Open Library of Humanities Journal

Erslev, M S 2026 The Ground Shook! Towards a Practice-based Poetics of the Transformer. *Open Library of Humanities*, 12(2): pp. 1–31. DOI: https://doi.org/10.16995/olh.28661

Open Library of Humanities

The Ground Shook! Towards a Practice-based Poetics of the Transformer

Malthe Stavning Erslev, Assistant Professor, School of Communication and Culture – Department of Comparative Literature and Rhetoric, Aarhus University, Denmark, stavning@cc.au.dk

This article argues that literary approaches to language models should insist on practice-based, media-specific engagement with model components. Specifically, the article focuses on the attention mechanism of the transformer neural network architecture and argues that this mechanism should be approached and rendered in practice-based ways that are native to (digital) literary culture. Theory can gesture toward literary and poetic potentials of language models yet still fail to escape conjecture without hands-on intervention. We need to think smaller, dig deeper, and stop letting corporate interfaces define what these systems are allowed to be. To make this argument, the article draws on theoretical perspectives on the relation between language models and literature, technical descriptions of the functionality of the attention mechanism, and (not least) a practice-based inquiry taking the form of *BERTIFIER*, a literary interface to the BERT language model that acts as both method and example in search of a poetics of the transformer architecture.

Open Library of Humanities is a peer-reviewed open access journal published by the Open Library of Humanities. © 2026 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/. Articles published by the Open Library of Humanities may also contain other CC-licensed material, or material used or reproduced under an exception or limitation to copyright. Such material is not subject to the article's license. It is either marked with its own license, or marked that it has been reproduced with permission or under fair-use.
OPEN ACCESS

Introduction

ChatGPT is boring. In its currently normalized form, so-called generative artificial intelligence may be the most boring thing that has happened to literature – to anything – in a long time. Problematic, too, in so many ways. These technologies are boring not because of any inherent essence in their technological build that is imbued with boringness a priori, but because of the incessant optimization towards sycophancy and corporate bullshit that may be the most defining characteristic of top-tier, as it were, off-the-shelf chatbots. They are problematic for many reasons, the direst of which may be the destructive logic of scaling – indeed hyperscaling – that guide the private, for-profit companies that develop and deploy them. In the interest of focus, I here restrain myself to focusing on the boringness of the situation, notwithstanding its problematic aspects.

As Jill Walker Rettberg and Hermann Wigers (2025) have documented, stories generated with OpenAI's gpt-4o-mini model overwhelmingly display stability over novelty (or even much meaningful variation). They argue that generative artificial intelligence exhibits a form of narrative bias, which favors specific and limited forms of narrative structure and converges on a smaller set of basic stories that are endlessly repeated with slightly varying vocabulary. For instance, the figure of the lighthouse keeper Elias Thorne has been observed to be stable stock in generated text, popping up with peculiar persistence across contexts, which is likely a result of post-training techniques that optimize chatbots for corporate automation (Hamilton & Mimno, 2026). In other words, popular chatbots tend to display a kind of 'formal stuckness', a rigid repetition of an idealized formalism (Heuser, 2025).

My point is not that writing with these technologies *cannot* be exciting and potent. On the contrary, there is most definitely literature being produced with generative artificial intelligence that is far from boring. For example, take recent work published in Ensemble Park (n.d.), Re·mediate magazine (n.d.), The AI Literary Review (n.d.), and Michigan Quarterly Review (2025). These works build on a longer strand of critical literary work engaging with machine learning techniques, including work by Allison Parrish (n.d.), David 'Jhave' Johnston (2019), and others (*cf.* Bertram & Montfort, 2024). The figure of Elias Thorne (mentioned above) has even become the subject of a web-based piece of digital literature, taking the broad prevalence of the name as a starting point for a literary exploration of identity (Elias Thorne, n.d.).

Likewise, theorists have many intriguing thoughts – some of which I will survey later – about what kinds of literary implications these technologies *could* have (Hayles, 2025; Weatherby, 2025; see also Bajohr, 2025). In practice, however, what the use of these technologies so often results in is a kind of textual output that is as bland as the circumstances of its production are numbingly opaque.

The issue, I will argue, is that theoretical approaches often forego contentious practice and tend to rely on descriptions of (or surface interaction with) off-the-shelf systems as the object of study. In contrast, most of the genuinely interesting literary work involving these technologies work *against* the grain of off-the-shelf products to escape the boringness that haunts them (*cf.* Erslev, 2023). To bridge theory with practice we should insist that writing with computers is more than just producing humanesque output and instead grapple with the contingencies of process.

In this article, I argue that the question of generative artificial intelligence and its relation to literature should be approached from a practice-based position. Digital literature is uniquely suited to tackle critical issues of how these technologies exist within digital culture, how we think about them, and how they propagate in our communities. However, we can only take on those issues if we do so from a practice-based position, rather than a reactive one that primarily orients itself towards commercial chatbot products. That is to say, as long as we keep letting large, for-profit companies hold the power to define how these technologies work and how we can interface with them, they will continue to be boring and problematic, no matter how vividly we theorize about them.

My core argument is that we need to dig deeper and think smaller in order to locate actually exciting literary potentials in generative artificial intelligence that bridge theory and practice. In making this argument, I take a practice-based position myself. A lot of the work that has led to the arguments presented here has unfolded through my work on *BERTIFIER*, which is an alternative, literary interface to transformer-based large language models (Erslev, n.d.). *BERTIFIER* is both a work of digital poetry and a form of situated inquiry into the transformer architecture, focusing on the attention mechanism. The methodological point of departure is that there are modes of knowing that emerge in practice and that are difficult if not impossible to cultivate without it (*cf.* Rettberg & Saum-Pascual, 2020). This, in turn, is also an important part of my argument. Using my own experimentation with *BERTIFIER* as both methodology and example, I will argue that it *is* possible to zoom in on individual components of language models via practice, and that there is a poetics to such a practice.

In short, this article, and *BERTIFIER*, are attempts to be prosaic in the search for the poetic. The article moves as follows. First, I discuss the conceptual problems we face when we discuss generative artificial intelligence as an object of study and argue that these problems are largely due to a divergence in focus, since the technology looks markedly different depending on what part of it you focus on. Second, I zoom in on a single component of the transformer architecture, namely the *attention mechanism* and argue that any proper understanding of the attention mechanism – at least from a

techno-cultural perspective – must render it via practice. Third, I introduce *BERTIFIER* as a concrete experiment that takes the attention mechanism as its poetic driver to both make the operation of the attention mechanism legible in practice and, in the process, produce poetry.

Architectures and outputs

As is the case with most new things, generative artificial intelligence is subject to a nontrivial divergence in terms of what people mean when they use the term. In general, we tend to shift (most of the time without noticing) between two different ways of understanding the technology. One orients itself towards our experience of engaging with the output, the other towards the specifications of the models themselves. Each position implies the relevance of the other, but they do differ in terms of what they focus on. The former position evaluates the status of large language models according to a consideration of the underlying architecture and its reliance on probabilistic sampling. The latter evaluates the status of large language models according to how the output is experienced. Although I will argue that we should strive to integrate these two positions, their distinction is useful in order to more clearly configure the conceptual landscape within which my inquiry exists. To this end, consider a provisional distinction between an architecture-centric and an output-centric position in relation to generative artificial intelligence.

The architecture-centric position holds that understanding large language models is a matter of revealing and discussing structures within the models themselves. This is related to a consideration of the individual components of models, their optimization paradigms, and the way they are implemented. This view stipulates that output alone cannot sufficiently give insight into large language models and, as such, should not be trusted as an epistemic vehicle. As an example, the characterization of large language models as *bullshit machines* belongs to this position (Hicks, Humphries & Slater, 2024). The notion of bullshit machines holds that as long as large language models work via probabilistic next-token prediction, they are inherently unrelated to matters of truth or truthfulness, no matter how verifiable their output is. Interestingly, the obverse understanding of generative artificial intelligence known as *mechanistic interpretability* (a branch of *explainable artificial intelligence*) is still architecture-centric (*cf.* Templeton et al., 2024; *cf.* Kamath et al., 2025). Mechanistic interpretability does not characterize models as bullshit machines but instead as structures that encode something akin to human-understandable concepts, which can be mapped by observing patterns in the activations within the model as it generates output. By seeking to represent the model's

internal representations, this approach takes the model's structure as it generates output, rather than the output itself, as the epistemic driver.

The output-centric position holds that understanding large language models is a matter of evaluating the quality of outputs. At its most extreme, this view holds that it is simply unimportant how a large language model is constructed or how its components work so long as it produces desirable outputs. In this context, *desirable* can also mean friction-filled or contentious; the important thing is the experience of the person interacting with the model, not the model itself. This is to say that the output-centric view considers causal explanations of model behavior insufficient to the understanding of large language models in their various contexts. For example, studies that evaluate how generative artificial intelligence affects the workforce rarely focus on intra-model operations and instead seek to understand the effects of the models on workers (*cf*. Lee et al., 2025). The practice of benchmarking, where models are evaluated according to how well their outputs solve a given task, is also output-centric. Often, benchmark scores are one of the primary ways in which differences between models from various providers are discussed and, as a consequence, benchmark scores have become advertisements in-themselves. This is the case for both quantifiable benchmarks (*cf*. Raji et al., 2021) and so-called vibes-based benchmarks, where users vote on which model output they prefer, given two options for each prompt (Zheng et al., 2023).

The status of generative artificial intelligence vis-à-vis digital literature has also been subject to debate, and the contours of an architecture-centric versus an output-centric position are discernible. Whereas John Cayley (2023) takes a somewhat architecture-centric position in arguing that the configuration of these models themselves makes their use largely unjustifiable within digital literature, Scott Rettberg (2023) takes an output-centric position and argues that the field needs to engage with them on the level of output.

These two positions, one architecture-centric and one output-centric, are of course not mutually exclusive; rather, they direct their focus towards different aspects of a complex thing. It is perfectly reasonable to be interested in intra-model operations in one situation and, in another, to focus on the impact of model output on a given cultural context. Yet the complexity of the situation can result in significant confusion when discussing generative artificial intelligence across contexts. As I will argue, it is necessary to devise modes of working that combine the architecture-centric and the output-centric stances in order to properly study the materiality of transformer models. This involves revisiting the notion of media-specificity and considering its status in the 2020s, relative to artificial neural networks.

Media-specificity and neural networks

As Katherine Hayles argues, generative artificial intelligence 'pose[s] especially urgent challenges to literary studies, theory, and criticism, fields focused on the interpretation of verbal texts' (2025: 2). In these fields, media-specificity refers to the view that any rigorous interpretation of output must involve an understanding of the underlying technique (Hayles, 2004). This is why 'the question of whether these systems optimize well, whether they perform to benchmarks, no longer seems satisfactory. We have to know what these systems *mean*' (Weatherby, 2025: 7, original emphasis).

Media-specific interpretation seeks to illuminate how works of digital literature often 'do not limit themselves to acting within the realm of thematic representation or conceptual meaning-making alone' but instead center on a 'performative entanglement among actants (reader, player, systems, game, languages)' which 'closes the gap between the world and its representing object' (Saum-Pascual, 2025: 408). Such performative entanglements render digital culture in legible ways without taking away the complexity of the situation. Instead of seeking to find a single measure that defines a specific issue related to digital culture, these performative entanglements explore the particularities of techno-cultural encounters via practice. In other words, media-specific approaches to generative artificial intelligence must involve both output-centric and architecture-centric aspects in order to be able to make sense of the performative entanglements at play in any meaningful way.

However, when it comes to generative artificial intelligence, establishing a clear connection between a model and an output is easier said than done. As Hannes Bajohr has argued, the algorithmic form of computer-generated literature – based on clear rules and instructions executed in a sequential manner – 'seems to deliberately reflect its underlying technology' thus inviting or at least enabling media-specificity (2022: 209). Literature generated with language models complicates this in at least two ways. First, the model itself is not readable in the same sense as an algorithm and second, a single pretrained model can result in an incredibly wide array of outputs. Second, the implementation of off-the-shelf chatbots adds several layers of opacity that make analysis even more difficult. These layers include black-boxing the specific architectural construction of the models used, adding hidden pre-user system prompts, engulfing the model in various guardrails, etc. When we receive an output from a standard chatbot, we have virtually no way of making sense of its media-specific situation – apart from scrutinizing it at the level of the user interface, which is surely worthwhile yet not sufficient for a thorough media-specific consideration.

It is for this reason that we end up with the split into an architecture-centric and an output-centric position, also within digital culture and literature: when we cannot establish a clear connection between output and model, we turn our attention to one of them. We either infer something from our interactions with the model at the output-level or deduce something from its underlying architecture.

To enable media-specific analysis in relation to language models, it will be helpful to untangle the overarching notion of *generative artificial intelligence* to ask, what material configuration are we dealing with? What are the components, and how can we render them in meaningful ways? In the following section, I zoom in on the transformer architecture, where the *attention mechanism* presents itself as an optimal site for media-specific inquiry. However, as I will argue, it is not sufficient to simply describe or even visualize attention. Instead, media-specificity warrants that the attention mechanism is put into play via interaction with specific models, thus integrating an architecture-centric stance with an output-centric approach.

Attending to attention (in theory)

Transformers and other deep neural networks are often thought to be black boxes that cannot be scrutinized in detail and therefore are treated only vaguely by noting their general reliance on probabilistic sampling and their lack of algorithmic traceability, without even attempting to go into more detail. The consequence of this generality is that we end up basing our discussions of these systems on conjectures that exist somewhere between hyped-up marketing speech and half-baked rumors. I will not go over the difficulties associated with interpreting transformers (or any deep artificial neural networks) in detail, since they are well-described elsewhere (cf. Burrell, 2016; Dobson, 2023; Bajohr, 2022). Instead, I will focus on the aspects of these models that we actually *can* study and work with. We may not be able to account for every aspect of the processes that take part in transforming an input into output, but, in the interest of media-specificity, we can at least try to account for *some* aspects of *some* of those processes.

As James Dobson notes in his survey of modes of interpretation of transformer models, '[t]he two major architectural features to target for critical readings of Transformer models would be the hidden layers and the attention heads' (2023: 442). In the time following Dobson's survey, there have emerged numerous breakthroughs in the field of mechanistic interpretability, which seeks to map structures of so-called *features* and *circuits* within a large language model, as I also mentioned above (Templeton et al., 2024; Kamath et al., 2025). Mechanistic interpretability adds new sites for study of transformers but does not do away with the centrality of attention and the hidden layers. In addition, mechanistic interpretability adds new levels of complexity, since

it works by applying other machine learning techniques to the mix in order to render recognizable features and circuits, meaning that there also emerges a need to study the methods of mechanistic interpretability themselves. All this notwithstanding, I am here zooming in on attention, following the lead of Dobson and many others who identify the attention mechanism as an ideal site of study of the transformer architecture.

Indeed, both Hayles (2025) and Leif Weatherby (2025) argue (in different ways) that it is the way context is computed via the attention mechanism that is the defining aspect of large language models. To Hayles, the attention mechanism 'provides both focus and contextual analysis' within the model (2025: 142). To Weatherby (following the theoretical work of Roman Jakobson), '[t]he attention mechanism is an entry point of computational language, not into poetry as art, but into the poetic function of language as such' (2025: 155). In both cases, attention is what establishes a nontrivial relation between large language models and literature.

Let us, then, consider the attention mechanism as the potential crux of understanding transformer models. To the end of deepening this inquiry, let us take a closer look at how the attention mechanism works in theory. Such theoretical consideration is part of, but is not in-itself, a media-specific approach.

The attention mechanism

In language modeling lingo, the attention mechanism is often explained in terms of how tokens *attend to* one another. However, this formulation leaves a lot of space for misinterpretation of what that means. In transformers, attention does not only identify which tokens are related — it also *changes* the numerical representation of tokens based on those relationships. In deep artificial neural networks, tokens are represented by (often quite long) series of numbers, known as embeddings. The central operation of the neural network is to modify these embeddings based on model weights. In the transformer architecture, the attention mechanism runs many times in parallel in each block of the neural network to *both* identify various relations between tokens *and* modify each token's embedding, which is in turn transferred to subsequent blocks. In technical terms, this happens by first comparing a *queries* vector to a *keys* vector to determine relations between tokens and then applying changes to the input embedding using a *values* vector.

When we talk about how much one token attends to another, we are referring to the interaction of queries and keys, but the actual functional impact of the attention mechanism is the application of the values vector. There is no external variable that stores information about which tokens are related – even if such a value may easily be calculated, it is not used in the generation of output. Instead, the information regarding

how much tokens attend to one another is used to calculate how much the embedding of a given token should be changed. The attention mechanism is applied in each block of the model, across many different attention heads (that is, parallel operations of the attention mechanism in the same block, known as multi-head attention). Each attention head can manifest a distinct pattern, or projection, which has been defined during the training process. This means that each attention head may render different kinds of relationships between tokens. All this happens in concert with operations that do not involve attention at all, such as the *multi-layer perceptrons*, sometimes called the hidden layers. The modified token embedding, along with the original embedding and a cache of previous keys-values pairs, gets fed to subsequent blocks, where it continues to be modified in each block and, ultimately, decoded into an output.

No single attention head, and not even a single multi-head attention component, contains all relevant information about relations between tokens. Instead, it is the total amount of embedding modifications performed by attention heads and multi-layer perceptrons (along with other model components including the layernorm and softmax functions) across the whole of the model that comprise the transformation of input into output. In contemporary models, these core components of the standard (so to speak) transformer architecture are supplemented by guardrails and other alignment measures, and the weights of the model have been finetuned during post-training in order to optimize for specific kinds of output. It is the post-training that instantiates the chat-format of off-the-shelf chatbots, optimizes the models to have certain voices and stylistic tendencies (including corporate sycophancy), and makes them output so-called reasoning traces.

My point is that even though attention is a central component of the transformer architecture – perhaps its centerpiece – it is ill-advised to study it out of its material context. Looking at attention heat maps (calculated from queries-keys relations) and treating that as a stand-in for the operation of the transformer architecture as such will lose track of the contingencies of how the system works and, by extension, will offer little insight to those seeking to live and write with and against these technologies.

Theorizing attention

The attention mechanism enables transformer models to both preserve cohesion across long stretches of text and pick up on local linguistic structures (*cf.* Bajohr, 2024). This happens because attention patterns across all the tokens in the context window are computed myriad times across many attention heads each time the model generates a single new token, meaning that each new token is the result of virtually uncountable calculations of linguistic dependency across the entirety of the context window.

Hayles (2025) theorizes attention as the central part of what she considers the *nonconscious cognition* of language models. This form of cognition is not limited to artificial neural networks but is instead a non-anthropomorphic account of cognition that takes place in both humans and nonhumans, including technical systems. To Hayles (2017), the defining aspect of nonconscious cognition is that a nonrandom decision is made in a context where that decision makes sense, which applies to both the direction of growth in trees and the distribution of bits on the internet – and, noticeably, to the way language models generate output. According to Hayles (2025), the linguistic dependencies computed via attention should be understood as an important form of the materiality of language models, to such an extent that we can think of them as the crux of nonconscious cognition in these systems. In effect, close reading the output of language models is meaningful to Hayles, since such output displays a nonconsciously-cognitive mapping of tendencies across vast datasets, as computed by a specific neural network architecture.

Another intriguing way to theorize the attention mechanism is, as Weatherby (2025) proposes, to view it as an operationalization of Jakobson's famous *poetic function* of language, that is, the ways in which language points to itself, inquires into itself, not only in philosophy and poetry but also (most often) in quotidian dialogue. In Weatherby's words, '[b]y passing sequences through self-attention in this fashion, LLMs learn how the message relates to itself. It is only from this poetic basis – in which words similar in relevance are drawn together in meaning – that other functions, like reference, precipitate. The LLM grounds all other functions in the poetic' (2025: 162). The poetic function is often understood to be a key driver of linguistic creativity, where grammar and representational semantics give way to new juxtapositions to form a new linguistic idea. Per Weatherby, we 'don't have informational knowledge laid out in referential terms and only *then* access to the web of rhythm and phrase that makes up language. What gets domesticated as "natural," as meaning something, depends to a large extent on this poetic function – whether it operates by sound, text, or matrix' (2025: 160, original emphasis).

That all sounds good – in theory. If the attention mechanism operationalizes the poetic function of language and the poetic function, in turn, is a key part of linguistic creativity (if not, as Weatherby argues, the very foundation of any kind of language-activity as such), then the attention mechanism should enable transformer models to generate ample potent prose and poetry, where new connections emerge and the boundaries of literature are pushed to their limits. In theory, at least. Yet the output from these models that we see in practice is, more often than not, 'generic, centrist, normative, and banal' (Gillespie, 2024: 11). We cannot theorize our way out of this;

instead, we need to engage these models in ways that allow us to render, and sensitize, their operations in meaningful ways – and hopefully create interesting texts in the process.

We must insist on media-specificity, even when models can produce text that is largely indistinguishable from the human-written kind. As we have seen, there already exist some deeply considered theoretical and analytical frameworks for doing so. However, this is not a task that can be left to theory, analysis, and critique alone (although we also cannot do without them). No amount of thinking about a model-output relation will make it suddenly manifest, at least not in ways that move beyond theoretical conjecture.

To be fair, both Hayles and Weatherby seek to move beyond theory by considering visualizations of the attention mechanism, such as BertViz (Vig, 2019). These tools visualize attention patterns, often as lines of varying thickness connecting words in the sequence, representing how much one token attends to other tokens within a given attention head. To both Hayles and Weatherby, such visualizations effectively become their most materially grounded engagement with the attention mechanism.

Such tools are certainly useful in some specific cases (cf. Rogers, Kovaleva & Rumshisky, 2020). Yet they can deceive since they render attention patterns for a single attention head, not for the full model. Moreover, such visualizations suggest that attention is about tracing relationships, when in fact it is about changing the internal numerical representation of each token embedding. Other, more recent forms of probing tools – such as Neuronpedia (Lin, n.d.) – integrate quantitative measurements of attention in the material context of the model and could potentially be repurposed to serve literary interpretation, yet that is beyond the scope of my focus here. My point is that literature (and digital culture more broadly) cannot contend itself with visualizations developed for other domains.

Instead, what we need is to develop a practice-based poetics that enables media-specificity. If the standard, off-the-shelf chatbots dismantle the relation between model and output, the task for digital art and literature must be to hack it in order to restore it, and explore it.

To do so, let us move on to focus what practice can do for us here. In the following, I turn to my own practice-based experimentation with *BERTIFIER*, which seeks to rethink how literature interacts with transformer models by stripping away the bells and whistles of consumer-facing products and seeking to render a media-specific situation. In developing *BERTIFIER*, one guiding principle has been to harness digital poetry as a methodology to work out alternative model configurations and discover

modes of working with transformers within experimental writing. An outcome of this process has been a program that also allows readers to zoom in on the aesthetic operation of the attention mechanism, to try to make legible the operation of one of the most central components in any transformer.

BERTIFIER

Let us pause here to take stock of the argument so far. The attention mechanism cannot be properly studied out of context. It needs to be executed, to be situated within the full material context of the transformer architecture. However, consumer-facing, off-the-shelf chatbots do not provide us with the means required to zoom in on attention in practice, much less to rework our interactions with the models within which attention is implemented. So, to move forward, let us look for opportunities to work with attention in context, in ways that allow us to scrutinize the operation in more granularity.

Smaller, local, open models offer an opportunity for inquiry into the attention mechanism in a way that is contextualized via its implementation in a working model and at the same time allows more specialized inquiry into the operation of the mechanism itself. Many of these smaller models mimic the setup of commercial chatbots, but for our purposes, it is worthwhile to zoom in on a type of model that foregoes the chat interaction altogether yet enables us to scrutinize the attention mechanism with a literary orientation. I am here referring to the BERT model. In the case of literature, however, the interesting thing is not the model itself but the way it can be implemented as a form of electronic literature (*cf*. Rettberg, 2019) that renders attention in a way that emerges from, and orients itself towards, literary culture. I will first describe the BERT model and then my literary implementation of it in *BERTIFIER*.

BERT (Bidirectional Encoder Representations from Transformers) was released by Google in 2018 (Devlin et al., 2019). It was the first wide-ranging implementation of the transformer architecture, in terms of use, further development, and scholarship focusing on a single model (*cf*. Rogers, Kovaleva & Rumshisky, 2020). As one defining aspect of scholarship into BERT, researchers looked primarily on attention patterns within the model (Clark et al., 2019; Kovaleva et al., 2019). Immediately following its release, BERT quickly became virtually ubiquitous within the field of natural language processing and BERT-style models are still used for many purposes within that and related fields.

The decision to work with BERT was in part motivated by its somewhat privileged status as the first wide-ranging implementation of the transformer architecture. There is a media archaeological (*cf*. Huhtamo & Parikka, 2011) interest at play here that I will not tease out in this context. This is also the reason why *BERTIFIER* works with the original BERT model, rather than later, higher-performance, versions.

BERT is based on many of the same core ideas as the popular GPT-style generative models, but it also differs from them in a set of crucial ways. Contrary to GPT-style models, BERT is not optimized to generate new text strings based on a prompt. Instead, BERT is optimized for two distinct tasks, *next sentence prediction* and *masked language modeling*. The former, which I will not discuss further here, is the task of predicting whether or not two given sentences logically follow one another. The latter, which I will focus on, centers on (re-)inserting a fitting token in a given place in a sentence. In other words, when performing masked language modeling, BERT receives a sequence of text with a token missing (marked with a special [MASK] token) and is then tasked to predict which token could likely fit in the associated position in the sequence.

Unlike GPT-style models, which always predict the *next* token in a sequence and thus can only perform calculations on prior tokens, BERT is optimized to predict a masked token in any position. In doing so, it computes the prediction based on both the tokens that come before *and* after the masked token, hence the model is bidirectional (the 'B' in BERT). Interestingly, the attention visualization tools discussed above are almost exclusively based on BERT-style models, rather than GPT-style models, which is why these tools are able to show how much a given token attends to a token that comes later in the sentence.

Another difference between BERT and GPT-style models is that BERT is deterministic rather than stochastic: given the same input, it always returns the same output. It computes and outputs the most likely candidate for the masked token. This deterministic aspect makes BERT popular in situations where it is crucial to be able to reproduce results or when it is desirable to guarantee the same output across multiple iterations.

Despite the nontrivial differences between BERT and GPT, both models utilize attention, and although there is a difference regarding whether the attention mechanism is mono- or bidirectional in these models, the core logic of the mechanism – and the fact that the models use self-attention rather than cross-attention – is the same.

In the following, I explain how *BERTIFIER* implements BERT in ways inspired by digital literature, and how that relates to a poetics of the transformer architecture.

How it works

BERTIFIER is a program that implements the BERT model in a nontraditional way to interface with the attention mechanism in ways that are not possible using a GPT-style model and much less so in the conventional chat interface. The central procedure of the program is as follows. *BERTIFIER* takes an input sentence, masks the first word, uses BERT to predict a possible replacement to that word, moves on to the next word, uses BERT again, and so on. When the end of the sequence is reached, the program

loops back to the first word and does the same operation again. Since the first iteration most likely changed at least some words, the new input will not be exactly the same, meaning the attention mechanism performs its computations based on another set of vector representations of tokens in the context window. In turn, this means that BERT will likely return a different output. Each changed word may impact the prediction, so that even seemingly subtle replacements can mean a surprisingly new output, as some of the examples analyzed below show. The process continues to loop until the sentence has not changed for two iterations, which would mean that the sentence has reached a final state: the most bertish version of the input sentence possible.

Consider the example below. This article's title originates from this output from *BERTIFIER*. The input "Hello world!" results in the final output "the ground shook!" Most people who have taken any kind of programming class will be familiar with the input phrase used here, which is canonically the first thing you get a program to output when learning a new programming language. Following the example, I will explain the program in some more detail.

```
{
     "original_text": "Hello world!",
     "iterations": {
            "1": [
                  "[the] world!"
                  "the [truth]!"
                  "the truth [is]!"
            ],
            "2": [
                  "[the] truth is!"
                  "the [thing] is!"
                  "the thing [moved]!"
            ],
            "3": [
                  "[the] truth is!"
                  "the [ground] is!"
                  "the ground [shook]!"
            ]
            "4": [
                  "[the] truth is!"
                  "the [ground] is!"
                  "the ground [shook]!"
            ]
     },
     "bertified_text": "the ground shook!"
}
```

At every step of the bertification process – with each predicted word – we can read the attention mechanism in context. During execution, individual predictions that may, viewed in isolation, seem insignificant or subtle, snowball to create semantic mutations. It is not the correctness of the individual prediction but the larger mutation trail (as one might call it for lack of a better phrasing) that is interesting. The mutation trail points towards the attention mechanism in a much more specific way than what we might get from chatting with a consumer-facing chatbot, yet also in a much more situated way than attention visualization tools afford. The structures of text modeled by BERT become vividly available for literary engagement – and for readings deep and shallow. Since *BERTIFIER*'s mutation trace is both a literary output and a performative index of the operation of the attention mechanism, it integrates the architecture-centric and output-centric positions, discussed above.

Note that the example above displays a bug that inserts words before punctuation due to an error in the way I originally chose to handle punctuation in the program, which at times slightly warps sentence structures. Curiously, this insertion means that certain words, such as "again" or "now", are quite common as the last word in *BERTIFIER* output since only a few words can be added to the end of sentences while keeping every other word intact. This bug was not intentional yet results in strangely open-ended sentences and slightly warped sentence structures, which I have found to sustain more interesting reading experiences, and is accordingly kept in the program despite being easily fixable. Here, the poetics of working with BERT involves not only zooming in on technical specificity but also authorial judgment based on practice and tinkering.

This bug is nontrivial to the output in the example above: since the sentence is as short as it is, adding the extra word shifts the expressive possibilities. We can see this clearly in the first iteration. First, "hello" is replaced with "the" and as a result, the next word must be a noun ("truth"). Since the program inserts an extra token ("is"), the space of semantic possibility grows significantly. In the following iteration, we see the sentence morph into a noun-and-verb pair (curiously, in past tense). No single replacement is sudden, or even surprising, yet the difference between the original input and the final output is noticeable.

Another noticeable difference is the lack of capitalization in the output. BERT is available in a cased and an uncased version (the difference is whether the training data contains upper-/lowercase or only lowercase sentences). Most versions of *BERTIFIER* use the uncased version and one adaptation (so far) uses the cased version. The difference is not only one of appearance, but it also impacts the operation of the program: the cased BERT model will use letter-case to perform a prediction whereas

the uncased version will not, which may lead to different predictions. I mention this because paying attention (no pun intended) to these seemingly unimportant details in model configuration is crucial to reading *BERTIFIER*.

In this spirit, it is worth noting that the version of *BERTIFIER* discussed here removes whole words and replaces them with [MASK] tokens, instead of only removing singular WordPiece tokens (which would be a better fit for BERT's tokenizer). Of course, the BERT model still predicts tokens, and *BERTIFIER* then continues to predict and assemble predictions into words if the first prediction is a subword token. This implementation is done to make the process more easily legible to humans, and to allow the model to produce different kinds of words instead of being forced to predict a token that had to fit into a specific subword position.

To explicate how *BERTIFIER* picks up on small variations, consider the output that results from an almost completely similar input to the one presented above: "Hello, world!" (the same two words but with an added comma).[1]

```
{
     "original_text": "Hello, world!",
     "iterations": {
           "1": [
                 "oh my, god!!"
           ]
     },
     "bertified_text": "oh my, god!!"
}
```

The added comma is consequential to the output: instead of getting an article ("the") as the first replacement, we get "oh", which is much more likely to be followed by a comma. Once again, the aforementioned bug inserts extra tokens to nontrivial effect, but what is interesting here is that the comma ends up in a curious position. Syntactically, it should follow the "oh", but since *BERTIFIER* inserts new tokens *before* punctuation, the program is stuck in a situation where it needs to predict a word in-between the "oh" and the comma. The "oh" seems to have a higher attention score in the given situation, since "my" is predicted notwithstanding the awkwardness of the following comma. After the collocation of these two words – "oh" and "my" – it is completely unsurprising that the model predicts "god", again notwithstanding

[1] In the interest of saving space, all the following example outputs from BERTIFIER have been abridged by removing intermediate steps in the process as well as the final, repeating, iteration that marks the completion of the bertification process. The actual outputs follow the structure shown in the first example.

the peculiar position of the comma. Finally, we get an extra exclamation mark. The dominance of the collocation of the three words is displayed clearly by the fact that this version reaches peak bertishness after just one iteration.

It is crucial to emphasize that *BERTIFIER* is in no way an explanation of the way a transformer model computes input to produce output; it is not even causally explanatory with regard to the attention mechanism itself. We do not get any verifiable, technically causal insight from reading *BERTIFIER*'s gradual transformation of input into output across the iterations. Instead, we get to read our way to an understanding of these models, in a way that emerges from, and orients itself towards, digital culture.

In arguing that *BERTIFIER* has to do with poetics, I am specifically thinking of poetics as a form of practice, a way of doing – or making, from the Ancient Greek *poiesis* – literature and other forms of art. In other words, the argument is not that *BERTIFIER* is a poetics in-itself or that its outputs are poems *ipso facto* but that this system is an example of a kind of thing that results from literary experimentation with language models, and that such experimentation can be thought of in terms of poetics. Following the work of Nathan Jones, I here understand poetics to be literary experimentation that construes 'language as a system of logics that can be bent and broken to deviate from the conditions of the sayable delimited by existing norms' (2022: 22). In contemporary digital culture, Jones argues, poetics has to do with a form of practice that 'provokes us to consider the transformed situation for speaking, writing and thinking' in our current techno-cultural context (2022: 22). Poetics, in short, is not a noun but a verb.

In the following readings of *BERTIFIER* output, it is therefore crucial to keep in mind that if any of the outputs close in on something poetic, this is not just the result of a successful design of the program itself but, rather, of a productive relation between the program, the interpreter (in this case me), the reader, and a broader and continuously changing literary culture. As digital poet Allison Parrish (2021) has argued, the output of language models should always be seen as a form of linguistic material, which may (or may not) become literature as it interfaces with communities of readers.

In sum, *BERTIFIER* is less a single program and more a general approach to working with transformers and specifically BERT within the study and practice of digital culture. It points to a practice-based, literary form of engagement that does not take off-the-shelf products as the only viable configuration of technology but instead seeks to rework and understand our techno-cultural moment by doing something with(in) it. In the following, I turn to analyzing some selected *BERTIFIER* outputs to demonstrate the relation between the program and the attention mechanism.

Reading BERTIFIER

In many cases – particularly if input sentences are longer than the ones I include here – the output from *BERTIFIER* is excessively long, since it may require many iterations until a sentence has been bertified, and since one iteration may consist of many lines. As a consequence, I choose to focus on relatively short examples that are pragmatically possible to include within this format. This is not to say that the cases analyzed here are the most interesting ones: many of the lengthier examples are equally interesting, sometimes even more so from a literary perspective. These longer exchanges will, however, need to find another outlet, and my analysis of *BERTIFIER* will in this context limit itself to demonstrating its relevance to the questions considered here. The selection process behind the included examples has, accordingly, been guided by an interest in examples that do not require too many iterations while still being useful to demonstrate a connection between BERTIFIER and the attention mechanism in practice. A broader archive of some of my own personal favorite *BERTIFIER* interactions is available on https://gratisdigt.neocities.org.

Below, I offer a few tentative readings of *BERTIFIER* output. They are meant to signal trajectories for studying the operation of the attention mechanism in transformer models in ways that are both situated and particular – and they exemplify a kind of (electronic) literature that is specific to large language models.

On animals, colors, sizes

The first case is, like the initial example above, drawn from a canonical sentence within digital culture, this time the famous font-testing go-to sentence, containing every letter in the English alphabet: "The quick brown fox jumps over the lazy dog".


```
{
     "original_text": "The quick brown fox jumps over the lazy dog.",
     "iterations": {
          "1": [
                "a small white cat hovered over the sleeping girl again."
          ],
          "2": [
                "the big black bird hovered over the little girl again."
          ],
          "3": [
                "the little black bird hovered over the little girl again."
          ]
     },
     "bertified_text": "the little black bird hovered over the little
     girl again."
}
```

My focus here lies on the curious sets of animal-color-size relations that come to the fore. In the first iteration, we see the "quick brown fox" turn into a "small white cat". This seems to be a sign of a strong relation between the size, the color, and the animal. Note that in the original sentence, the first adjective has nothing to do with color, but it gets turned into a matter of color via BERT's prediction. In the following iteration, the "small white cat" turns into a "big black bird". Since *BERTIFIER* works in a sequential manner from left to right, the relation is, here, somewhat hierarchical: the size is predicted before the color, before the animal. If something is "small", then it is likely "white", and if it is both "small" and "white", it is likely a "cat"; if something is "big", it is likely "black", and if it is both "big" and "black", it is likely a "bird" (a blackbird or, possibly, a crow or raven). The final, third, iteration, showcases some aspect of the attention pattern, since the animal continues to be a "black bird" but is now little. The collocation of "black" and "bird" weighs more than the size, which is changed to fit the animal in the final iteration. (Although the notion of a "big black bird" is far from unrealistic, "little black bird" is more attuned to the cultural image of a blackbird.)

There is potentially a racial bias at play here: "small" seems to coincide with "white" whereas "big" coincides with "black", unless we are talking about a "black bird", in which case it triggers the size to be "little". These patterns are reminiscent of well-documented cases of racial bias in language technology (*cf*. Noble, 2018). In this sense, this case illustrates how the attention mechanism seems to work as a kind of nonconscious cognition, as Hayles (2025) has argued, where the output is a result of nonrandom (and highly normative) decisions. Note that the output here displays another manifestation of the bug that inserts tokens before punctuation: it appends an "again" to the sentence, which does not change much about its structure yet does add a slightly ominous feeling to the output. We could comment on other aspects of this output – like the strangeness of the verb, "hovered", which emerges already in the first iteration despite the fact that cats do not usually hover. However, in the interest of brevity, let us move on to the next case.

Frankenlove

For the next case, we drift into a more explicitly literary arena: the input is taken from Mary Shelley's *Frankenstein*. The case demonstrates a tendency for *BERTIFIER* to indulge in repetitions – which is a well-known effect of earlier generations of language models. In the output, the repetitions result in a sense of gentle, almost naive reflection. In addition, if all the iterations are read as verses belonging to a single poem, the slight variations in the mutation trail of *BERTIFIER* result in an almost narrative progression in the poem.

```
{
     “original_text”: “Life, although it may only be an accumulation of
     anguish, is dear to me, and I will defend it.”,
     “iterations”: {
           “1”: [
                 “that love, though it may well be an expression of
                 love itself, is important to me now, and i must
                 accept it now.”
           ],
           “2”: [
                 “true love, though it may not be an act of love itself,
                 is important to me now, and i can feel it now.”
           ]
           “3”: [
                 “true love, though it may not be an act of love anymore,
                 is happening to me now, and i can feel it now.”
           ]
           “3”: [
                 “true love, though it may not be an act of love itself,
                 is happening to me now, and i can feel it now.”
           ]
     },
     “bertified_text”: “true love, though it may not be an act of love
     itself, is happening to me now, and i can feel it now.”
}
```


First, we see how already in the first iteration, “love” appears twice, as does “now”, which gives the sentence a kind of symmetry. The appearance of the first instance of “love”, as well as both appearances of “now” are the result of the (by now well-known) bug that inserts words before punctuation. The second appearance of “love” is likely a direct consequence of the words that come immediately before it, “an expression of”, since “love” would be a very common word to follow that string. Once again, the inserted word “itself” adds nontrivial semantic weight, since it infers a distinction between “love” and (the expression of) “love itself”, which (at least to me) bears some contemplation. That distinction becomes more poignant in the following iterations, where “true love” is taken to *not* be “an act of love itself”. Of course, it is easy to infer that true love is indeed not an act of love but instead the cause of other acts of love, yet the structure of the sentence goes further as it begs the question: if not an act of love, then what is it an act of?

The slight variations between the iterations mark a shift from love being "important" to later simply "happening", as a result of the change between the need to accept it and the sudden feeling of it. In the meantime, "itself" changes to "anymore" in the third iteration, and then back again, indicating a change in the way love is conceived as the narrator begins to feel it and, simultaneously, stops thinking of it as important and simply observes that it is happening. In this case, we can see how the attention mechanism picks up on the changes in words in different places and, as a result, *BERTIFIER* adjusts some of its predictions until it reaches the final, stable, form. Read as a poem, these slight variations contain a multitude of possible trajectories of interpretation that we shall not follow through to completion here, but which indicate the literary possibilities of opening up the generation process and tinkering with output formats rather than accepting the helpful assistant-imbued output format of popular chatbots as a default.

Exclamation marks galore

We stay in literary territory with the next case, where the input is taken from Allen Ginsberg's 1956 poem *Howl*. The case shows how BERTIFIER seems to support the idea that language models operationalize the poetic function of language, and how the program tends to retain a similarity in terms of style, theme, or voice – or, perhaps more succinctly, in terms of vibe (*cf*. Grietzer, 2025).


```
{
     "original_text": "Moloch! Solitude! Filth! Ugliness! Ashcans and
     unobtainable dollars! Children screaming under the stairways! Boys
     sobbing in armies! Old men weeping in the parks! Moloch! Moloch!
     Nightmare of Moloch! Moloch the loveless! Mental Moloch! Moloch
     the heavy judger of men!",
          "iterations": {
               "1": [
                    "solitude!! misery!! misery!! misery!! cheap,
                    cheap!!! men crying on the streets!! men fighting
                    in war!!! men fighting in the streets!!!!!!!! sons
                    of men!!!!!!!!!!!!!! sons of men!!"
               ],
               "2": [
                    "misery!!!!!!!!!! cheap!!!!! men fighting in the
                    streets!! men fighting for money!!! men fighting
                    in the streets!!!!!!!! sons of men!!!!!!!!!!!!!!
                    sons of men!!"
               ],
```

```
                "3": [
                        "!!!!!!!!!!!!!!!!! men fighting in the streets!!
                        men fighting for money!!! men fighting in the
                        streets!!!!!!!! sons of men!!!!!!!!!!!!! sons
                        of men!!"
                ]
        },
        "bertified_text": "!!!!!!!!!!!!!!!!! men fighting in the streets!!
        men fighting for money!!! men fighting in the streets!!!!!!!! sons
        of men!!!!!!!!!!!!! sons of men!!"
}
```


The input sentence displays the signature staccato rhythm and exclamation-based tone of *Howl*, and the words portray the industrial ruins of so-called modern life in 1950s urban USA. The figure of Moloch, repeated many times in the relatively short input, is a central to Part II of *Howl*, where it functions as a personification of a destructive logic of industrialism. As we saw in the case above, *BERTIFIER* has a tendency to produce or cling onto such repetitions, yet in the output, many of the instances of Moloch are replaced with exclamation marks, already in the first iteration. It seems altogether that *BERTIFIER* is highly sensitive to the punctuation in this case, as it overwhelmingly predicts repeated exclamation marks. Most notably, the first 18 characters of the final output are exclamation marks, which may seem odd from a syntactic perspective yet works well to establish the general vibe of both the input sentence and the output.

The men fighting for money in the streets is perhaps less evocative than the original segment of *Howl*, but they do fit well with the theme of the input, not least in the repetition and the generous addition of exclamation marks. Insofar as language models operationalize the poetic function of language – as Weatherby (2025) argues – and insofar as that poetic function is to be observed in those cases where language points to itself, this output becomes a poetic investigation of the exclamation mark, or more precisely, of the difference between two, three, eight, 14, and 18 exclamation marks in a row.

Glitch or imposter?

The final case I will consider here is perhaps less canonical than the previous cases. The input sentence '*Being represented by data is like losing a part of yourself*' is the title of a video artwork from Danish artist Ada Ada Ada, which circles themes of representation, datafication, gender, and sexuality (2024).

```
{
     “original_text”: “Being represented by data is like losing a part
     of yourself.”,
     “iterations”: {
           “1”: [
                 “being surrounded by people was like being a part of
                 life again.”
           ],
           “2”: [
                 “being surrounded by them was like being a part of
                 them again.”
           ]
     },
     “bertified_text”: “being surrounded by them was like being a part
     of them again.”
}
```


One theme in the work is the curious feeling of thriving in the role of the glitch, as the instance that does not fit into binary categories. To embrace existence as a glitch. However, already in the first *BERTIFIER* iteration, we see the theme shift from the jarring experience of “being represented by data” to a more comforting description of “being a part of life” by way of “being surrounded by people”. In the following iteration, the sentence turns into a tale of assimilation, or possibly of impostering: no longer surrounded by “people” but by “them” gives an experience that is “like being a part of them again”. Like the previous case, the “again” was added by the bug adding tokens before punctuation in a way that carries nontrivial semantic weight. In the final sentence, we have a tale of someone who used to be part of a specific group but no longer is; being surrounded by members of that group, then, does not so much make the person part of the group (which would turn the “them” into an “us”), but is instead “like” being part of the group “again”.

In some senses, this case changes the least of the ones analyzed here, yet exactly because of the subtlety of the changes that do happen, the semantic shift is nontrivial. Instead of describing the somewhat marginal position as a glitch, which is a place wherein the artist thrives, the sentence becomes a story of impostering or, potentially, of being interpellated as a member of a (likely) majority group despite not being (or at least not identifying) as being part of that group.

A practice-based poetics

Digital literature – and, more broadly, digital culture – currently finds itself in a curious and almost paradoxical situation with respect to large language models and their various multimodal variants. On the one hand, these systems are eagerly discussed in popular discourse as well as academic scholarship (including in this very piece). Our fascination with this technology is likely at least in part due to the seemingly insurmountable (and equally intriguing) difficulty associated with studying it, leading to nontrivial discussions of cornerstone assumptions and approaches within our fields. On the other hand, proprietary chatbots usually produce incredibly boring output. Indeed, interest in the outputs of off-the-shelf chatbots is dwindling fast. As Tuuli Hongisto (2025) demonstrates, we have witnessed a rapid decrease in the amount of self-published books on Amazon that list ChatGPT as an author. Of course, this does likely not mean that people are not using chatbots to produce self-published books, but that doing so is considered not-very-interesting and thus not something that warrants declaration.

This decreasing interest in the output of consumer-facing chatbots is not limited to Amazon but also shows itself clearly within communities invested in computer-generated literature. This despite the fact that, on the surface, the topic of computer-generated literature has never before been subject to as much discussion as is currently the case. As two cases in point, consider two fairly recent calls for computer-generated writing (from, respectively, Ensemble Park and Michigan Quarterly Review). Both discourage submissions taking the form of output from a chatbot without a clear literary purpose or significant conceptual development. As the call from Michigan Quarterly Review makes clear, a submission should 'do more than demonstrate the ability of certain software and programs to generate cogent text' (Michigan Quarterly Review – A Special Issue on Computer-Generated Text, 2025). Similarly, Ensemble Park states that although chatbot-based submissions are welcome, they should 'reflect on the role of the model in the writing process, and push the "incorporating suggested generated text" paradigm into new forms' (Ensemble Park, n.d.).

The clear-cut matter-of-factness of these tendencies – the decreasing amount of ChatGPT-authored books on Amazon and the discouragement of chatbot output in journals of computer-generated writing – clearly demonstrate that theory is not enough. My point is by no means that the philosophical and conceptual questions raised by scholars are irrelevant. Yet our intellectual fascination with these technologies seems at times to be slightly out of tune with what is going on in practice. As I also noted in the introduction, there are ample interesting, stimulating, and potent examples of poetry and prose written with large language models, including GPT-style chatbots.

But our scholarly engagements with these technologies seem to be largely cut-off from the situated, literary experimentation taking place. The problem is not that we have insufficient theories as such or that it is impossible to create interesting things with language models, but that theory and practice are oddly separated. In effect, we seem to be stuck trying to understand proprietary, deliberately opaque products rather than insisting on rethinking the whole situation by performing both theory and practice differently.

This is why we are in dire need of a *practice-based* poetics of the transformer architecture – and, in turn, of other components of contemporary language models or of the transformer architecture as such. As Hannes Bajohr argues in one of the earliest treatments of media-specificity with respect to artificial neural networks (pre-transformer), we must insist on 'a non-anthropocentric [...] hermeneutical understanding that is aimed [...] at comprehending the process of the work's material production' (2022: 210). This is, however, not only a matter of our analytical approach but also of how we work with those systems in practice to make them available for such analysis. Bajohr seems to know this, or at least to enact it by taking a practice-based position, bridging literary production and theory. Yet, whereas Bajohr (2023) has argued that the availability of systems that can produce human-comparable text leads to a *post-artificial* situation where distinguishing human-written from machine-generated texts becomes a meaningless project, I believe a more likely conclusion would be that we will see a thoroughgoing blending of the acts of reading and writing, where practice and process – not only among professional writers but in general – will become even more central than it already is. In other words, post-artificiality means not that artificial text cancels all other forms, but that we find ourselves in a situation where we need to figure out how to exist meaningfully in a world riddled with corporate bullshit.

The assumption that large language models cancel or subvert media-specific considerations of the modes of production of computer-generated literature simply because it is *possible* to generate text that cannot be reliably identified as being generated by a computer misses the point. People who are invested in showing readers their methods, their processes, their algorithms do not do so out of necessity, that is, because they were unable to produce text that passes as human-written. On the contrary, such authors often go out of their way to make the media-specificity of their work legible, both as paratexts and as recognizable linguistic imprint in the texts themselves. The point has never been to fool the reader but to indulge them in all the poetic weirdness that characterizes writing with computers – and to revel in the outcome.

This point is also related to the politics of digital literature. As I mentioned in the introduction but have not scrutinized in detail, current off-the-shelf chatbots are

problematic in many ways. I will not open this issue up now (it is, due to limitations of space, beyond the scope of this article), but I will note that the effort to enforce a culture that does not care about the process of writing so long as the product is human-comparable has strong corporate undertones (*cf.* Dini, 2025). Although we may not be able to confidently tell where a text comes from, that does not mean the materiality of the technologies and infrastructures used in its creation and propagation is unimportant. The ties between current mega-infrastructures of generative artificial intelligence and power are deep, and deserve contentious attention, and resistance, at every opportunity (Klein & D'Ignazio, 2024). An important role of art and literature in this regard is to grapple with these issues and to point to alternative configurations.

BERTIFIER is an effort to recenter critical questions of media-specificity with regard to large language models, particularly transformers, within digital culture. The task for *BERTIFIER* – and I would hope much more critical scholarship within digital culture and critical digital humanities in the coming years – is to reestablish media-specificity and material contention as principal ways of operating, and to assume a proactive position vis-à-vis technological mega-structures. Rather than sticking to interpreting what the model's internal operations might mean for language and poetics, it operationalizes those dynamics as a site of writing.

BERTIFIER is not intended to be a competing product in the chatbot economy as such, but it is intended to function as part of a host of literary and artistic alternatives that center practice and demonstrate its continued importance. In other words, *BERTIFIER* is both a method and an example of a poetics of the transformer architecture. Viewed as poetics, the work surrounding the *BERTIFIER* program 'is inherently concerned with the conditions and means of its production and use: the systems, standards, politics and tendencies of language' in the context of generative artificial intelligence (Jones, 2022: 55). If corporate optimization gives us narrative sameness and formal stuckness, literary alternatives should work not as optimizations towards literariness but as sites for contentious engagement, cultural exchanges, and a practice-based poetics. This requires media-specificity – the integration of an architecture-centric stance with an output-centric approach – as well as a willingness to look beyond the latest releases from Silicon Valley.

The ground may be shaking under our methodological feet, but we should not take that as a sign that it no longer exists. Let us, instead, shake along and shake back, to establish new or rediscovered connections amid the quake. Let us reopen the question of what large language models are within digital culture and how we interface with them.

– – –

As an instructional addendum, I offer one last output from *BERTIFIER*. I will not share the full mutation trail of this output and neither the input text, yet I will disclose that it comes from a pre-computational example of algorithmic writing that involves a newspaper, a pair of scissors, and a bag.

> do it very carefully. use your pen again. take out all of the pages from the book you used to write this article in. take out your pen again. very carefully take out all of the pages that make up this article and put them back in the book again. think very carefully. then take out the pages one after the other again. think carefully about the order in which you put the pages back. the memories that haunt you now. and then you have a very clear view of the world again, as if surrounded by a vast cattle herd.

Acknowledgements

I thank my colleagues who participated in the XDN Seminar in Bergen in October 2024 for generous discussion of the ideas presented here. Thanks also to my colleagues in the HAIC-III project, Søren Pold and Kristin Veel, for constructive comments during the writing process. Finally, I thank the anonymous reviewers for their help in the tightening the argument.

This work was partially supported by the Research Council of Norway through its Centers of Excellence scheme, project number 332643 (Center for Digital Narrative), and its SAMKUL project scheme, project number 335129 (Extending Digital Narrative). The work is also partially supported by the Independent Research Fund Denmark, grant ID 10.46540/4256-00095B (Human-AI Collaboration: Imaginaries, Interventions, Interfaces), and the Danish National Research Foundation, grant ID DNRF193 (TEXT: Center for Contemporary Cultures of Text).

Competing Interests

The author has no competing interests to declare.

References

Ada Ada Ada 2024 Being represented by data is like losing a part of yourself. Available at: https://ada-ada-ada.art/projects/being-represented-by-data-is-like-losing-a-part-of-yourself [Last Accessed 3 June 2025].

Bajohr, H 2022 Algorithmic Empathy: Toward a Critique of Aesthetic AI. *Configurations*, 30(2): 203–231. https://doi.org/10.1353/con.2022.0011.

Bajohr, H 2023 Artificial and Post-Artificial Texts: On Machine Learning and the Reading Expectations Towards Literary and Non-Literary Writing. *Basel Media Culture and Cultural Techniques Working Papers*: 1–31. https://doi.org/10.12685/bmcct.2023.007.

Bajohr, H 2024 Cohesion Without Coherence: Artificial Intelligence and Narrative Form. *Transit*, 14(2) https://doi.org/10.5070/T714264658.

Bajohr, H (ed.) 2025 *Thinking with AI: Machine Learning the Humanities*. London: Open Humanities Press.

Bertram, L-Y and **Montfort, N** 2024 *Output: An Anthology of Computer-Generated Text, 1953–2023*. MIT Press.

Burrell, J 2016 How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1): 205395171562251. https://doi.org/10.1177/2053951715622512.

Cayley, J 2023 Modelit: eliterature à la (language) mode(l). *Electronic Book Review* https://doi.org/10.7273/2BDK-NG31.

Clark, K, Khandelwal, U, Levy, O and **Manning, C D** 2019 What Does BERT Look at? An Analysis of BERT's Attention. In: *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics. p. 276–286.

Devlin, J, Chang, M-W, Lee, K and **Toutanova, K** 2019 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv. http://arxiv.org/abs/1810.04805 [Last Accessed 15 October 2024].

Dini, R 2025 Literature Is Not a Vibe: On ChatGPT and the Humanities. *Los Angeles Review of Books*, 31 October 2025. Available at: https://lareviewofbooks.org/article/literature-is-not-a-vibe-on-chatgpt-and-the-humanities [Last Accessed 29 July 2026].

Dobson, J E 2023 On reading and interpreting black box deep neural networks. *International Journal of Digital Humanities*, 5(2–3): 431–449. https://doi.org/10.1007/s42803-023-00075-w.

Ensemble Park n.d. 'Ensemble Park: A Journal of Human and Computer Writing'. Available at: https://www.ensemblepark.com/ [Last Accessed 7 March 2025].

Erslev, M S 2023 *Machine Mimesis: Electronic Literature at the Intersection of Human and Computer Imitation*. Ph.D. Dissertation, Aarhus University.

Erslev, M S n.d. 'BERTIFIER: reaching peak bertishness one token at a time'. Available: https://gratisdigt.neocities.org/ [Last Accessed 17 December 2025].

Gillespie, T 2024 Generative AI and the politics of visibility. *Big Data & Society*, 11(2): 20539517241252131 https://doi.org/10.1177/20539517241252131.

Ginsberg, A 1956 Howl. In: *Howl and Other Poems*. San Fransisco, CA, USA: City Lights.

Grietzer, P 2025 A Theory of Vibe. In: Bajohr, H (ed.) *Thinking with AI: Machine Learning the Humanities*. London: Open Humanities Press. p.20–32.

Hamilton, S and **Mimno, D** 2026 Elias in the Lighthouse, Again? Diagnosing Low Diversity in LLM Stories. *arXiv*. http://arxiv.org/abs/2605.26492 [Last Accessed 29 July 2026].

Hayles, N K 2004 Print Is Flat, Code Is Deep: The Importance of Media-Specific Analysis. *Poetics Today*, 25(1): 67–90.

Hayles, N K 2017 *Unthought: The Power of the Cognitive Nonconscious*. University of Chicago Press.

Hayles, N K 2025 *Bacteria to AI: human futures with our nonhuman symbionts*. Chicago, IL: University of Chicago Press.

Heuser, R 2025 Cultural Collapse: Toward a Generative Formalism for AI Cultural Production. *Anthology of Computers and the Humanities*, 3: 575–588. https://doi.org/10.63744/USvuyzSlapvy.

Hicks, M T, Humphries, J and **Slater, J** 2024 ChatGPT is bullshit. *Ethics and Information Technology*, 26(2): 38. https://doi.org/10.1007/s10676-024-09775-5.

Hongisto, T 2025 Advertising with AI – On the presentation of authorship of ChatGPT-generated books. *Electronic Book Review*. https://doi.org/10.64773/c3f2-7c45.

Huhtamo, E and **Parikka, J** (eds.) 2011 *Media archaeology: approaches, applications, and implications*. Berkeley, Calif: University of California Press.

Johnston, J D 2019 *ReRites: human + A.I. poetry; Raw output: A.I. trained on custom poetry corpus; Responses: 8 essays about poetry and A.I.* Montreal, Quebec: Anteism Books.

Jones, N 2022 *Glitch Poetics*. S.l.: Open Humanities Press.

Kamath, H, Ameisen, E, Kauvar, I, Luger, R, Gurnee, W, Pearce, A, Zimmerman, S, Batson, J, Conerly, T and **Olah, C** 2025 Tracing attention computation: Attention connects features, and features direct attention. *Transformer Circuits Thread*. Available at: https://transformer-circuits.pub/2025/attention-qk/index.html.

Klein, L and **D'Ignazio, C** 2024 Data Feminism for AI. In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. p. 100–112. https://doi.org/10.1145/3630106.3658543.

Kovaleva, O, Romanov, A, Rogers, A and **Rumshisky, A** 2019 Revealing the Dark Secrets of BERT. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics. p. 4364–4373. http://doi.org/10.18653/v1/D19-1445.

Lee, H P, Sarkar, A, Tankelevitch, L, Drosos, I, Rintel, S, Banks, R and **Wilson, N** 2025 *The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers*. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Yokohama, Japan, 26–30 April 2025. New York: ACM. https://doi.org/10.1145/3706598.3713778.

Lin, J n.d. Neuronpedia. *Neuronpedia*. Available at: https://www.neuronpedia.org [Last Accessed 28 July 2026].

Michigan Quarterly Review – A Special Issue on Computer-Generated Text 2025. Available at: https://mqr.submittable.com/submit/318008/a-special-issue-on-computer-generated-text [Last Accessed 7 March 2025].

Noble, S U 2018 *Algorithms of oppression*. New York University Press.

Parrish, A 2021 Language models can only write poetry. *Decontextualize*. Available at: https://posts.decontextualize.com/language-models-poetry [Last Accessed 14 December 2021].

Parrish, A n.d. 'Hi, I'm Allison Parrish'. Available at: https://www.decontextualize.com/ [Last Accessed 27 July 2026].

Raji, I D, Bender, E M, Paullada, A, Denton, E and **Hanna, A** 2021 AI and the Everything in the Whole Wide World Benchmark. *arXiv:2111.15366 [cs]*.

re•mediate lit mag n.d. *re•mediate lit mag*. Available at: https://remediatelitmag.xyz/p22811773/ [Last Accessed 17 December 2025].

Rettberg, J W and **Wigers, H** 2025 AI-generated stories favour stability over change: homogeneity and cultural stereotyping in narratives generated by gpt-4o-mini. *Open Research Europe*, 5: 202. https://doi.org/10.12688/openreseurope.20576.1.

Rettberg, S 2019 *Electronic literature*. Cambridge, UK; Medford, MA, USA: Polity Press.

Rettberg, S 2023 Cyborg Authorship: Writing with AI – Part 1: The Trouble(s) with ChatGPT. *Electronic Book Review*. https://doi.org/10.7273/5SY5-RX37.

Rettberg, S and **Saum-Pascual, A** 2020 Introduction: Electronic Literature as a Framework for the Digital Humanities. *Electronic Book Review*. https://doi.org/10.7273/vyqe-dq93.

Rogers, A, Kovaleva, O and **Rumshisky, A** 2020 A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, 8: 842–866. https://doi.org/10.1162/tacl_a_00349.

Saum-Pascual, A 2025 Material Algorithms: Bridging the Gap between the Virtual and the Physical in Latin American Digital Literature. *Journal of Latin American Cultural Studies*, 34(3): 389–412. https://doi.org/10.1080/13569325.2025.2473387.

Templeton, A, Conerly, T, Marcus, J, Lindsey, J, Bricken, T, Chen, B, Pearce, A, Citro, C, Ameisen, E, Jones, A, Cunningham, H, Turner, N L, McDougall, C, MacDiarmid, M, Tamkin, A, Durmus, E, Hume, T, Mosconi, F, Freeman, C D, Sumers, TR, Rees, E, Batson, J, Jermyn, A, Carter, S, Olah, C and **Henighan, T** 2024 Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *Transformer Circuits*. Available at: https://transformer-circuits.pub/2024/scaling-monosemanticity/ [Last Accessed 8 May 2025].

The AI Literary Review n.d. 'Welcome to the A.I Literary Review'. Available at: https://ailiteraryreview.co.uk/ [Last Accessed 17 December 2025].

Thorne, E n.d. 'Elias Thorne'. Available at: https://eliasthorne.org/ [Last Accessed 29 July 2026].

Vig, J 2019 BERTVIZ: A TOOL FOR VISUALIZING MULTI-HEAD SELF-ATTENTION IN THE BERT MODEL.

Weatherby, L 2025 *Language machines: cultural AI and the end of remainder humanism*. Minneapolis London: University of Minnesota Press.

Zheng, L, Sheng, Y, Chiang, W L, Zhang, H, Gonzalez, J E and **Stoica, I** 2023 Chatbot Arena: Benchmarking LLMs in the Wild with Elo Ratings | LMSYS Org. *LMSYS Org*. https://lmsys.org/blog/2023-05-03-arena [Last Accessed 28 May 2025].