[Under review by Bloomsbury Academic as part of *New Directions in Digital Textual Studies: Scholarly Editing, Book History, and Curation in Conversation*, edited by Christopher Ohge and Kristen Schuster]

Honey, AI Shrunk the Archive:

Artificial Intelligence as Compression Algorithm

Jon Ippolito

The digital age reshaped the foundations of knowledge; generative AI promises to overturn them completely. The advent of academic repositories like ARTstor and Europeana alongside the quotidian world wide web challenged the primacy of scholarly narrative in the tradition of authors like Erwin Panofsky and Jacques Derrida. Yet while these database-driven methods offered new ways to discover knowledge buried in primary sources, large language models threaten to replace sources with a compacted surrogate that generates rather than discovers knowledge. Understanding the statistical engine behind these lossy substitutes for the archive can help us choose the best strategies for combating the homogenising tendencies of transformer models.

How AI Bends the Archival Arc

Scholarship has always had a presumed trajectory that starts with a collection of data. Exactly what the researcher discovers in her purview depends on the discipline: an astronomer might find evidence of a black hole in a star's spectrographic signature, while a historian might find evidence of fierce anti-clericism in something William Blake scribbled in a manuscript margin. Domain specifics notwithstanding, the researcher is likely then to hunt for additional evidence in the same range, or contrast the discovery with discoveries from other ranges, or wrap the discovery in an explanatory framework. Regardless of which 'archive' a scholarly process starts with, its endpoint has traditionally been some sort of publication, whether gracing the pages of the *Antioch Review* or the *Open Journal of Astrophysics*.

From Galileo's telescope to Anne Kelly Knowles' viewshed analyses, emerging technologies and techniques have forged spurs alongside this well-worn path and expanded the scope of valid evidence. In 1999, media scholar Lev Manovich[1] pointed to the displacement of a traditional narrative approach to humanities scholarship by the increasing influence of databases, which culminated a decade later in an academic obsession with 'big data'. From ArcGIS to the blockchain, data structures became both a new ground for archival research as well as a potential archive in which to publish it.

[1] Lev Manovich, 'Database as Symbolic Form', *Millennium Film Journal* No. 34 (Fall 1999). http://www.mfj-online.org/journalPages/MFJ34/Manovich_Database_FrameSet.html, accessed 12 November 2024.

Creators have traditionally fed the opposite stream of this cultural trajectory. Photos were shot with a camera and then digitised, or born digital when snapped with someone's phone; images selected by the photographer then circulated through the Internet via Instagram or Flickr; those deemed sufficiently notable to preserve for the future made their way into an archive or museum. At that final destination, conservators ingested them into databases or preservation software, and thus presumably into the historical record. Rather than starting with a collection and producing original research as researchers do, artists start by producing original research,[2] publishing what they deem worthy of public view in a gallery or online, and then (with sufficient talent and luck) selling or donating their work to a public or private collection.

As artist Eryk Salvaggio points out, the advent of generative AI could more than any other recent technology bend the archival trajectory for both scholars and creators back upon itself.[3] This bend is not a distortion due to bias in the archival collection--a valid concern in itself--but a flip of the arrow of contingency. Whether trained on a highly curated photo collection or the billion web pages of the Common Crawl, large language models start by atomizing content in the archive and then compressing it into an engine that can produce new artefacts derived from that original content. Salvaggio has described generative AI as 'an archive in reverse', in that culture produced by AI actually starts with curation (an archive of source material) and ends with creation (an artefact generated from this archive). The implications of a twist in the archival trajectory could be dramatic for both scholars and creators alike, calling into question deeply held beliefs about ground truth and creativity.[4]

The Shifting Repositories of Information

The word 'scholarship' has its roots in the concept of an illuminating conversation[5], so it's natural that the original paradigm for conveying knowledge would be discursive. Memory palaces and other artifices provided information architectures--almost literally--to help bards recall the scenes of a story or parts of an argument. Over time, as these narratives evolved into written forms, their genres multiplied and became codified, but the overt goal was generally to convey complex ideas in an accessible manner (even if the covert goal might have been to dominate a debate or rise in an intellectual hierarchy).

Manovich's 'Database as Symbolic Form' argues that the rise of software shifted the paradigm of knowledge stockpiles from narrative to database. He posits that the emergence of

---

[2] Artists draw on whatever experience is available to them, like memories or feelings or what they've seen in museums. While AI apologists compare this to scraping existing archives for inspiration, artistic research is far more nebulous and expansive. For more on artistic research, see Joline Blais and Jon Ippolito, *At the Edge of Art* (London: Thames & Hudson, 2006).

[3] Technically Salvaggio only described the reversal of the archival trajectory for creators: 'Generative AI is digital humanities in reverse. Any description of an archive becomes a formula for the production of similar content. This reversal makes digital humanities an essential lens for understanding the cultural questions surrounding AI. To do that, we have to start imagining archives in reverse.' Threads, 3 September 2024. https://www.threads.net/@cyberneticforests/post/C eCLsUIR6S, accessed 16 November 2024.

[4] This fact complicates both sides of the copyright debate, adding a new wrinkle to the debates over the nature of authenticity and attribution. At the time of writing US law provides no recourse or control over the creation and distribution of these automated offspring.

[5] The Greek word σχολή (skholḗ), which originally meant spare time, later became 'conversations and the knowledge gained through them during free time; the places where these conversations took place.' 'Scholar', Wikipedia, https://en.wiktionary.org/wiki/scholar, accessed 16 November 2024.

databases decentered the academic narrative in favour of a more modular, data-driven approach to scholarship. Manovich states, 'If after the death of God (Nietzsche), the end of grand Narratives of Enlightenment (Lyotard) and the arrival of the Web (Tim Berners-Lee) the world appears to us as an endless and unstructured collection of images, texts, and other data records, it is only appropriate that we will be moved to model it as a database'. In Manovich's view, the database becomes a matrix for grounding scholarship, for which coding an interface to a database on Salman Rushdie might be more significant than writing an essay about his literary relevance.[6]

In the mid-2000s, cryptocurrency enthusiasts popularised distributed ledgers, decentralised kinds of databases that evolved into programmable blockchains like Ethereum. These platforms enabled artists and scholars to create smart contracts that modify, fork, or merge records upon access, transforming transactions into dynamic components of the creative process. For the most experimental crypto-artists, these transactions are not mere exchanges but integral to their medium. While some archives experimented with blockchains, many faced challenges due to their computationally intensive nature and inherent risks.[7]

The latest evolution in accumulated knowledge is generative AI, epitomised by large language models like ChatGPT. These models consolidate vast amounts of online knowledge into standalone chatbots, offering a new paradigm of interaction. However, this shift raises concerns about Euro-ethnic bias and the limitations of accessing the deep web.[8] Large language models predict the likelihood of a word following a given sequence of words, using complex neural networks to analyse and generate human-like text. More cogent to a discussion about the evolution of cultural repositories, however, is the fact that ChatGPT's one-field interface presents it as an all-knowing oracle[9], threatening to replace academic scholars and creative professionals with a superior intelligence that digests and synthesises the sum of human knowledge. Yet this personification conceals mechanistic frailties of actual large language models, from data poisoning to reinforcing stereotypes to numerous other downsides.[10]

---

[6] Notable examples of digital humanities databases include the Perseus Digital Library, a collection of ancient texts and artefacts, and the Walt Whitman Archive, which provides comprehensive access to Whitman's works and related materials. Beyond database interfaces, computational humanities projects like the Mining the Dispatch project use data analysis to extract insights from cultural records, presenting findings through charts and visualisations that reveal patterns and trends that otherwise might be hidden in traditional narratives. See also Miller Prosser's essay in this volume (Chapter 6) for more on the database paradigm in textual editing.

[7] Notable examples include the Zentrum für Kunst und Medien (ZKM), which accidentally burned two valuable NFTs by sending them to an inaccessible wallet; and ARCHANGEL, an initiative of the University of Surrey with the UK National Archives that explored blockchain's potential as a 'trusted archive of digital public records' before recognizing its limitations.

[8] Even if AI harvests the narrow slice of human experience that is posted online, LLM training data does not include the 'deep web' inaccessible to web crawlers, which some researchers estimate to be 500 times the size of the harvestable web. Michael K. Bergman, 'The Deep Web: Surfacing Hidden Value,' BrightPlanet, 2000. https://resources.mpi-inf.mpg.de/d5/teaching/ws01 02/proseminarliteratur/deepwebwhitepaper.pdf, accessed 16 November 2024.

[9] For more on how chatbot interfaces obscure the vagaries of generative AI models, see Jon Ippolito, 'Why you Should Generate AI Images in Your Classroom', Still Water blog, 8 January 2024, https://blog.still-water.net/why-you-should-generate-ai-images-in-your-classroom, accessed 16 November 2024.

[10] For a compact list of potential AI downsides, see the IMPACT RISK framework created by the author at https://AI-Impact-Risk.com, accessed 15 November 2024.

Beyond the Animate AI Metaphor

To reconcile the interdependence of generative AI and the archive, we need to challenge some prevailing assumptions about the lifelike qualities of large language models. When DALL-E 2 and ChatGPT exploded on the scene in April and November 2022, respectively, these tools seemed unlike anything that came before--not just because they vaulted ahead of prior attempts to generate images and text with artificial intelligence, but also because the astronomical scale of these systems--trained on billions of documents found in the web--was all but incomprehensible to users accustomed to managing a few thousand documents on their hard drives. The marketing rhetoric of OpenAI and its successors in the AI industry fed the impression that these tools had no precedent, often falling back on anthropomorphic metaphors to reinforce the impression that never-before-seen beings were emerging from torrents of matrix multiplication. In Silicon Valley's view, AI was 'growing up fast' and 'would understand exactly what you wanted, and it would give you the right thing'.[11]

At the same time, critics like Emily Bender[12] and Maha Bali[13] have warned that envisioning algorithms as children or animals can foster misunderstandings about their capabilities and limitations. When we describe AI systems as 'thinking' or 'understanding', we risk attributing emotions, intentions, and consciousness to systems that are no less mechanistic than a dishwasher. It's a short hop from presuming AI has consciousness to presuming it has a conscience. That could result in misplaced trust in AI agents to make moral or ethical decisions without human oversight, whether betting on the stock market or reading an x-ray.

Anthropomorphism also obscures accountability. For years, vehicle companies neglected to mention the remote assistance provided to their so-called 'autonomous' vehicles by humans behind the scenes.[14] The very term 'self-driving cars' insinuates that these machines have a self, and presumably therefore free will and possibly even an ethical compass. On the flip side, humanising AI also risks exacerbating fears about it surpassing human intelligence, fueling dystopian narratives about killer robots that can hinder productive discourse about the integration of AI into society.

As complex as a trillion-parameter large language model might be, we are only obscuring its mechanism by comparing it to an organism or superbeing. Organic life is complex on a vast array of levels, from the electron-transport chain in the mitochondria of a cell to insulin's effect on the bloodstream to the unique body language of Italian gestures. Chatbots are complex on

---

[11] Quotes from Diane Ackerman and Larry Page in Bernard Marr, '28 Best Quotes About Artificial Intelligence,' Forbes 25 July 2017. https://www.forbes.com/sites/bernardmarr/2017/07/25/28-best-quotes-about-artificial-intelligence, accessed 15 November 2024.

[12] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell, 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜', FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (March 2021), pp. 610-23. https://doi.org/10.1145/3442188.3445922, accessed 15 November 2024.

[13] Anuj Gupta, Yasser Atef, Anna Mills, and Maha Bali, 'Assistant, Parrot, or Colonizing Loudspeaker? ChatGPT Metaphors for Developing Critical AI Literacies,' arXiv:2401.08711, 15 Jan 2024. https://arxiv.org/abs/2401.08711, accessed 15 November 2024.

[14] Compare to Cade Metz, et al., 'How Self-Driving Cars Get Help From Humans Hundreds of Miles Away,' The New York Times, 3 September 2024. https://www.nytimes.com/interactive/2024/09/03/technology/zoox-self-driving-cars-remote-control.html, accessed 15 November 2024.

pretty much one level: the billions of matrix parameters involved in training and generating outputs. To some observers, LLMs exhibit emergent linguistic or even conceptual structures. But these structures are not the product of a heterogeneous architecture. They are the product of a homogenous mechanism that has more in common with the messiness of a bowl of spaghetti than the intricacy of a living cell. And since these models are ultimately assemblages of wire and silicon, comparing their behaviour to actual machines might be more appropriate than reaching for animistic comparisons. And their mechanistic counterparts have the virtue of being a lot simpler than organic ones.

AI as a Cup of Coffee

Emily Bender's provocative disparagement of large language models as 'stochastic parrots'[15] has the benefit of countering the narrative of large language models as superhuman sentients. Unfortunately this metaphor's association with another life form--even if one associated more with meaningless chatter than intelligent conversation--distracts from the mechanistic qualities of generative AI. The 'stochastic' half of her moniker, on the other hand, reflects a core tenet of how generative AI works. Stochastic systems are composed of so many random bits that its microstates can't be predicted but its overall statistics can. So let's consider something less animate and more stochastic for our metaphor: a cup of coffee.

This metaphor goes deeper than caricaturing AI models as 'full of hot air' or capable of 'burning their users'. Imprecise metaphors aside, a cup of coffee, bowl of soup, or even a balloon shares mathematical properties with a large language model that help reveal what's going on inside generative AI's 'magic box'. Mugs of coffee and large language models are both thermodynamic systems that have achieved a state of equilibrium thanks to trillions of probabilistic interactions. For coffee, these interactions are molecules of caffeine and milk that move in seemingly random directions in a physical cup. For an LLM, they are numerical weights that move in seemingly random directions in a mathematical space.

While such desultory beginnings would seem to yield nothing more than useless randomness, applying a constraint to such 'stochastic' systems can rein in these haphazard processes and, surprisingly, deliver the entire state to a stable equilibrium. For a drink, this constraint is nature's commandment that drives all systems to get cooler and more disordered; it's why your morning brew gets cold and the 'latte art' poured by your barista eventually dissipates.

Although a large language model is a human construct rather than a natural system, AI engineers can still 'train' the system into equilibrium by adding an artificial constraint. To do this, they test to see how well the model identifies some input--say, photographs of cats--while at the same time randomly tweaking the weights. A 'cost' function measures the incremental gain or loss in accuracy after each tweak; beneficial tweaks are retained and detrimental ones discarded. Automating these tweaks across the ten-thousand CUDA cores on each of ten-thousand GPUs enables the model to explore a huge space of possible weights, just as the septillion (10^24) molecules in a cup of coffee are free to explore a vast space of possible positions.

A coffee or chatbot may seem a single entity to macroscopic creatures like us, but both contain the possibility of a myriad different microstates, whether the specific jumble of

[15] Bender et al, ibid.

molecules in a latte or the specific word sequences produced by a chatbot. (Here the 'micro' in microstate here doesn't refer to a system that is small in scale, but refers to the entire drink or model at the deterministic level of individual molecules or numbers in matrices.)[16] The microstate a coffee gravitates to when we stir it, or a chatbot when we query it, depends on a few properties common to all stochastic systems.

Energy is one such property. For a liquid, it's a measure of internal movement, that is, how much its molecules are deviating from a condition of complete rest. Energy for an LLM can be thought of as how far the model deviates from real-world data. In nature, the arrangement of molecules will naturally tend toward the lowest possible energy, which is why your coffee cools off.[17] For a large language model to be accurate, however, requires that engineers train it into a state with as low an energy as possible. While training large language models has a somewhat more structured series of steps than pouring espresso into a cup, in the end an LLM also reaches a point where the energy is at a minimum as measured by how close its output fits the training data.

The disorder within a thermodynamic system, meanwhile, is represented by its entropy. The mathematical definition of entropy for a liquid counts how many possible microstates look like the current microstate. A latte with all the milk molecules bunched together in one half and all the coffee molecules bunched together in the other would be unlikely indeed--not because there aren't multiple microstates that could look like this, but because the number pales by comparison to the myriad of possible configurations of molecules in which the milk and coffee molecules are intermixed. (That's why 'latte art' doesn't spontaneously form in your macchiato.) Entropy for a large language model, meanwhile, corresponds to the uncertainty and variability of its possible outputs, which influences the diversity of generated content.

While energy and entropy are properties of the entire system, they can be observed indirectly by sampling. Measuring the momentum of the septillion molecules in a coffee cup is beyond the reach of contemporary scientific instruments, but a few drops sprinkled on your tongue can tell you if the entire cup is too hot to drink. Likewise, if a chatbot gives a different answer every time you ask it for the best tennis player of all time, that would be a condition of high energy, while a chatbot with zero energy might always respond 'Serena Williams'.

Natural systems naturally aim for a balance of minimising energy and maximising entropy.[18] In practice, AI engineers also have to strike a balance between outputs that are creative yet trustworthy. It's great to have a chatbot that responds 'Paris' every time it's asked for the capital of France; but if 'princess' is the only response you get from prompting a chatbot to complete the sentence, 'Once upon a time there was a ', the energy is too low to qualify as 'generative' AI.

---

16 It is possible for a deterministic system to exhibit stochastic behaviour. Lennart Carleson, 'Stochastic Behaviour of Deterministic Systems,' IUI Working Paper, No. 233, The Research Institute of Industrial Economics (IUI), Stockholm, 1989. https://www.econstor.eu/bitstream/10419/95223/1/wp233.pdf, accessed 16 November 2024.

17 Technically a cup of coffee could be an open system if you apply heat or cold externally, but left to its own devices any closed physical system will seek its lowest energy state.

18 This law can be generalised as the concept of Gibbs Free Energy. 'Gibbs Free Energy,' Wikipedia, https://en.wikipedia.org/wiki/Gibbs free energy, accessed 16 November 2024.

Bringing a drink or a model to equilibrium requires an interplay between energy and entropy. If all the milk is at the top of your drink and you don't have a spoon, heating up your coffee will increase the kinetic energy of its molecules, causing them to scatter more evenly, eventually mixing the ingredients thoroughly and maximising your drink's entropy. Likewise, when a model always returns the same results, it has been 'overfitted' to its training data and needs more entropy. During training, engineers can respond by injecting more energy into the model with techniques like dropout and weight decay that add randomness to the system. This increases the number of available microstates and thus diversifies the possible answers to a prompt. Add a smidgen of energy and your chatbot might offer 'Once upon a time there was a queen'; add a lot and you might get 'Once upon a time there was a donut'.

Even after a model has been trained, it's possible for users to coerce it into more diverse responses by adjusting the model's 'temperature' setting. For a coffee, temperature is the average kinetic energy of its molecules. For a model, it can be thought of as the 'creativity' or randomness in LLM outputs. Far from a vague analogy, the concept of temperature has an exact parallel in the mathematics of liquids and chatbots. Near equilibrium, the energies of molecules in a liquid or gas are distributed according to an equation first formulated by physicist Ludwig Boltzmann:

$$\mathrm{P}(Ei) = \frac{e^{-\frac{Ei}{kT}}}{\sum j\, e^{-\frac{Ej}{kT}}}$$

[Note to editor: this and the following equation are in LaTex format]

The exponential function in the equation, symbolised by Euler's number e, is a mathematician's way of exaggerating the difference in a spread of numbers. e to the power of 0 is 1; e to the power of 2 is 2.7; e to the power of 3 is 20; e to the power of 4 is 55. The numbers 0 through 4 may only be four units apart, but when cast as exponents the 3 and 4 get much further apart than the 1 and 2. Because the exponent in the formula is negative, that imposes a dramatically lower probability for high-energy states than low-energy ones. For coffee, this bias means a cup left alone is likely to cool off. For large language models, this bias favours only the word predictions with the highest correlations, and can even prevent a chatbot from ever returning anything but the best fit. That can be desirable when you're asking for a factual answer like the capital of France, but this 'degenerate' condition can never produce creative or unexpected results.

That's where temperature comes in, symbolised by the T in the denominator of the exponent. The higher this value is for a liquid, the more the distribution of molecular energies will spread out. At absolute zero degrees, all the molecules would be still, corresponding to an extremely tall and narrow curve near zero. At high temperatures, the curve widens, allowing molecules to explore higher energies far from the zero state. In that case, some of the molecules might still be slow, but others will be moving fast enough to burn your tongue.

Not by coincidence, temperature is the parameter that controls the sharpness of a function called at the end of an AI inference that steers the likely outputs to an AI prompt. As in the

formula for the distribution of molecular energies, this AI 'Softmax' function includes a temperature parameter in the denominator of the exponent to adjust the distribution's spread:[19]

$$P(xi) = \frac{e^{\frac{zi}{\tau}}}{\sum j\, e^{\frac{zj}{\tau}}}$$

AI tools with knobs you can turn to control the temperature--common in image generators and open-weight models--allow users to constrain outputs to only the most likely outcomes or to expand the aperture to 'roll the dice' for more fanciful results.[20] This is in marked contrast to the oracular interface of ChatGPT and the like, which as noted above make it seem like AI utterances come from some omnipotent oracle.

The Importance of AI's Thermodynamic Pedigree

To see how AI companies have buried generative AI's thermodynamic pedigree, look no further than the 2024 Nobel prize for physics. Angry physicists booed when the 'Godfather of AI' Geoffrey Hinton shared the Nobel prize in Physics, while Silicon Valley cheered confirmation that AI has become the driver of all innovation. Both sides got the lesson exactly wrong.

Observers who didn't bother to read the official Nobel award citation probably assume Hinton, a computer scientist, was honoured because his AI contributions made possible so many wonderful achievements in physics. Computing has certainly opened entirely new directions of physics research, but that was true long before generative AI. Contrary to the narrative pushed by AI boosters, the fact that the 2024 Physics Nobel was shared by a computer scientist and a physicist is less proof of computing's influence on physics than of physics' influence on computing.

The chain of discoveries recognized in this Nobel award starts with Hinton's fellow recipient, physicist John Hopfield, whose research describing how magnets emerge from lattices of electrons led to pattern-recognizing networks called Boltzmann machines. These statistical models are named after the same Boltzmann who founded thermodynamics, laying the groundwork for the probability distribution common to coffee cups and large language models described above. Hinton name-checked Boltzmann because finding a pattern in the noise turned out to be equivalent to finding the lowest-energy state of a system. (As Salvaggio reminds us, image generators like Midjourney render the pope wearing Prada or a capybara on Mars by looking for those images in noise.)[21] Hinton nabbed the Nobel because a technique he added, backpropagation, dramatically improved the pattern recognition of these networks. *

*** Correction: Hinton developed his Boltzmann machines independently of Hopfield; see https://new.nsf.gov/news/nsf-congratulates-laureates-2024-nobel-prize-physics. The author is grateful to Timothy Dasey for pointing this out.**

[19] For more on how the temperature changes chatbot output, see ML Tech Lead, 'What is this Temperature for a Large Language Model?', YouTube, 17 May 2024, https://www.youtube.com/watch?v=FMPzS2gQrNI, accessed 16 November 2024.

[20] The parallels between the Boltzmann distribution and softmax function don't stop with temperature. Both equations are 'normalized' so that the sum of all microstates will be 100%--whether those are possible energies of a molecule or words in the response--ensuring that every probability is correctly accounted for.

[21] Eryk Salvaggio, 'Flowers Blooming Backwards into Noise,' YouTube, 1 June 2023, https://www.youtube.com/watch?v=zNA7sPm-zlQ, accessed 16 November 2024.

Why should we care about whether generative AI came from physics or computer science? Because seeing large language models as statistical engines rather than artificial brains reminds us of their fallibility. And thermodynamic concepts like energy and temperature can help explain why they seem at turns big-brained or pea-brained depending on the context.

One paradox that concepts like temperature can help explain is that models scoring highest on hard problems requiring mental creativity and lateral thinking may ironically score the lowest on easy problems requiring short, fact-seeking answers. As a case in point, GPT-4o, a model OpenAI touted for its superior reasoning ability, scored less than 40% on OpenAI's own SimpleQA benchmark.[22] We can explain this using the homomorphism between large language models and thermodynamic systems. The higher you raise the temperature, the more solution states a model will sample, but that also means the more likely it is to settle in one that is incorrect. To put it crudely, the 'smarter' they are, the more they can be wrong.

The thermodynamic view of chatbots reminds us that they are stochastic systems whose effects vary dramatically based on hidden properties of the system like entropy and temperature. The manifest commonalities between chatbots and mechanistic systems break the often implicit parallel between electronic neural networks and biological brains. And the thermodynamic revelation that microscopic states may obey universal laws even if they cannot be directly apprehended debunks the common perception of large language models as inscrutable black boxes that cannot be explained or tuned.

AI as a Compression Algorithm

There are of course ways in which a large language model is not at all like a cup of coffee. Training an LLM involves numerous intertwined processes that can't be captured simply by the activity of molecules bouncing around a mug. Additionally, a cup of hot liquid achieves equilibrium on its own thanks to deterministic physical laws, while a large language model must be trained by engineers on data created by free-spirited humans; the first embodies natural laws, the second emulates them. This makes generative AI's outcomes less predictable and more varied than those in thermodynamic systems near equilibrium.

As useful as concepts like energy and temperature are for understanding AI models, it's unclear at first how they would help assess AI's threat of supplanting the archive. As stochastic engines, large language models can accommodate all possible microstates, but archives don't have the shelf or server space for that. Jorge Luis Borges' famous combinatorial library aside, archivists have to curate their selections of dusty books or retro MP3s because they can't accommodate everything. Even the World Wide Web, for all its strange nooks devoted to Flat Earth theories and Rent-a-Chicken enterprises, is but a tiny subset of every possible web page.

Since large language models can essentially accommodate every linguistic utterance real or imagined, researchers looking to expand their range of scholarly inquiry may choose to consult AI instead of a librarian (or Google). Science fiction tells us where this line of thought is going, in examples like the synthetic Librarian of Neal Stephenson's novel *Snow Crash* (1992).

---

[22] See OpenAI, 'Introducing SimpleQA,' https://openai.com/index/introducing-simpleqa, accessed 15 November 2024, and Victor Tangermann, 'OpenAI Research Finds That Even Its Best Models Give Wrong Answers a Wild Proportion of the Time,' Futurism, 2 November 2024 https://futurism.com/the-byte/openai-research-best-models-wrong-answers, accessed 15 November 2024.

But again we don't need to resort to an animate metaphor to imagine a mechanism capable of digesting a library full of documents into a concise summary. Another precedent for AI models captures its relationship to a human-made archive in a way that is simple yet rigorous. And that precedent is as ubiquitous in our digital workplaces as hot beverages are on our desks.

Today's digital workflows would slow to a crawl without algorithms like JPEG and AAC to share images and stream songs, or the ZIP file format we use to speed attachments to and fro. The Internet's architects have entrusted one particular utility, gzip, with the responsibility for compressing web traffic, data archives, software downloads, container images, database backups, and file transfers. So it may come as a surprise that generative AI models can be viewed as an advanced equivalent of the humdrum compression algorithms that we use every day.

For the lay user, PNG and JPEG may be the most familiar compression formats. A raw digital photo is essentially a two-dimensional array of numbers, each corresponding to the colour of a pixel. The resolution of your smartphone camera is on the order of 100 megapixels, so reducing redundancy is essential to making room for all those vacation and pet snapshots on your phone. You can think of PNG compression as counting the number of contiguous pixels with the same colour. To represent a row of 5 white pixels in a logo background, a raw photo might include the block 'white white white white white '; PNG replaces that 30-character string with the 8-character string 'white 5x'. JPEG, on the other hand, is meant more for photos than graphics, so it aims to replace transitions from one colour to another rather than uniform colour blocks. In a sunset snapshot, JPEG might represent the gradual fade from a blue sky to an orange horizon with a uniform gradient between those two colours.

Regardless of the compression format, what's important is that such interpolations deliberately discard any deviation from uniformity. When you load a JPEG onto your phone, the photo uncompresses back into an array of pixels. But this new, streamlined version may omit some tiny blips in the original gradient that were in fact birds in the far distance. This interpolation can be counterproductive when image compression blurs over subtle fractures in medical images or licence plates in crime scene photographs. But it is a known cost of compressing raw data into a digested deliverable.

Science fiction author Ted Chiang describes the parallel in his essay 'ChatGPT Is a Blurry JPEG of the Web':

> Imagine that you're about to lose your access to the Internet forever. In preparation, you plan to create a compressed copy of all the text on the Web, so that you can store it on a private server. Unfortunately, your private server has only one per cent of the space needed; you can't use a lossless compression algorithm if you want everything to fit. Instead, you write a lossy algorithm that identifies statistical regularities in the text and stores them in a specialized file format. Because you have virtually unlimited computational power to throw at this task, your algorithm can identify extraordinarily nuanced statistical regularities, and this allows you to achieve the desired compression ratio of a hundred to one.
>
> Now, losing your Internet access isn't quite so terrible; you've got all the information on the Web stored on your server. The only catch is that, because the text has been so highly compressed, you can't look for information by searching for an exact quote; you'll never get an exact match, because the words aren't what's being stored. To solve this problem,

> you create an interface that accepts queries in the form of questions and responds with answers that convey the gist of what you have on your server.
>
> What I've described sounds a lot like ChatGPT, or most any other large language model.[23]

AI in education enthusiast and Wharton professor Ethan Mollick puts it more simply:

> We now have the world's most advanced compression system for knowledge. Anyone can download, for free, a 235 GB file that can answer questions in many languages based on a vast swath of all human writing (even if it makes some errors, unsurprising as compression isn't perfect).[24]

These comparisons may sound like figures of speech, but taking the compression analogy literally can help us shed some of the hype around AI models to see them with greater clarity. Take GPT-3, an early version of ChatGPT for which OpenAI has disclosed more details than with subsequent models. Training the GPT-3 model distilled word correlations in 300 billion tokens of source data down to 500 matrices with a total of 150 million parameters. In principle, GPT-3 is able to reproduce information digested from the original data on demand. This compression is lossy, which explains why retrieved content can feel boring for some queries yet farcical for others.[25] Despite their occasional misfires, generative AI models are a remarkably compact and efficient representation of the original material. Their formidable 2000-to-1 compression ratio demonstrates the ability of large language models to distil and generalise vast textual information into a manageable utility.

As in the case of the coffee metaphor, an examination of the underlying mathematics shows deep parallels between compression algorithms and generative AI. Although they were not originally intended to generate novel results, modern compression algorithms identify implicit features in source text, code, or media, and then map them into a vector space and compute the similarities among them.[26] To reduce storage requirements while preserving core information, a common compression technique involves replacing groups of data points with

---

[23] Ted Chiang, 'ChatGPT Is a Blurry JPEG of the Web,' *The New Yorker*, 9 February 2023, https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web, accessed 16 November 2024.

[24] Ethan Mollick, 'An underutilized perspective...', LinkedIn, October 2024, https://www.linkedin.com/posts/emollick_an-underutilized-perspective-on-ai-for-non-technical-activity-7243309260212699136-lzXx, accessed 16 November 2024.

[25] Whether the results are boring or fanciful depends on the density of good data near the midpoint of the query. See Jon Ippolito, 'AI Made Me Basic,' Still Water Blog, 13 November 2024, https://blog.still-water.net/ai-made-me-basic, accessed 16 November 2024.

[26] 'Compression algorithms implicitly map strings into implicit feature space vectors, and compression-based similarity measures compute similarity within these feature spaces.' D. Sculley and C.E. Brodley, 'Compression and Machine Learning: a new Perspective on Feature Space Vectors,' IEEE Explore, 10 April 2006, https://ieeexplore.ieee.org/document/1607268, accessed 16 November 2024.

their midpoints.[27] This has a striking parallel in the weighted vector sums by which inferences derive outputs from large language models, which in a crude sense are regressions to the mean.[28]

The most convincing validation of this view of generative AI models is the fact that some compression algorithms can be pressed into service as language models, even when they were originally developed purely to reduce a file's footprint when saved on a hard drive or streamed over the Internet. Indeed, a text classifier that combined gzip with the neighbourhood averaging technique described above actually outperformed a 2018 language model developed by researchers at Google.[29] gzip's surprising success as a word predictor further underscores the value of viewing generative AI as a sophisticated compression tool. Perhaps it is no wonder that before OS 10.5 Leopard, the Apple command for compressing a folder was 'Create Archive'.[30]

Conclusions from Mechanistic AI Analogues

Seeing a large language model as a compression algorithm is useful when considering the model's ability to summarise and condense vast amounts of information. Indeed, since the 1990s, some researchers have argued that compression is equivalent to intelligence.[31] However, the analogy between large language models and compression algorithms falls short in capturing the LLM's capability to generate novel content. This limitation can be explained, however, by combining our understanding of LLMs as compression algorithms with our understanding of them as thermodynamic systems.

We can view compression algorithms as extreme cases of a more general class to which large language models belong. Because its goal is to reproduce as closely as possible the original Word document or Quicktime movie, gzip aims to match the entropy of the original file. In this sense, gzip is a large language model with the temperature set to zero. Nevertheless, the fact that entropy is a key parameter in this process allows us to see how large language models deviate from this restriction by enabling the user to adjust a model's temperature. While you wouldn't want to do this for most compression contexts, large language models can be viewed as a generalisation of compression that permits deviations from exact reproduction of the original data.

What makes these analogues more useful than the stochastic parrot metaphor is that thermodynamic and compression operations do not merely replicate what is in their training data, but they transform it through a process of normalisation. A dollop of milk poured into espresso produces a homogenised new hybrid with a different flavour and texture than black coffee or

---

[27] See the discussion of k-means clustering in Vincent Cohen-Addad and Alessandro Epasto, 'Differentially Private Clustering for Large-scale Datasets,' Google Research, 25 May 2023, https://research.google/blog/differentially-private-clustering-for-large-scale-datasets.

[28] In another parallel, large language models privilege information repeated frequently in its training data; likewise, gzip encodes data with variable-length codes, assigning shorter codes to more frequent elements.

[29] Zhiying Jiang et al., 'Low-Resource' Text Classification: A Parameter-Free Classification Method with Compressors,' Findings of the Association for Computational Linguistics: ACL 2023, pp 6810–28, https://aclanthology.org/2023.findings-acl.426, accessed 16 November 2024.

[30] 'List of Built-in MacOS Apps: Archive Utility,' Wikipedia, https://en.wikipedia.org/wiki/List of built-in macOS apps#Archive Utility, accessed 16 November 2024.

[31] See the discussion in Yuzhen Huang, 'Compression Represents Intelligence Linearly,' arXiv:2404.09937v1, 15 April 2024, https://arxiv.org/html/2404.09937v1, accessed 16 November 2024.

milk alone. That said, while they are more expressive than hot liquids or compression algorithms, AI models may limit the amount of deviation due to the way they produce results by 'averaging' derivations from their training data.[32]

Apart from disarming the temptation to ascribe lifelike qualities to chatbots, the mechanistic analogues we've reviewed each help in their own way to reveal two diametrical flaws in large language models that make them ill-suited as replacements for an archive. As argued above, thermodynamic systems aim toward equilibrium, which minimises anomalous microstates. Stir your coffee (or just wait long enough) and you'll end up with a liquid homogeneous in makeup and temperature; left to its own devices, your cappuccino isn't going to spontaneously divide into espresso on one side and milk on the other, much less create a stunning work of latte art. Likewise, compression algorithms like gzip privilege the presentation of items that occur frequently in the training data; this makes gzip a stochastic system, which tends to repress outliers in favour of the most common features in the source material.

On the other side of the coin, large language models can invent bogus records not present in the original archive. Of course, JPEG is typically enlisted to compress and then reconstruct a photo for which it has been fed a complete array of pixels. But users can easily prompt generative AI to return information for which there is no training data, like 'Explain the Treaty of Versailles as a Choose-Your-Own-Adventure story' or 'Show the pope wearing a puffer coat.' This is like feeding JPEG a photo with big patches of pixels missing, and like any good compression algorithm ChatGPT and DALL-E will do their best to reconstruct the data void. While these examples are fun, no researcher wants her librarian to make up books that don't exist, which makes generative AI a poor substitute for an archive.

The Reverse Archive

If generative AI tends to smooth over outliers while fabricating interpolated points, what future could that imply for scholars and creators? As mentioned in the introduction, AI tools may twist the archival arc into an unfamiliar trajectory. Comparisons to mechanistic analogues like thermodynamic and compression processes suggest that large language models are neither superintelligences nor Xerox machines, but technologies that replace diverse content with normalised hybrids. By comparison, an archive has less entropy and more energy than a large language model; most of us go to archives to discover ground truth, however messy. Compressing an archive removes the unique factuality that makes them useful to us.

But what if the ease and ubiquity of chatbots inspires the public to consult them rather than direct sources? The danger is that a large language model is not meant to retrieve knowledge from an archive but to replace it with a homogenised surrogate, with the long-term risk of reconstruction turning into replacement. If a lay user needs a photo of the Colosseum for her website, it's going to be faster and cheaper to prompt Stable Diffusion to create one than to pay for a stock photo or a professional photographer.

Human creators and cultural heritage professionals may react to this forecast with panic, though it's reminiscent of a panic those of us who lived through the dot-com era have felt before. Back then we worried that digitising and uploading Guernica or the Grateful Dead would mean

---

[32] See Jon Ippolito, 'Why Your AI Outputs Feel Average,' Still Water Blog, 2 September 2024, https://blog.still-water.net/why-your-ai-outputs-feel-average, accessed 16 November 2024.

people spent less time at the real canvas or concert, and this has basically come to pass. It's conceivable that the same fate could befall the same digital media that usurped analog media; perhaps people in the future will interact less with actual photos and songs, and more with surrogates conjured up via impromptu conversations with chatbots. The original web chopped up the seamless pages of printed books and magazines into separate items like text, styles, and images. Generative AI goes a step further by atomizing the web itself into individual words that are recombined probabilistically when the reader asks for new content. While these models clearly wouldn't exist without archives, they also shatter the authority of the archive's status as the standard bearer for truth.

A preview of this disturbing vision that you can explore right now is Websim[33], a site that is effectively a version of the Internet created on-the-fly by user prompts. The site presents you with a fake web browser; type a web domain you want to visit, like malaysia.travel or pumpkin.recipes, in its location bar and behind the scenes a large language model generates the HTML and images for a website that seems to be a probable match for that web address. Clever prompters have even managed to get Websim to produce playable 2- and 3-D games without ever touching the code or making an actual website. The result is a sort of mirror of the web, created not by HTML coders and web designers but by the spur-of-the-moment desires of its viewers.

While Salvaggio describes generative AI as an 'archive in reverse,' AI may engender an even more contorted twist in the trajectory of cultural production. While there's broad consensus that scaling up the amount of data fed into these models has produced remarkable results from fairly simple mathematical constructs, companies like open AI and Google are running out of data to scrape. To feed future expansion of these ravenous models, some engineers have proposed using the models themselves to generate 'synthetic data' to train future releases. Even if AI companies are wary of depending on this lab-grown data, they may not be able to avoid training on AI-generated content. After all, the current landing place for AI-generated PNGs of unicorns with rollerblades and AI rehashes of popular news stories is social media and websites. At that point, the products of previous models posted to the Internet will in turn become fodder for training the next models.

In this dystopian curatorial ouroboros, archives will no longer be just the starting or end point of a cultural artefact, but both origin and destination. In the extreme case, this paradigm shift could make coming into contact with actual human-made artefacts even less likely for the average person. For archival sources that we depend on for ground truth, from Wikipedia to Reddit to the Internet Archive, the danger is becoming overrun with AI-generated slop. For the AI models themselves, the danger is overfitting or even model collapse. A 'synthetic archive' would offer a perpetual source of training data, but it also threatens to create evermore inbred generations of output. This vicious circle could amplify the numerous ways models can misrepresent the truth, whether by reproducing bias in their training data or by reducing the diversity of archival records to averages.[34]

---

[33] Websim, https://websim.ai, accessed 16 November 2024.

[34] When Salvaggio prompted an image generator for an obsolete photograph format, he saw another example of how AI can surface stereotypical content in an archive: 'I never prompted anything aside from Stereoview [but found repeated] images of palm trees and certain styles of dresses that had nothing to do with this medium per se.

Conclusion

As a general rule, collecting institutions haven't stayed abreast of the implications of generative AI as much as fields like education, software development, or even the arts. Some intrepid experimentation has demonstrated how useful AI tools can be for the digital humanities and data science; researchers have demonstrated the value of generic chatbots to automate dreary tasks like adding metadata to photographs[35] or cleaning up bioinformatic data in a spreadsheet.[36]

While these experiments are essential, we must not let them distract us from the looming threat that generative AI poses to the content and relevance of formal and informal archives. Keeping alert to the potential risks and benefits of generative AI means being clear-eyed about the fundamental mechanisms by which they work--and steering clear of animistic metaphors that obscure those mechanisms.[37]

Acknowledgement of AI use

The author used GPT-4o extensively for research, but not for sources, conclusions, or editing prose.

Bibliography

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜'. *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (March 2021): 610–23. https://doi.org/10.1145/3442188.3445922.

Bergman, Michael K. 'The Deep Web: Surfacing Hidden Value.' BrightPlanet (July 2000). https://resources.mpi-

But when I went into the training data for these image models and explored open repositories for where these images may have been sourced from, I discovered that there was a vast collection of stereoview images from the Library of Congress that referred to stereoview images — and specifically, as a tool for circulating images of the US occupation of the Philippines. The US circulated these images deliberately to tell a story glorifying that project. As expected for a cultural institution, the Library of Congress shared these images online, contextualizing them in an exhibition that showed how this storytelling was crafted and the goals it served. Nonetheless, when encountered simply as data, these images became strongly associated with the imagery of stereoview images, even to the extent that the very hallmark of the media format — two photos, side by side — would sometimes go away. At the same time, pictures of colonization would remain, as if the word stereoview was more strongly correlated to colonization than to a media format.' 'Infrastructures of Memory,' Cybernetic Forests, Oct 19, 2024, https://cyberneticforests.substack.com/p/infrastructures-of-memory, accessed 16 November 2024.

[35] Sara and Ben Brumfield, '10 Ways AI Will Change Archives,' YouTube, 27 October 2023, https://www.youtube.com/watch?v=Fmgbk1x6RSY, accessed 16 November 2024.

[36] Andy Stapleton, 'A Literal AI Game Changer for Research & Academia,' YouTube, 12 July 2023, https://www.youtube.com/watch?v=yklFHtlK4sQ, accessed 16 November 2024.

[37] It's important to be both open-minded about potential solutions as well as recognize their limitations. In particular, solutions that favour voluntary interventions by humans may prevail over attempts to legislate blanket compliance that leave too many loopholes. For example, mandating watermarks for AI-generated content is unenforceable, while adding digital signatures to genuine archival material is a tried-and-true way to ensure evidence is trustworthy or creations are human-made. Likewise, outlawing deep fakes will be less effective than supporting local journalism that can debunk fake news. More generally, standing up for diversity amidst the homogenising tendencies of generative AI will be a priority going forward.

inf.mpg.de/d5/teaching/ws01_02/proseminarliteratur/deepwebwhitepaper.pdf, accessed 16 November 2024.

Blais, Joline and Jon Ippolito. *At the Edge of Art*. London: Thames & Hudson, 2006.

Brumfield, Sara and Ben Brumfield. '10 Ways AI Will Change Archives'. YouTube, 27 October 2023. https://www.youtube.com/watch?v=Fmgbk1x6RSY, accessed 16 November 2024.

Carleson, Lennart. 'Stochastic Behaviour of Deterministic Systems'. IUI Working Paper, No. 233, The Research Institute of Industrial Economics (IUI), Stockholm, 1989. https://www.econstor.eu/bitstream/10419/95223/1/wp233.pdf, accessed 16 November 2024.

Chiang, Ted. 'ChatGPT Is a Blurry JPEG of the Web'. *The New Yorker* (9 February 2023). https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web, accessed 16 November 2024.

Cohen-Addad, Vincent and Alessandro Epasto. 'Differentially Private Clustering for Large-scale Datasets.' Google Research (25 May 2023). https://research.google/blog/differentially-private-clustering-for-large-scale-datasets.

Gupta, Anuj, Yasser Atef, Anna Mills, and Maha Bali. 'Assistant, Parrot, or Colonizing Loudspeaker? ChatGPT Metaphors for Developing Critical AI Literacies'. arXiv:2401.08711, 15 Jan 2024. https://arxiv.org/abs/2401.08711, accessed 15 November 2024.

Huang, Yuzhen. 'Compression Represents Intelligence Linearly'. arXiv:2404.09937v1 (15 April 2024). https://arxiv.org/html/2404.09937v1, accessed 16 November 2024.

Ippolito, Jon. 'AI Made Me Basic.' Still Water Blog, 13 November 2024. https://blog.still-water.net/ai-made-me-basic, accessed 16 November 2024.

——. 'Why you Should Generate AI Images in Your Classroom'. Still Water blog, 8 January 2024. https://blog.still-water.net/why-you-should-generate-ai-images-in-your-classroom, accessed 16 November 2024.

——. 'Why Your AI Outputs Feel Average'. Still Water Blog. 2 September 2024. https://blog.still-water.net/why-your-ai-outputs-feel-average, accessed 16 November 2024.

Jiang, Zhiying, et al. '"Low-Resource" Text Classification: A Parameter-Free Classification Method with Compressors'. *Findings of the Association for Computational Linguistics: ACL* (2023): 6810–28. https://aclanthology.org/2023.findings-acl.426, accessed 16 November 2024.

Manovich, Lev. 'Database as Symbolic Form'. *Millennium Film Journal* No. 34 (Fall 1999). http://www.mfj-online.org/journalPages/MFJ34/Manovich_Database_FrameSet.html, accessed 12 November 2024.

Mollick, Ethan. 'An underutilized perspective...'. LinkedIn, October 2024. https://www.linkedin.com/posts/emollick-an-underutilized-perspective-on-ai-for-non-technical-activity-7243309260212699136-lzXx, accessed 16 November 2024.

Salvaggio, Eryk. 'Infrastructures of Memory,' *Cybernetic Forests*, Oct 19, 2024. https://cyberneticforests.substack.com/p/infrastructures-of-memory, accessed 16 November.

——. 'Flowers Blooming Backwards into Noise'. YouTube, 1 June 2023, https://www.youtube.com/watch?v=zNA7sPm-zlQ, accessed 16 November 2024

Sculley, D. and C. E. Brodley. 'Compression and Machine Learning: a new Perspective on Feature Space Vectors.' IEEE Explore, 10 April 2006. https://ieeexplore.ieee.org/document/1607268, accessed 16 November 2024.

Stapleton, Andy. 'A Literal AI Game Changer for Research & Academia'. YouTube, 12 July 2023. https://www.youtube.com/watch?v=yklFHtlK4sQ, accessed 16 November 2024.

Tangermann, Victor. 'OpenAI Research Finds That Even Its Best Models Give Wrong Answers a Wild Proportion of the Time'. *Futurism*, 2 November 2024. https://futurism.com/the-byte/openai-research-best-models-wrong-answers, accessed 15 November 2024.