significant human resource costs of the research and educational activities we propose over the 36-month performance period. We have designed our outputs to rely on proven, low-cost technologies, most of which have been available to the public for more than a decade and have demonstrated their independent sustainability (GitHub, YouTube, Wordpress, etc.). Where we have relied on newer entrants (Hugging Face), there are sufficient commercial resources backing them that we are confident of at least their medium-term sustainability.

As a fallback for the popular publication platforms we have selected, we will also deposit our products in Cornell's institutional repository, which provides no-cost, long-term archival storage.

Other potential near-term sustainability needs are easily met. The project is self-documenting, in the sense that one of its major deliverables is a set of resources that will help researchers in the field replicate and extend our work. If the project is successful, there will be many people who will continue aspects of its work. (If it is unsuccessful – which we believe is unlikely, of course – then the question of sustainability is moot.) Each of the core project members is also embedded in a large, vibrant, and very well-resourced department that is committed to exploring the broad implications of new language technologies. While we are uniquely qualified to carry out the specific objectives of the BERT for Humanists project, we can count on collaborators and internal funding sources to fill any unanticipated needs that may arise.