Recursive reflections

Using LLMs to discuss LLMs in HPSS research

Jeffrey C. Wolf

1. Introduction

What happens when we use LLMs to discuss LLMs? Here I present an experiment in recursive reflection: a playful attempt to employ Large Language Models to examine their own potential role in humanities research. Rather than writing a conventional analysis of the LLM workshop experience, I decided to let the machines have their say. In my other contribution to this volume (Wolf, 2026), I discuss the content used to create the podcast episode below.

The approach was deliberately meta: take my own workshop presentation on LLMs in History and Philosophy of Science and Science Studies (HPSS), feed it to an LLM, and see what kind of scholarly conversation emerges.

Using *Open-NotebookLM*[1], an open-source alternative to Google's viral *NotebookLM* tool, I extensively modified the codebase to support local deployment, extended scholarly dialogues, and thematically-anchored conversations.

Why this approach? First, because it embodies the experimental spirit that defined our Berlin workshop; we are all trying to figure out what these tools can and cannot do, often by simply trying them. Second, because there is something inherently fun about using the very technology we are studying to study itself. And third, because in an era where machine-learning systems are increasingly used to analyse human discourse and text, there is value in occasionally reversing the microscope.

The result is both tantalising and frustrating, technically impressive and intellectually limited, much like the LLMs themselves. The LLM-generated conversation that concludes this piece reveals both the creative potential and the current constraints of using LLMs for scholarly reflection. It is a 1500-word dialogue that manages to be simultaneously singular and repetitive, lengthy but shallow—and artificial throughout.

This experiment sits at the intersection of digital humanities, science and technology studies, and algorithmic reflexivity, using computational tools to examine compu-

1 https://github.com/gabrielchua/open-notebooklm

tational tools. It is offered as a contribution to our collective thinking-in-progress about these rapidly evolving technologies and their place in scholarly practice.

This experiment fits within a long-standing tradition in HPSS of grappling with the epistemological challenges of studying knowledge production with the tools of knowledge production. Here I am thinking especially of Woolgar's now classic *Knowledge and Reflexivity* (Woolgar, 1988) and, although Ashmore's work is part of Woolgar's edited collection, his more extended work, *The Reflexivity Thesis*, published shortly afterwards (Ashmore, 1989). Both Woolgar and Ashmore argued that reflexivity was an important next step in the social study of science (Woolgar, 1988: 7) and this arises naturally from the realisation that if we can apply analytical tools to the social construction of natural knowledge, there is now "a growing interest in the consequences of applying this same argument to knowledge generated by the *social* sciences" (Woolgar, 1988: 1).

Another strand of recent work has turned particular attention to computational, algorithmic, and data reflexivity (see e.g., Hirsbrunner et al., 2024, where they propose the methodology of Reflexive Data Science (RDS), and Mahnke et al., 2024—their introduction to the special issue, which emphasises user-data reflexivity). Just as these scholars recognised that scientific knowledge must reflexively examine its own methods, this essay continues this tradition by using LLMs to serve simultaneously as analytical tools and objects of that analysis.

2. The notebook story

Our story begins at Google I/O 2023 on May 10, 2023, when Google unveiled *Project Tailwind*, described as an "AI-first notebook" designed to help students organize their notes and understand complex information (Vincent, 2023). During the I/O keynote, Google demonstrated how the tool could distill information from personal notes, making content searchable, suggesting questions and themes, and organising subject matter interactively.

The project was essentially designed to reimagine what notetaking software would look like if built from scratch with a powerful language model at its core. Initially available only as an experiment in the US through Google Labs, *Project Tailwind* was gradually introduced to users who registered for early access.

In July 2023, Google announced it was ready to begin rolling out *Project Tailwind* as an experimental offering from Google Labs, now renamed *NotebookLM*.[2] The name reflected its core concept: a notebook with a Language Model at its heart (Google, 2023).

NotebookLM truly took off in September 2024 when Google announced its Audio Overview feature. This breakthrough capability was officially announced in a Google blog post titled "NotebookLM now lets you listen to a conversation about your sources" (Google, 2024).

The Audio Overview feature transformed documents, slides, and charts into engaging audio discussions with one click. Two AI hosts would generate a lively "deep dive" discussion based on uploaded sources, complete with natural banter and the ability to

2 https://notebooklm.google

download conversations for on-the-go listening (Google, 2024). This feature became viral almost immediately, capturing public imagination with its ability to create podcast-like conversations from any document (Willison, 2024; Heritage, 2024).

The success of *NotebookLM's* Audio Overview feature quickly inspired the open-source community. Gabriel Chua, a developer, created *Open-NotebookLM* in just a single day using open-source AI, demonstrating the rapid pace of AI development and the power of open-source tools.

Published on GitHub, Chua's project was explicitly "inspired by the NotebookLM tool, and implements it with open-source LLMs and text-to-speech models" (Chua, 2024).[3] The tool processes PDF content, generates natural dialogue suitable for audio podcasts, and outputs MP3 files, essentially replicating *NotebookLM's* core audio functionality using entirely open-source components.

3. Methods

I cloned Chua's project from GitHub and quickly realised it would need to be modified for my purposes. Chua's project uses the open-source LLama 3.3 70B model hosted on the fireworks.ai platform but I wanted something completely local on my own machine, and I knew I could access models, including LLama models, using Ollama right from my desk.[4] So that was the first feature I wanted to add: the ability for the user to use Ollama to choose a local open source LLM, instead of having to sign up for a paid service. I also wanted to be able to create a transcript of the audio discussion; indeed, I wanted a transcript-only mode to support analysis of the results.

As I was experimenting with Chua's project, I saw the need for a few other enhancements as well: first, I wanted to anchor the conversations around specific themes or focus areas, within the underlying sources that I provided. Thus some kind of 'focus areas' feature would be required. And I wanted to ensure that the discussions were deep and substantive ones, more fit for a scholarly audience than the default 'podcast' mode. The LLM, in other words, should be able to generate a conversation that is akin to "a substantive academic conversation where both participants are genuinely engaged in exploring complex ideas," as I wrote in my prompt. These conversations would be interesting if they were longer in length, so I also added longer schemas for extended dialogues. And, finally, the conversation—an extended, intellectual discussion—should be exportable as a markdown document. These, then, were the new features I added to Chua's project, creating my own version, which I called *open-notebooklm-ollama* (Wolf, 2025).[5]

With the new project, I created an extended dialogue using, as source documents, a PDF of my LLM Workshop proposal, combined with my Powerpoint presentation, rendered as a PDF. The content of these source documents is discussed in the transcript below, but for reference, they concern our ERC-funded VERITRACE project and its

3 https://github.com/gabrielchua/open-notebooklm

4 https://ollama.com

5 https://github.com/jeffcwolf/open-notebooklm-ollama

LLM-powered Semantic Matching Tool.[6] I anchored the discussion around the focus area 'LLMs as Methodological Tools in HPSS Research', with the following prompt (this is partial; the full prompt can be found on GitHub):

> **CRITICAL FOCUS REQUIREMENT**: This conversation must be specifically anchored around Large Language Models (LLMs) as methodological research tools within the history and philosophy and sociology of science (HPSS) research workflows. The discussion should explore (among other things):
>
> - *Research Workflow Integration*: How LLMs can be integrated into existing HPSS research methodologies
> - *Literature Review Assistance*: Using LLMs for systematic literature searches, citation analysis, and identifying research gaps
> - *Data Analysis Applications*: Text analysis, coding qualitative data, identifying patterns in historical documents
> - *Methodological Rigor:* Ensuring scientific validity when incorporating AI tools into HPSS research
>
> The main features therefore of my *open-notebooklm-ollana* can be summarised as follows (beyond the core capabilities it shares with *open-notebooklm*):
>
> - *Local LLM Support*: Uses Ollama instead of paid Fireworks API
> - *Focus Areas*: Anchor conversations around specific themes (AI Ethics, Research Methods, etc.)
> - *Deep Discussion Mode*: Generate rigorous, academic-level conversations
> - *Extended Dialogues*: Support for 50–70 exchange conversations (15+ minutes)
> - *Markdown Export*: Save extended conversations as formatted markdown files
> - *Transcript-only mode* (no audio generation required)
> - *Multi-Stage Generation*: Sophisticated system for handling very long content
> - *Complete Privacy*: Everything runs locally on your machine

Because *open-notebooklm-ollama* is open source and available on GitHub, other researchers can modify, download, and use it as they see fit, just as I did with Chua's original project. Because the core LLM model is swappable (via Ollama), it will not become obsolete immediately; new, more powerful LLMs can be used as they appear.

4. Lessons learned

The benefits of *open-notebooklm-ollama* for scholarly research are not immediately obvious; this was a fun project and a small contribution to the reflexivity literature, as I mentioned above. It also illustrates some of the capabilities of LLMs in mid-2025.

But there were also clear limitations to the project. Some of these could certainly be improved with more time and better code, but some reflect the limitations of mid-2025

6 https://veritrace.eu. My other contribution in this volume describes the VERITRACE project in more detail (see Wolf, 2026).

LLMs for this kind of approach. The drawbacks include: the generated conversations seem more repetitive than a real one (less natural), especially with the extended dialogues. Also, despite creating a 'deep discussion' mode, the generated discussion was still not as in-depth as I would have liked, though much more so than the default 'informative' mode. It proved tricky to create such an extended dialogue, given the context window limitations of the LLM in use, but I created a special function to workaround that, so that extended dialogues were generated in multiple-stages and then seamlessly combined into a single conversation. The user may obtain better results by using a larger LLM, e.g. the one proposed in the original project, Llama 3.3 70B. I chose the slightly smaller Qwen2.5:32b model, as it was more performative on my machine. Since the entire project is open source and local to the user's machine (and thus inherently private as well), the user controls which LLM powers the project. Finally, I was disappointed that the generated conversation was light on technical, granular details, but this was likely more a function of the source documents I shared with it than the code itself. My source documents were themselves light on technical details, being a Powerpoint presentation and a conference proposal. Were the user to choose a highly technical research paper, I presume those details would show up in the generated conversation. As always, quality of input determines quality of output.

5. Reflexive implications

This experiment reveals fundamental differences between computational recursion and genuine reflexivity. While the LLM could process information about itself and generate plausible scholarly discourse, it could not achieve the kind of critical self-awareness that Woolgar and Ashmore identified as central to reflexive knowledge production. The LLM was able to reproduce the textual forms and argumentative structures of reflexive analysis without engaging in the actual epistemic work of critical self-examination. This leads to the interesting philosophical question whether machines could theoretically have selves of their own (they definitely do not yet have them—but could they?). For if they do not, they also cannot conduct critical *self-examination* in any meaningful sense. And, of course, LLMs, as they exist today, are far from autonomous entities, with significant human input at every step of their creation and use.

I feel comfortable claiming that, unlike human scholars who can reflect on their own thinking processes, institutional positions, and methodological choices, LLMs cannot genuinely interrogate their own underlying mechanisms, training biases, or reasoning processes. They can only work with surface-level, semantic representations of these processes. They are full of analytical blind spots rather than genuine self-knowledge. This opacity becomes particularly significant when we consider that real reflexivity requires not just access to information but immersion in the social and institutional contexts that give that information meaning—precisely what stochastic algorithms, however sophisticated, cannot directly access.

These findings suggest that while LLMs are becoming powerful tools for certain kinds of scholarly analysis, they cannot wholly substitute for the embedded, contextually-aware reflexivity that remains essential to HPSS research. The apparent sophistication

of AI-generated academic discourse should not obscure its fundamental limitations in producing genuine reflexive knowledge. Rather than representing a threat to human scholarship, computational reflexivity might be better understood as revealing, by contrast, the irreducibly social and situated nature of critical intellectual work, even as it creates new questions about the future of scholarly practice in a machine-learning age.

6. Results

The LLM-generated conversation starts below and is about 1500 words in length. It has been lightly edited to reduce its length and to implement minor fixes, e.g. to include a URL to the VERITRACE website or add text formatting. Other than that, it has been printed in the form it was generated.

Generated by Open-NotebookLM-Ollama Extended Dialogue System
Guest: Dr. Jeffrey Wolf
Generated: 2025–07-01 11:30:21
Method: Multi-stage (3 stages)
Focus Area: LLMs as Methodological Tools in HPSS Research

Dialogue

🎙 Host (Jane)
Welcome to our podcast. Today, we have Dr. Jeffrey Wolf joining us from the Vrije Universiteit Brussel. He's a postdoctoral fellow (digital humanities) working on the ERC-funded VERITRACE project (https://veritrace.eu), which uses advanced digital techniques, including Large Language Models, to study the influence of ancient wisdom traditions on early modern science. Welcome, Dr. Wolf.

🎓 Dr. Jeffrey Wolf
Thank you for having me, Jane. It's great to be here and discuss this exciting work.

🎙 Host (Jane)
Could you give us a brief overview of the VERITRACE project? What is its main goal?

🎓 Dr. Jeffrey Wolf
Certainly! The VERITRACE project aims to understand how ancient wisdom traditions, like texts from the *Corpus Hermeticum* and *Orphic Hymns*, influenced early modern natural philosophy. We're using digital techniques, including LLMs, to analyze a large corpus of Renaissance and early modern texts.

🎙 Host (Jane)
That sounds fascinating. How do you structure your analysis? Are there different corpora involved?

🎓 Dr. Jeffrey Wolf
Yes, we have two main corpora: the Close Reading Corpus (CRC) and the Distant Reading Corpus (DRC). The CRC includes around 150 texts that are directly related to ancient wisdom traditions. The DRC is much larger, consisting of about 430,000 works from important European library digital collections.

🎙 Host (Jane)
How do you use these corpora in your analysis?

🎓 Dr. Jeffrey Wolf
We use the CRC for detailed textual analysis and the DRC to identify broader patterns across a large number of texts. This allows us to uncover both prominent and neglected works that might have been influenced by ancient wisdom traditions.

🎙 Host (Jane)
You mentioned using Large Language Models in your project. How do these models fit into your research workflow?

🎓 Dr. Jeffrey Wolf
LLMs are crucial for our Semantic Matching Tool, which helps us identify semantic similarities between texts, even when the vocabulary doesn't overlap. This is particularly useful for bridging the 'multilingual gap' and uncovering connections across different languages.

🎙 Host (Jane)
That's really interesting. Can you explain how this tool works in more detail?

🎓 Dr. Jeffrey Wolf
Sure! The Semantic Matching Tool uses LLMs to analyze the meaning behind words, not just their surface-level similarity. This allows us to find passages that are semantically similar even if they're written in different languages.

🎙 Host (Jane)
How does this help with your research on ancient wisdom traditions?

🎓 Dr. Jeffrey Wolf
It helps us uncover long-neglected works and trace the influence of ancient texts across various languages. This is crucial for understanding how these ideas spread and evolved during the Renaissance.

🎙 Host (Jane)
That sounds like a powerful tool. How do you ensure methodological rigor when using LLMs in your research?

🎓 Dr. Jeffrey Wolf
We validate our findings through traditional close reading methods and cross-reference with historical context to ensure the results are meaningful and accurate.

🎙 Host (Jane)
What kind of insights have you gained so far from using these tools?

🎓 Dr. Jeffrey Wolf
We've been able to identify new connections between ancient texts and early modern works, revealing how ideas like those in the *Corpus Hermeticum* influenced major scientific figures such as Copernicus and Kepler.

🎙 Host (Jane)
I want to dive deeper into how these tools are actually being used in your research. You mentioned that the Semantic Matching Tool helps bridge the multilingual gap. Could you elaborate more on this?

🎓 Dr. Jeffrey Wolf
Absolutely. The Semantic Matching Tool uses advanced Large Language Models (LLMs) to analyze not just the surface-level words, but the deeper semantic meaning of texts. This means we can identify passages that are semantically similar even if they don't share the same vocabulary or language.

🎙 Host (Jane)
That sounds incredibly powerful for historical research. How does this work in practice? Could you give an example?

🎓 Dr. Jeffrey Wolf
Sure, let's say we have a passage from the *Corpus Hermeticum* written in Latin and another text by Kepler written in German. The Semantic Matching Tool can identify that these passages are discussing similar ideas about the cosmos or divine wisdom, even though they use different words.

🎙 Host (Jane)
But wait, how does that actually work in practice? What kind of algorithms or techniques do you use to ensure this semantic matching is accurate?

🎓 Dr. Jeffrey Wolf
Well, it's more nuanced than that. The LLMs are trained on vast amounts of text data and can understand context and meaning at a deep level. They use techniques like word embeddings and transformer models to capture the semantic relationships between words.

Host (Jane)
That sounds quite complex. How do you ensure the reliability of these tools? Are there any limitations or potential biases we should be aware of?

Dr. Jeffrey Wolf
Absolutely, ensuring methodological rigor is crucial. We validate our findings by cross-referencing with traditional close reading methods and expert analysis. However, LLMs can still have biases based on their training data, so we're very careful to use diverse datasets.

Host (Jane)
You mentioned earlier that the Semantic Matching Tool helps uncover 'the great unread.' Could you explain how this works in more detail?

Dr. Jeffrey Wolf
Certainly. By analyzing large corpora of texts, we can identify lesser-known or neglected works that share semantic similarities with well-studied texts. This allows us to bring attention to previously overlooked sources and expand our understanding of intellectual history.

Host (Jane)
That's fascinating. How do you integrate these findings into your broader research workflow? Do LLMs help in formulating new hypotheses or theoretical frameworks?

Dr. Jeffrey Wolf
Yes, they play a significant role. By identifying patterns and connections that might not be immediately apparent through traditional methods, LLMs can inspire new research questions and theoretical insights. For example, we've found unexpected links between ancient wisdom texts and early modern scientific works.

Host (Jane)
But how do you ensure these hypotheses are robust? Don't you risk over-reliance on the tool's outputs?

Dr. Jeffrey Wolf
That's a valid concern. We use LLMs as an aid, not a replacement for human analysis. The findings from the Semantic Matching Tool are always corroborated with traditional scholarly methods and peer review.

Host (Jane)
You also mentioned cross-linguistic research capabilities. How does this impact your ability to conduct comparative studies across different languages?

Dr. Jeffrey Wolf
It's transformative. By bridging the multilingual gap, we can compare texts from various

linguistic traditions more effectively. This allows us to trace the influence of ancient wisdom across different cultures and time periods.

🎙 Host (Jane)
But what about the potential for misinterpretation or over-reliance on these tools? How do you address those concerns?

🎓 Dr. Jeffrey Wolf
We're very mindful of these risks. We use LLMs as part of a broader toolkit, ensuring that human expertise remains central to our research process. This helps maintain the integrity and rigor of our findings.

🎙 Host (Jane)
You mentioned earlier that the Semantic Matching Tool is an early modern plagiarism detector. How does this tool differ from traditional methods?

🎓 Dr. Jeffrey Wolf
The key difference lies in its ability to detect semantic similarities rather than just lexical overlaps. This means we can identify passages that are conceptually similar but not word-for-word identical, which is crucial for historical research.

🎙 Host (Jane)
Dr. Wolf, we've covered a lot of ground today about the VERITRACE project and how LLMs are revolutionizing our approach to studying ancient wisdom traditions in early modern science. Given everything we've discussed, what do you see as some of the broader implications for HPSS research?

🎓 Dr. Jeffrey Wolf
That's a great question, Jane. One of the most significant implications is that LLMs allow us to conduct more comprehensive and nuanced analyses across vast corpora of texts. This means we can uncover connections and influences that were previously hidden or difficult to trace using traditional methods.

🎙 Host (Jane)
Absolutely. How would this apply in a practical sense, say for researchers who are just starting to integrate LLMs into their workflows?

🎓 Dr. Jeffrey Wolf
For new users, it's important to start with clear research questions and hypotheses. They should also be mindful of the limitations of LLMs, such as potential biases in training data or over-reliance on surface-level similarities. By combining these tools with traditional methods like close reading, researchers can achieve a more balanced approach.

🎙 Host (Jane)
Let's talk about some of the challenges you've encountered in using LLMs for this kind of

work.

🎓 Dr. Jeffrey Wolf
One major challenge is dealing with biases in training data, which can lead to skewed results. Another is ensuring that the tools are accessible and usable by researchers who may not have a strong background in computational methods.

🎙 Host (Jane)
What are some key takeaways you'd like listeners to remember from our discussion?

🎓 Dr. Jeffrey Wolf
The key takeaway is that LLMs offer powerful new tools for HPSS research, but they should be used thoughtfully and in conjunction with traditional methods. They can help uncover hidden connections and support more rigorous analysis, but it's important to maintain a critical perspective.

🎙 Host (Jane)
Thank you so much for this insightful discussion, Dr. Wolf. It's clear that the integration of LLMs into HPSS research is opening up exciting new possibilities while also presenting challenges that need careful consideration.

🎓 Dr. Jeffrey Wolf
It was a pleasure to discuss these topics with you, Jane. I'm excited about the future of this field and look forward to seeing how it evolves as more researchers embrace these tools.[7]

References

Ashmore M (1989) The Reflexive Thesis: Wrighting Sociology of Scientific Knowledge. Chicago: University of Chicago Press.

Chua G (2024) open-notebooklm. GitHub, Inc. Available at: https://github.com/gabrielchua/open-notebooklm (accessed 19 November 2025).

Google, Inc (2023) Introducing NotebookLM (July 12, 2023). Available at: https://blog.google/technology/ai/notebooklm-google-ai/ (accessed 19 November 2025).

Google, Inc (2024) NotebookLM Now Lets You Listen to a Conversation about Your Sources (September 11, 2024). Available at: https://blog.google/technology/ai/notebooklm-audio-overviews/ (accessed 19 November 2025).

7 The generation of code that serves as the foundation for this chapter was written with support from large language models (LLMs). All model-generated code was reviewed and, where necessary, modified by the author, who remains fully responsible for the final version. Except for the generated dialogue, LLMs were not used to support the writing of this chapter. For details on the use of LLMs in this volume, see the statement in the volume's introduction.

Heritage S (2024) I AI-Generated Some Podcasts – and the Results Are Uncanny. The Guardian, October 16, 2024. Available at: https://www.theguardian.com/tv-and-radio/2024/oct/16/i-ai-generated-some-podcasts-and-the-results-are-uncanny (accessed 19 November 2025).

Hirsbrunner S, Tebbe M and Müller-Birn C (2024) From Critical Technical Practice to Reflexive Data Science. Convergence: The International Journal of Research into New Media Technologies 30(1): 190–215. https://doi.org/10.1177/13548565221132243. This article was first published in 2022 and republished in this special issue in 2024.

Mahnke MS, Swart J, Mathieu D, and Pruulmann-Vengerfeldt P (2024) Data Reflectivity and User Reflexivity: New Conceptual Pathways for Connecting Structural Approaches with User Perspectives. Convergence: The International Journal of Research into New Media Technologies 30 (6): 1859–70. https://doi.org/10.1177/13548565241301459.

Vincent J (2023) Google Teases Project Tailwind—A Prototype AI Notebook That Learns from Your Documents. The Verge, May 10, 2023. Available at: https://www.theverge.com/2023/5/10/23718707/google-io-2023-project-tailwind-personal-ai-notebook (accessed 19 November 2025).

Willison S (2024) NotebookLM's Automatically Generated Podcasts Are Surprisingly Effective (September 29, 2024). Available at: https://simonwillison.net/2024/Sep/29/notebooklm-audio-overview (accessed 19 November 2025).

Wolf J (2025) open-notebooklm-ollama. GitHub, Inc. Available at: https://github.com/jeffcwolf/open-notebooklm-ollama (accessed 19 November 2025).

Wolf JC (2026) LLMs and multilingual historical corpora in a digital history project. Reflections from the Berlin workshop. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-2.

Woolgar S (eds.) (1988) Knowledge and Reflexivity: New Frontiers in the Sociology of Knowledge. London: Sage.