The Last Word: Speaking with our Sources— The Possibilities and Pitfalls of AI Language Models in Historical Research

Jacob Forward and Xioachen Zhu

"The task of American citizens is to preserve an air of invulnerability that can lead to peace and freedom, even where the threat of war looms." This original and eloquent insight into Cold War culture was not written by a historian, nor even by a human. It was written by a form of artificial intelligence (AI) called a language model, which had been trained on a large number of State of the Union Addresses. Language models (LMs) offer historians a potentially powerful, though not unproblematic tool for augmenting traditional discourse analysis and working with much larger corpora of sources than before. This is particularly promising for modern and contemporary historians, who face, as Paul S. Boyer observes, "a dizzying abundance of potential sources."

LMs are digital neural networks that excel at adopting the style and content of the writing they are trained on and can generate original text output. One model, called ChatGPT, has made headlines around the world since its release by OpenAI in November 2022. Last autumn there was real excitement among digital historians and digital humanists about the potential of this technology for humanities research, and our project, which uses an AI language model to study the discourse of State of the Union Addresses, is one of the very first examples of how they might be used to augment the research capabilities of historians.

We chose State of the Union Addresses because these annual speeches from the incumbent president to the legislature, the nation, and the world, are arguably the heart of American political rhetoric and we reasoned that a language model specialized on them could yield excellent insight into the themes of recent American history. We chose a publicly available LM called GPT-2, which we trained on a corpus consisting of every State of the Union Address from the end of WWII to the present day.

In essence, LMs work by analysing large amounts of text and learning the patterns and structures of the language. Once trained, the model can then generate new text that is similar in style and structure to the input it was trained on. It does this by predicting which words will come next in light of the preceding context. We provided the context in the form of a short prompt on which the LM could then build a few sentences. Due to the extensive training of LMs they learn deep patterns in language that can make their outputs appear uncannily human.

Having trained the LM on our corpus of speeches we gave it several prompts to write about key topics in American history. The LM wrote in the first person, just like the speeches, and mimicked the dramatic, stylized, and idealistic language of presidential rhetoric. For example, in response to the prompt in italics it wrote, "*The state of the union is* strong. We have a new, unified vision and a new, hopeful spirit, one that will stand strong against the specter of cynicism, the shadow of fear." The LM also used a justificatory tone— "that is why," "for this reason,"— reflecting the president's task of persuading the legislature to pass certain laws in the coming year. Sometimes the LM expressed a latent idea from the corpus so bluntly that it could be quite humorous. Consider this unusually frank assessment of US power: "America has never been more determined than tonight to shape change that doesn't always pan out." This suggests that LMs can be useful for studying the tone, and some of the hidden assumptions and associations in historical sources.

This capacity to render explicit what was implicit by rephrasing and summarizing broad themes in the corpus proved particularly insightful in the case of America's role in the world. We were surprised by the extent to which cynical, imperialist sentiments came to the fore, for instance: "*America in the world means* nothing. We need to take every chance in the world to meet our obligations to the people of the United States." In particular, it emphasized the imperial export of values: "*The American people* are demanding more. We are fighting the same battles we have already won, not just at home, but for the world." This neatly reflects the sentiment of much US foreign policy in the past half-century, which has often rationalized military intervention in terms of the defense of supposedly universal values, most notably in the War on Terror.

As these examples show, the special utility of LMs, in comparison to other digital tools for macro-textual analysis, is that they reveal associated ideas, not merely associated words. This is partly thanks to their ability to write, but it's also a result of their capacity to discern deeper patterns in the text, patterns of thought. This allows historians to explore how concepts, not just terms, are related to each other in a corpus.

Digital history methodologies have come a long way since the statistical research of the 1960s and '70s, and the current suite of tools available for macro-textual analysis, as seen for instance in Voyant Tools, can offer us useful insight into historical sources. However, where current tools can show us that the words 'American' and 'citizens' are used together seven times in our corpus of speeches, LMs can

An AI art model's interpretation of the prompt "State of the Union Address," made on the NightCafe platform using an open-source stable diffusion algorithm

produce a range of written responses that meaningfully re-contextualize these terms, for instance, *"the task of American citizens is* to preserve and strengthen the ideals of freedom and justice," or *"the task of American citizens is* to lead a world in which freedom, justice and peace can meet the aspirations of men everywhere." This strongly confirms the work of State of the Union Address scholar Deborah Kalb, in her finding that citizenship is rhetorically constructed in allegiance to national values.

However, we should be clear that this technology is far from flawless, and its application to historical research poses many ethical and methodological questions. The power of LMs is measured in the number of parameters (micro-rules about language) that it has learned during training. We had access to the standard version of GPT-2 which is a 117 million parameter model, much less capable than GPT-3 which is a 175 billion parameter model, and this affected the fluency of the generated text. As a consequence, numerous outputs were misspelled, nonsensical or strangely truncated. Additionally, the LM could sometimes be overwhelmed by the frequent recurrence of certain ideas in the corpus, leading to strangely repetitive outputs such as "the world is changing. The world is changing more and more rapidly; the world is changing as well."

From a historian's perspective, there are major issues with the reproducibility and transparency of research with LMs. They are effectively a 'black box' of weights and variables that generates a slightly different response each time to the same prompt. The archivist Rick Prelinger raised the concern that with AI we might "synthesize a past that never existed." But if historians treat the output of LMs not as definitive answers to research questions, nor as synthetic sources, but as invitations for the further exploration of themes with the traditional close reading of the text, then we can still benefit from the insights this technology has to offer us.

While Michael Moss, writing in the early 1990s, worried that digital history was inevitably quantitative and that it would become increasingly "difficult to see the thesis for the data," LMs, and AI more generally, are good at deriving patterns from the data, using more data than scholars could possibly process manually. Even if, as computational linguist Patrick Juola puts it, most of their outputs are "flat gibberish, the one in a thousand may include interesting and provocative readings that human authors have missed." As more powerful language models become publicly available, we can expect more like one in ten outputs to be of real research value.

There is still so much more to explore, for instance by training multiple LMs on different time periods we could trace changes in discourse over time. LMs could be adjusted into history specialist chatbots, which promises a boom in public engagement with archives and history. We might even resurrect historical personages from their written remains, transforming them into digital entities capable of conversation. Truly, we are at the start of the AI turn in history. Moreover, language models fit the brand of history, as neither art nor science, for they utilize a precise and intricate series of calculations to yield a product that is ephemeral, unreproducible, contingent, and even beautiful. The very existence of such a powerful means for mapping language, let alone what it can generate from our sources, is surely cause for us to take a new perspective on history.