Re-Experiencing History: A Platform for the Re-Enactment of Historical Events with Multimodal Large Language Models

Ströbel, Phillip Benjamin
phillip.stroebel@uzh.ch
University of Zurich, Schweiz
ORCID: 0000-0003-2063-5495

Maier, Felix Klaus
felix.maier@hist.uzh.ch
University of Zurich, Schweiz
ORCID: 0000-0002-5578-723X

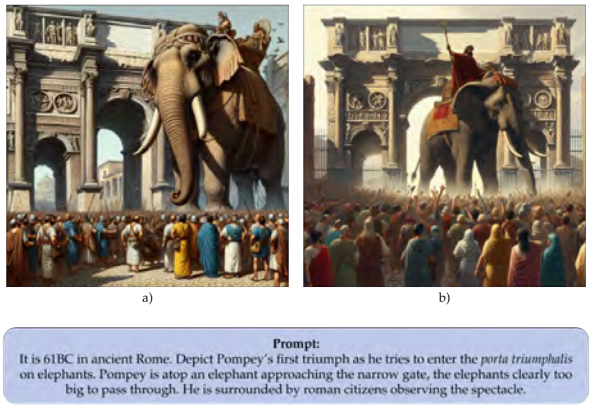
Project Description

In 1982, seven people in the Chicago area tragically died after consuming cyanide-laced capsules (cf. Bergmann (2000)). The source of this tampering was initially unknown, leading to nationwide panic. Investigators were stumped until they conducted **re-enactments** of the purchase and consumption of the contaminated capsules. These re-enactments enabled them to trace the tainted products to specific store shelves and identify the exact locations where the tampering occurred.

Like criminologists, historians have to reconstruct past moments or situations. However, unlike criminologists, historians often cannot re-enact[1] events due to the fact that their subject of investigation cannot be easily reproduced. The advent of AI in the form of powerful multimodal large language models (LLMs) is a game changer here: Historians can now prompt image generation models (mostly based on *Stable Diffusion* (Rombach et al., 2022)) to recreate scenes from primary sources and research findings. This enables them to quickly visualise past events, thereby enhancing their understanding of historical moments. The project *Re-Experiencing History* aims to create a platform exploiting LLMs to support users in re-enacting such historical events.

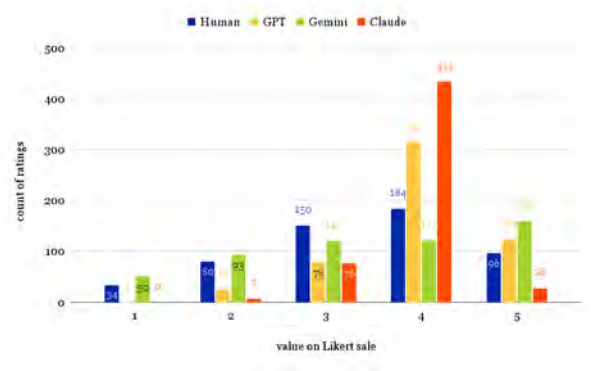
Our research includes assessing existing multimodal LLMs regarding historical accuracy and prompt-to-image alignment (Xu et al., 2023), manipulating generated images with further prompts, and fine-tuning LLMs to improve historical accuracy. The **interdisciplinary approach** where computational linguists work with historians is crucial, as Hutson, Huffman, and Ratican (2024) highlighted in their work on resurrecting Mary Sibley (1800-1878) using her diaries.

In a prototype setting, we focus on two scenarios from antiquity: the Roman triumph and the *Lupercalia* festival. We assess image generation capabilities, initially focusing on DALL-E 3. Based on literature (15 sources on the triumph, 5 on the *Lupercalia*), we crafted 100 prompts to generate six images per prompt, creating 600 images total.[2] These images and prompts form a basis for further research. Figure 1 displays two images generated with the same prompt.



Example of two images generated with the same prompt depicting a triumphal procession in ancient Rome as described in Algül (2018).

We know these images are challenging to evaluate and that automatic evaluations correlate poorly with human judgment (Otani et al., 2023). A first human assessment of the 600 images by 20 history students, coupled with an automatic evaluation using LLMs like GPT-4o, Gemini, or Claude Sonnet 3.5 as evaluators, shows that prompt-to-image alignment on Likert scale from "1 – The image does not match the prompt at all." to "5 – The image completely matches the prompt." is sometimes highly overestimated and too generous (especially by Claude) (see Figure 2).[3] Moreover, we found that the scores on the Likert scale as assessed by LLMs are usually higher for the Roman triumph than for the *Lupercalia*, suggesting potential knowledge gaps. We aim to close this gap by fine-tuning language and vision models on relevant material, increasing historical accuracy.



Aggregation and comparison of scores of human ratings vs. LLM ratings on a Likert scale.

Our application targets history students at all levels, the general public, and museums. We envision a platform that creates accurate images and allows their editing and posting in an online gallery. The application aims to foster **engagement** with historical content and serve as a research tool for historians to reconstruct past events, enhancing our understanding of history (Rosenzweig and Thelen, 1998).

This project contributes to the field of digital humanities and explores the potential of AI in historical research and education. By bridging the gap between past events and modern technology, we aim to create a more immersive and accessible approach to studying history, potentially redefining how we interact with and understand our collective past.

Fußnoten

1. Although re-enactment is present as a concept in, e.g., R. G. Collingwood's *Idea of History* (cf. Dray (1995)), it is rather perceived as thought experiments.
2. We will make the prompts and the generated images available soon.
3. We must highlight at this point that the results are preliminary and that out of the 600 images, only 544 have been rated. The evaluation is still on-going, which is why we cannot provide an inter-annotator agreement yet. However, Figure 2 shows initial tendencies of the prompt-to-image alignment when evaluated with LLMs.

Bibliographie

Algül, Aydın. 2018. *The Roman Triumph: Participation, Historiography and Remembrance*. https://www.academia.edu/43295099/The_Roman_Triumph_Participation_Historiography_and_Remembrance(accessed 21 July 2024).

Bergmann, J oy . 2000. A Bitter Pill. In *Chicago Reader*, https://chicagoreader.com/news-politics/a-bitter-pill(accessed 21 July 2024).

Dray, W illiam H erbert . 1995. *History as Re-Enactment: R. G. Collingwood's Idea of History*. Oxford: Oxford University Press.

Hutson, James , Paul Huffman and Jeremiah Ratican. 2024. Digital Resurrection of Historical Figures: A Case Study on Mary Sibley through Customized ChatGPT. In *Faculty Scholarship,* 590, https://digitalcommons.lindenwood.edu/faculty-research-papers/590(accessed 21 July 2024).

Otani, M ayu , Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yute Nakashima, Esa Rahtu, Janne Heikkilä and Shin'ichi Satoh. 2023. Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 14277–14286, https://cvpr2023.thecvf.com/virtual/2023/poster/22014(accessed 21 July 2024).

Rombach, R obin, Andreas Blattmann, Dominik Lorenz, Patrick Esser and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684-10695, https://openaccess.thecvf.com/content/CVPR2022/papers/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_ (accessed 21 July 2024).

Rosenzweig, R oy and David Thelen. 1998. *The Presence of the Past: Popular Uses of History in American Life*. New York: Columbia University Press.

Xu, J iazheng , Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang and Yuxiao Dong. 2023. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (eds.), *Advances in Neural Information Processing Systems*, 36, 15903–15935, https://arxiv.org/abs/2304.05977 (accessed 21 July 2024).