**The Retrodiction of the Urtext**

Patrick Downey

DIGI406 Essay

November 2024

For my independent project for a Postgraduate Certificate in Digital Humanities, I used a custom fine-tuned large language model (LLM) to complete fragmentary sections of lost plays by the 5th Century BC Greek Tragedian Euripides. I then evaluated the resulting texts for their plausibility in style and content. Traditionally, the reconstruction of fragmentary works has been conceived as a combination of textual scholarship and creativity (pg4, Lamari et al). In order to establish the legitimacy of my method of reconstruction using a LLM, this essay will ask two interrelated questions about the nature of LLMs and their outputs: 1. What is the artisic and creative value of LLMs? And 2. What is the scholarly value of LLMs for literary-historical research? The case study of my project will be used to show that LLMs have interesting potential uses in both these areas, albiet while challenging traditional notions of authorship, and historicity. My discussion and arguments assume some knowledge of LLM architecture and applications.

**Part 1: LLMs and creativity**

Ted Chiang’s essay from August 2024 in the New Yorker magazine criticises generative AI and its potential to create art. For Chiang, an accomplished science fiction author whose own works speculate about the impact and and ethicial questions around possible future technologies, generative AI and LLMs in particular are a ‘fundamentally dehumanising technology’ that ‘reduces the amount of intention in the world.’ This assertion is made against the background of Chiang’s general definition of art as something that results from making a lot of choices – he gives the example of writing fiction with involves choice about every word – tens of thousands of choices as compared to the much fewer choices involved in writing a prompt for an LLM to write a story. He also reminds us that LLMs such as ChatGPT have no true intention to communicate despite their aparent skill with language, and dismisses them as a ‘turbo-charged autocomplete’. (Chiang)

I find his argument persuasive, as far as the assessment of the artistic value of generative AI outputs when they are considered in isolation. No one wants to listen to an entire suno.ai song or read chatgpt poems after the novelty has worn off. However, when such work is valid and interesting, and attracts attention and discussion, it is often because it is presented in tandem with the prompts used to create it. In such works, the hybrid collaborative nature of the process betwen human and AI is highlighted and celebrated, and the object of aesthetic and critical interest is this collaboration, not merely the deracinated output. An example is one of the first exhibitions of generative AI art using modern neural networks, *Neurography* by Mario Klingemann, which presented generated visual art while documenting his process in generating it (The Photographers Gallery). An emphasis on process over final product is not an uncommon theme in contemporary art (Tate).

Although my project did not begin with a goal to produce ‘art’ as such, the fact that I am attempting to imitate and expand on the work of one of history’s most celebrated playwrites neccessitates an engagement with Chiang’s aesthetic and technical critique of generative AI outputs. It is a very limited kind of creativity that I use the LLM for in my project, and one that is tightly bound by the constraints of imitating a specific author from a particular time and place, and by the context of the existing fragments. Contra Chiang’s assumption, my process did involve many choices as I selected material for reconstruction and refined my techniques of fine tuning and prompting. The passages that I have fully reconstructed are made from a large proportion of original text, which itself is part of the prompts, and an example of the hybrid collaborative works that I argue above as being aesthetically valuable and interesting.

The transformer architecture on which LLMs are based on uses a probability distribution of possible tokens for these predictions, a process that has an inherent ramdomness that can be somewhat callibrated by a parameter called temperature. This randomness has been used to critique LLMs as ‘stochastic parrots’ with outputs that are inherently meaningless due to their lack of communicative intent (Bender et al., 2021). However, it is clear from the history of art in the twentieth century that the deliberate use of randomness is a legitimate artistic method, for example Marcel Duchamp’s work 3 Standard Stoppages, which used the random shapes created by a dropping a thread onto a canvas, or Jackson Pollock’s paint splatter abstract expressionism.

**LLMs and Authorship**

As for the lack of communicative intent behind the text produced by LLMs, this does not necessarily mean that it lacks meaning. There was an influential movement in twentieth century literary theory that sought to de-centre the role of the author in literary analysis, represented by Roland Barthes’ famous and provocative essay titled The Death of the Author. This was a intentional break with the ordinary understanding of the role of the author, which still more or less persists today, that Barthes saw as ‘the epitome and culmination of capitalist ideology’ with its focus on the prestige of the individual. (pg143, Barthes) Barthes claims that:

*‘We know now that a text is not a line of words releasing a single ‘theological’ meaning (the ‘message of the Author-God) but a multi-dimensional space in which a variety of writings, none of them original, blend and clash. The text is a tissue of quotations drawn from the innumerable centres of culture’*(pg146,Barthes)

This passage seems remarkably prescient and descriptive of LLMs and their voluminous training data that is recombined to produce new text. Barthes denies that we can interpret anything about an author’s interiority from writing that they have produced as this is subsumed and blurred into the characters and language (which itself is collection of impersonal signs and rules). We are separated from the author by space and time and have to create our own meaning as readers:

*‘…but there is one place where this multiplicity is focused and that place is the reader, not, as was hitherto said, the author. The reader is the space on which all the quotation that make up a writing are inscribed without any of them being lost; a text’s unity lies not in its origin but in its destination.’*

This theory explains well the surpising lack of uncanniness in the experience of reading LLM generated text. Although we know that there is no traditional author behind such ‘writing’, we easily interpret and ascribe meaning to it due to its fluency, coherence, and accuracy in combining the ‘signs’ of language. The popular chatbot format is very intuitive, and seems like a significant breakthrough in human-computer interaction design. The ease with which we read meaning into LLM outputs leads to a tendency towards anthropomorphise these models, to project and imagine an author behind them, even to invite speculation about consciousness or sentience of models. Barthes’ theory helps us reject this anthropomorphising by decoupling language use from traditional personal notions of authorship.

**Many Homers**

A relevent example is from the study of Homer and the so called ‘Homeric Question’ regarding authorship of the Iliad and the Odyssey. An influential theory in this field is the ‘oral-formulaic theory’, which asserts that the repititive poetic formulas found in these poems are proof that they come from oral traditions and represent the work of a collective of poets rather than one single author. The formulas act like modular components that the oral poets of ancient Greece would recombine and expand on in semi-improvised performances to tell the familiar stories of the Trojan and war and its aftermath. ‘Homer’ may have been just the name given to this oral tradition. (Struck)

It seems significant that these very important works that influenced the development of all of European literature, especially the greek tragic poets such as Euripides, were likely written in a collaborative way without a single author. Indeed the process described by the oral-formulaic theory of recombining and improvising with modular formulaic language is somewhat analogous to the way that LLMs recombine tokens from their training data to produce new language. This example and analogy show that far from such language being meaningless because it lacks a singular author figure, it can in fact be enormously influential and significant.

Although there is an observable tendency towards standardisation of Homers works in the manuscripts that have come down to us from antiquity, there remain differences and inconsistancies among the multiple versions. There is no one true version of the these works. It is the role of scholars to investigate and compare these differences through the techniques of textual criticism, not to discover an original ‘urtext’, which will remain only an ideal, but to understand them in their state of multiplicity. This is the project of scholars such as Gregory Nagy and his Homer Multitext Project, which leverages digital technology to collect and display different versions of the poems (The Center for Hellenic Studies).

This is the approach to scholarship that is neccessary to understand the value of my Euripidean LLM project. Unfortunately, manuscripts of Euripides’ works are not as numerous as Homer’s. Of the 90 or so plays that he is thought to have written, only 18 are complete and extant, and a number of fragments remain for some of the others (pg5, Cropp et al.). So one problem that scholars of Euripides are confronted with is a lack of evidence - a lack of raw material for the work of textual criticism. The next part of this essay will argue that LLMs have potential as a tool for generating useful hypotheses about the text of the lost works of Euripides, and other similar tasks in literary-historical research.

**Part 2: LLMs as simulation tools for scholarship**

I will begin by introducing a theoretical framework for understanding LLMs as *simulators* and their output as *simulacra,* and then discuss why this view of LLMs is useful for describing their practical value for literary-historical research. I will argue that my project produces simulacra of Euripides’ lost plays that is analagous to the computational modeling and simulation methods used in the natural sciences, where predictions are made about the future state of natural phenomena or systems based on known inputs. My thesis is that if we understand language as an abstract system that is analogous to the systems of material phenomena studied by the natural sciences, then LLMs can perform a similar predictive function by using known inputs to generate new text, and that this can then be used to speculate about the possibility space of unknown text.

**Simulators in theory**

The idea of understanding LLMs as simulators and their output as simulacra has been proposed by a blogger with the nom de plume Janus writing on the LessWrong website. Janus refers to Nick Bostrom’s work on defining ontologies for understanding different kinds of AI systems – the oracle (question answerer) the genie (command executor), the soveriegn (full agent) and the tool (no goal oriented behaviour) (pg177, Bostrom), and finds that LLMs can superficially appear like each of these depending on the prompts used to interact with them, but that none of these categories fully capture the nature of LLMs. By making a clear ontological distinction between the LLM itself and its output-instances, Janus resolves the problem of this variability in LLM characteristics, describing the former as a powerful simulator that produces simulacra as outputs, which are the things that then display these different characteristics. LLMs are optimised with machine learning methods towards accurately modeling their training distribution – that is, to simulate parts of their dataset. ‘They are able to not only repoduces the behaviour of its demonstrators but produces the behaviours of an inexhaustable number of counterfactual configurations’ (Janus). Importantly, this is a process that happens over time, with new inputs, (in the case of a chatbot these would be the user’s responses) causing the LLM to make adjustments to its simulation of a conversational partner.

If we apply this framework to my Euripides fragment completion LLM, we can understand the prompts that I feed to the model as setting the parameters for the simulation process, which proceeds as I add each subsequent lines, either authentic fragment or LLM generated, producing a simulacrum of a complete Euripidean text. To what extant can these texts be said to be accurate? What is the precise relationship between the Euripidean simulacra that I have produced and the original lost texts? These are questions of the epistomological and ontological status of simulacra and even representation in general, which have been examined by philosophers since Plato, and influentially in the twentieth century by Jean Baudrillard and Gilles Delueze, among others.

Janus uses the term ‘simulacra’, quoting from Baudrillard to emphasise that LLM outputs do have not indepedent existence and do not refer to a real object. For Baudrillard, simulacra are a false representation that bear ‘no relation to any reality whatsoever’ and create a confused state of postmodern culture called the ‘hyperreal’. (pg3, Baudrillard) This is similar and perhaps derived from the position of Plato who sees any representation as an initial departure from the true reality of the ideal forms, and simulacra, which he discusses in *The Republic* as *mimesis* or imitative art, and in *The Sophist* as *phantasma* – the copy of the copy, as another distortion that is even further removed from the real (Plato).

I wish to challenge these views by showing that we can use LLMs as legitimate tools for historical research by assigning their simulacra outputs a hypothetical or ‘virtual’ reality, that can open up new possibilities for scholars to evaluate. This is the approach of Delueze to the question of simulacra, who sees them as a way to break down the hierarchical schema and singular notion of truth that philosophy has inherited from Plato. ‘The simulacrum is not a degraded copy. It harbours a positive power.’ (pg262, Delueze). This is ‘the power to affirm divergence and decentering’ (pg265 ibid) and to ‘institute the chaos that creates’ (pg266 ibid).

One of Deleuze’s foremost modern interpreters is Manuel Delanda, who develops Delueze’s idea of the *virtual* to describe a system’s possible states and behaviours, distinct from the *actual*  which is what exists when these potentials have become material and concrete (pg30, Delanda). This can be applied as a way of understanding the simulation process that LLMs enact, in which the virtual is the space of possible text outputs which over the course of the simulation becomes actualised as the LLM generates text. This space of possible outputs in an LLM is called the *latent space* and is constructed of vector embeddings that represent the relationships between tokens (Robinson et al). I understand it to be virtual in a Deluezean sense in that it is abstract but contains real potentialities.

Consider a dialogue between a user and an LLM based chatbot, during each point in a conversation an LLM chatbot has a large number of different possible simulacra that it can generate that are consistent with the preceding context. According to AI researcher Murray Shanahan, ‘the stochastic nature of autoregressive sampling means that, at each point in a conversation, multiple possibilities for continuation branch into the future.’ (pg5, Shanahan). Shanahan uses the concept of a ‘superposition of simulacra’ in reference to Janus’ simulation ontology to emphasise that the simulation is a dynamical process across time in which new possibilities are being opened and closed with each user input.

*‘At each node, the set of possible next tokens exists in superposition, and to sample a token is to collapse this superposition to a single token. Autoregressively sampling the model picks out a single, linear path through the tree. But there is no obligation to follow a linear path. With the aid of a suitably designed interface, a user can explore multiple branches, keeping track of nodes where a narrative diverges in interesting ways, revisiting alternative branches at leisure.’* (pg4, Shanahan)

This is a description of the traversal of a Deluezean virtual space. It is also a description of part of my LLM assisted method of completing Euripidean fragments, as I often used multiple generations in an iterative process to produce final text that I was satisfied with. It is possible to imagine a user interface that makes this process much easier and more explicit, and it is to such practical questions that I will now turn.

**Simulators in Practice**

Now that I have established a theoretical foundation for understanding LLMs as simulators I want to describe the practical uses that such a tool can have for research in the humanities. Modelling and simulation methods already have an important place in the toolkit of the natural sciences and it seems to me that LLMs can fill a similar role in fields that are primarily concerned with language such literary history and textual criticism.

Consider the problem that my Euripidean LLM project engages with – gaps in the textual record, caused by ‘mould, worming, water and fire … politics, cultural upheavals, migration and wars.’ (pg5, Lamari et al.) Such lacunae constitute a kind of negative space, a literary-historical dark matter, that we can see the shadow and influence of in other ancient writers and in the fragments that have been preserved of the plays. As classicists Collard, Cropp and Lee write in their commentary on Euripides’ fragmentary plays ‘speculative reconstruction of such plays is wholly justifiable, indeed an almost irresistible challenge’(Collard). Traditionally, reconstruction has been approached as a combination of textual scholarship and creativity, with various balances between the two, but usually with a goal of ‘imagining the intention of its producer’ (Lamari et al.). Informed by Barthes, we can dispense with this naive notion of authorship and orient on the text itself. If we consider the problem of fragmentary works to be an empirical problem, where language is the material substrate and the reconstruction of a complete text is the end goal of an experimental process, then LLMs can be a tool that assists this experimentation by stochastically generating hypotheses that are bounded by existing knowledge (the training data), including a rigorously quantified but impersonal model of style. The continuation of sequences of tokens, i.e. language, in a plausible way is the very goal that LLMs have been optimised for, and so the problem of a fragment of text – an incomplete sequence of language, seems to perfectly matched to this tool. Recall Chiang’s criticism of LLMs as mere autocomplete machines, here is an appropriate use for such a machine - his critique is already incorporated into my project.

In the natural sciences, computer simulation methods are well established as a way to investigate complex phenomena and predict the behaviour of dynamical systems. For example, the discipline of meteorology uses simulations of weather systems using models that are provided with known variables such as temperature and pressure in order to make predictions about the future state of the weather system. My claim is that LLMs are doing something analogous, although in the specific case of completing Euripidean fragments, are retrodicting a past state rather than predicting a future state. This claim is stronger than Janus and Shanahan’s initial framing of LLMs as simulators and their output as simulacra, because I am saying that the simulacra actually have some empirical reality albeit a probabilistic and virtual one. A meteorological model produces a simulation that becomes actualised as a forecast that is then considered by and influences the decisions of farmers and beachgoers, in the same way, an LLM simulacrum of a Euripides play can be a useful and influential object to be considered by Classicists.

This claim rests on a number of assumptions, including of the nature of language as a system that can be quantified and modelled separately from actual language users. This is not uncontroversial in linguistics and philosophy of language, and I discussed in the first part of this essay some of the demands for communicative intent that are made against it. However the remarkable success of LLMs at creating coherent language seems to provide some evidence for Barthes’ understanding of language as ‘know[ing] a subject, not a ‘person’, and this subject, empty outside of the very enunciation which defines it, suffices to make language hold together, suffices, that is to say, to exhaust it.’ (pg145, Barthes)

I make no promise to strict accuracy, if accuracy is understood as the discovery of a definitive original urtext. Medievalist and critical philologist Bernard Cerquilini writes ‘in praise of the variant’, and it is in this way that we must understand the role of scholarship to explore and describe difference and variability. LLMs are a new and interesting technical means to produce the plausible variability that is the necessary material for this scholarship. This legitimacy of this method of scholarship can be summarised by the following premises:

1. We cannot access the intentionality of an author, but must rely on the text itself.
2. Text can be used as data to train machine learning models such as LLMs
3. Premise 1 and 2 imply that language is an abstract system that can be analysed without reference to actual language users
4. LLMs effectively simulate language and are continually generating language based simulacra
5. These simulacra have a ‘virtual’ ontological status that is analogous to potential outcomes of complex physical systems
6. They deserve attention and consideration as speculative hypotheses

How does this argument connect with the discussion of LLM creativity in the first part of the essay? In response to Ted Chiang’s critique of the artistic value of generative AI, I introduced the idea of human-AI collaborative work that does not hide the process of generation as characterising the most interesting ‘AI art’. This collaborative and process-based approach can also describe the effective scholarly use of LLMs that I have elaborated. The two aspects of reconstruction of lost works, creativity and scholarship, can both by stimulated by this new technology, but the benefits can only be realised with substantial human input and supervision.

**Bibliography**

Baudrillard, Jean. *Simulacra and Simulation*. 1995, https://doi.org/10.3998/mpub.9904.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. *FAcct ’21*. <https://doi.org/10.1145/3442188.3445922>

Chiang, Ted. “Why A.I. Isn’t Going to Make Art.” *The New Yorker*, 31 Aug. 2024, www.newyorker.com/culture/the-weekend-essay/why-ai-isnt-going-to-make-art/

Collard, et al., editors. *THE PLAYS OF EURIPIDES SELECTED FRAGMENTARY PLAYS: I*. Aris and Phillips Classical Texts, 1995.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. 2014.

DeLanda, Manuel. *Intensive Science and Virtual Philosophy*. Bloomsbury Publishing, 2013.

Delueze, Gilles. *Logic of Sense*. Bloomsbury Publishing, 2004.

Cerquiglini, Bernard. *In Praise of the Variant: A Critical History of Philology*. JHU Press, 1999.

Janus. *Simulators*. 2 Sept. 2022, www.lesswrong.com/posts/vJFdjigzmcXMhNTsx/simulators.

Lamari, Anna A., et al. *Fragmentation in Ancient Greek Drama*. Walter de Gruyter GmbH and Co KG, 2020.

“Marcel Duchamp. 3 Standard Stoppages. Paris 1913-14 | MoMA.” *The Museum of Modern Art*, www.moma.org/collection/works/78990.

“Mario Klingemann - Neurography | the Photographers Gallery.” *The Photographers Gallery*, thephotographersgallery.org.uk/whats-on/mario-klingemann-neurography.

Plato*. The Internet Classics Archive | the Republic by Plato*. classics.mit.edu/Plato/republic.11.x.html.

Robinson, Michael, et al. “The Structure of the Token Space for Large Language Models.” *arXiv.org*, 11 Oct. 2024, arxiv.org/abs/2410.08993.

Shanahan, Murray, et al. “Role Play With Large Language Models.” *Nature*, vol. 623, no. 7987, Nov. 2023, pp. 493–98. https://doi.org/10.1038/s41586-023-06647-8.

Struck, Peter T. *Greek and Roman Mythology - Homer*. www2.classics.upenn.edu/myth/php/homer/index.php?page=comp.

Tate. “Process Art | Tate.” *Tate*, www.tate.org.uk/art/art-terms/p/process-art.

“The Homer Multitext Project - the Center for Hellenic Studies.” *The Center for Hellenic Studies*, 28 Jan. 2023, chs.harvard.edu/curated-article/gregory-nagy-the-homer-multitext-project.