Carlin Soos & Levon Haroutunian. 2023. On the Question of Authorship in Large Language Models (LLMs). NASKO, Vol. 9. pp. 1-17.

Carlin Soos – University of California, Los Angeles (UCLA)
Levon Haroutunian – University of Washington

On the Question of Authorship in Large Language Models (LLMs)

Abstract

The adoption of pre-trained large language models (LLMs), like ChatGPT, across an increasingly diverse range of tasks and domains poses significant challenges for authorial attribution and other basic knowledge organization practices. This paper examines the theoretical and practical issues introduced by LLMs and describes how their use erodes the supposedly firm boundaries separating specific works and creators. Building upon the author-as-node framework proposed by Soos and Leazer (2020), we compare works created with and without the use of LLMs; ultimately, we argue that the issues associated with these novel tools are indicative of preexisting limitations within standard entity-relationship models. As the growing popularity of generative AI raises concerns about plagiarism, academic integrity, and intellectual property, we encourage a reevaluation of reductive work/creator associations and advocate for the adoption of a more expansive approach to authorship.

1. Introduction

OpenAI's release of ChatGPT in November 2022 triggered near-immediate concerns across college campuses. Perhaps still on guard from a reported rise in cheating attributed to the remote-instruction phase of the COVID-19 pandemic (Jenkins et al. 2022; Dey 2022), administrators and faculty quickly began speculating about widespread misuse of the chatbot. An abundance of news coverage questioning ChatGPT's ability to "replace humans" (Lock 2022) no doubt exacerbated these anxieties, leading to concerns about "radical consequences for teaching and learning" (Dolan 2023). As is common with these types of technological innovations, the panic subsided almost as quickly as it emerged, leaving in its wake a lingering malaise and ambivalence. Although concerns persist about the use of generative AI for cheating, university talking points now strike a balance of offensive disciplinary policies and practical recommendations (e.g. University of Washington 2023; UCLA 2023; University of Wisconsin-Madison 2023).

As the pedagogical value of ChatGPT and its competitors continues to be explored (Kasneci et al. 2023), educators are finding new ways to utilize the unique capabilities of these large language models (LLMs) without compromising the integrity of their classrooms. While some schools are attempting to outright ban all applications of generative AI, its utilization by students, staff, and faculty seems inevitable. As distinct entities, ChatGPT-like tools can be too tempting to resist, and their allure is only heightened by an increase in social, professional, and economic pressures looming over learners and their instructors. Complicating matters further is the embedding of these models into preexisting information retrieval systems, such as Microsoft's use of ChatGPT for Bing or Google's new "collaborative AI service," Bard (Elias 2023). As the line between "chatbot" and "search engine" is further blurred, determining where "researching"

ends and "cheating" begins will likely become increasingly more difficult. Viewed from this perspective, initial warnings about how ChatGPT will "upend longstanding concepts of plagiarism, authorship, ownership, and learning" (McCarthy 2023) are not entirely unfounded. However, with these new challenges comes an opportunity to revisit each of these concepts, question our preexisting assumptions about their conceptual validity, and develop new perspectives to match the current moment.

1.1. Paper Goals and Structure

This project seeks to address the implications of LLMs for authorial attribution and other knowledge organization (KO) practices. In the following section, we provide a technical overview of LLMs and introduce relevant literature from the field of natural language processing (NLP). Next, we review theories of authorship within KO, focusing primarily on Soos and Leazer's concept of the "author-as-node" (2020). Building upon this network theory, we proceed to discuss the various authorship issues introduced by generative AI and describe how the nature of pre-trained models—as well as their creative outputs—complicate the supposedly firm boundaries separating specific works and creators. With these considerations in mind, we expand the author-as-node framework using the concept of "communicative intent" (Bender and Gebru et al. 2021). To conclude, we reiterate and reaffirm previously acknowledged concerns about "the author" as a distinct categorical entity while maintaining the importance of idea attribution for personal development and community accountability.

2. Technical Overview of LLMs for KO

Language modeling is the task of computationally representing how humans use language. In practice, language models typically predict and generate a sequence of words given another sequence as context. These models are a useful component of nearly every kind of NLP system, from automatic speech recognition to machine translation to natural language generation.

Language models based on neural networks are far and away the most common types used today. The simplest type of neural language model is a feed-forward neural network (Bengio et al. 2003), which is composed of a number of layers containing sets of units typically referred to as "neurons." The first layer is an embedding layer, which converts the individual words of an input into vectors of numeric values. Every unit—or neuron—of this embedding vector is connected to every neuron in the next layer through a weight and a bias value. The first stage of computation applies those weights and biases to the initial vector; a nonlinear function (such as a sigmoid function) is then applied to the initial vector to determine the values of each neuron in the second layer. The neurons in the second layer are similarly connected to the neurons of the third layer, and so on. More complex types of neural models incorporate different types of connections between neurons, which are necessary to account for the sequential nature of text data. Weights

and biases are generally referred to as "parameters," and a model's size is usually described by its number of parameters.

During training,[1] neural language models are optimized on token prediction tasks, where they must predict output text based on input text. Input text is first split into a sequence of tokens, which are typically words or sub-word pieces. These tokens are then mapped to corresponding vectors, which are then run through the matrices that comprise the model's parameters. The output of this computation is another sequence of vectors. This sequence of vectors can then be compared against the expected output, which is tokenized and mapped to a sequence of vectors in the same manner the input was. The result of this comparison is a loss score, which determines the degree to which the model output differs from the expected output. Using the Chain Rule from multivariable calculus, the training routine updates the model's parameters in a way that reduces the loss score; in other words, the parameters are modified to push the output closer to the expected output. This process repeats for every input/output pair in the training data. Typically, training concludes after many full passes (called "epochs") over the training data.

Neural language model training results in what is referred to as a parametric memory: language models distill and "memorize" their training data in their parameters to produce output that aligns with the observed data patterns. This parametric memory is the sum total of the "knowledge" that a language model has; after training, a language model has no access to its training data or information from any other source.

Once a language model is trained, one can produce output from it by taking an input text, splitting that text into a sequence of tokens, mapping those tokens into corresponding vectors, and running those vectors through the model's parameters. The model produces output by iteratively predicting the vector of the next token in the sequence; in other words, it predicts the most likely continuation of the input, based on the information stored in its parametric memory.

2.1. Scaling Up: The Birth of Large Language Models

In 2017 (Vaswani et al.), the advent of a specialized type of neural network, called a Transformer, gave rise to a new era in language modeling. One of the first examples of a Large Language Model is BERT, a Transformer-based language model that advanced the state of the art on many common NLP benchmark tasks (Devlin et al. 2019).

The shift to Transformer-based language models marked an increase in both the size of models and the data used to train them. BERT has approximately 110 million parameters—which is relatively massive compared to its contemporaries—and a similarly large training corpus: English Wikipedia, which included 2.5 billion words in the version the authors used; and BookCorpus (Zhu et al. 2015), which contained around 800 million words pulled from

[1] This description is a simplified explanation of a typical training routine for a neural language model. For a more detailed overview, see Jurafsky and Martin (2023).

approximately 11,000 unpublished books scraped from Smashwords. Following BERT was a flood of pre-trained language models, with the notable examples of ERNIE (Zhang et al. 2019), GPT-2 (Radford et al. 2019), XLNet (Yang et al. 2019), BART (Lewis et al. 2020), T5 (Raffel et al. 2020), and GPT-3 (Brown et al. 2020).

The creation of GPT-3 in 2020 marked the apex of increases to model size; it has a whopping 175 billion parameters, almost 1,600 times larger than BERT. GPT-3's gigantic scale came along with, of course, a gigantic training set, which includes English Wikipedia and BookCorpus along with the CommonCrawl dataset, which is a web crawl dataset consisting of the text from billions of web pages. Like BERT before it, GPT-3 showed impressive performance gains on a variety of NLP tasks.

GPT-3 also marked the beginning of a new LLM paradigm. Previous LLMs were usually not directly applied to specific tasks of interest; instead, researchers would download a pre-trained language model like BERT and train its parameters further on a smaller set of task-specific data—a process known as fine-tuning. Because GPT-3 was released closed-source, its users could not simply download the model and train it further. However, GPT-3 achieved impressive performance *without* being fine-tuned, through a method called in-context learning (ICL; Brown et al. 2020). To apply ICL, a user supplies a "prompt" to the model that includes a handful of in-context demonstrations (e.g. a few examples of English sentences paired with their French translations) along with their input to the model (e.g. a new English sentence), and the model is expected to produce output in format given by the demonstrations (e.g. the French translation of the input). In this way, GPT-3 functions as a general-purpose LLM; it is intended to be used on a wide variety of tasks, with no need (or option) to customize it.

2.2. Data

For the reasons described above, massive corpora are a necessity to creating large language models: neural language models get their power from their parametric memory, and their parametric memory comes from the data the models ingest during training. Unfortunately, the sheer size of these corpora means that researchers who create them or use them cannot be fully aware of what they contain (Paullada et al. 2020). The opacity of many of these large datasets is due to what Bender and Gebru et al. (2021) call "documentation debt," which is "a situation where the datasets are both undocumented and too large to document post hoc" (615). Numerous audits of large machine learning datasets have found that they contain non-trivial amounts of unwanted content (Dodge et al. 2021), copyright violations (Bandy and Vincent 2021), sexually explicit material (Birhane et al., 2021), and hate speech (Gehman et al. 2020). For a more detailed critique of practices surrounding the collection and use of machine learning datasets, see Paullada et al. (2020).

Take for example CommonCrawl, which is one of the datasets used to train GPT-3 and its successors. CommonCrawl is an effort by The Common Crawl Foundation to "[democratize]

access to web information by producing and maintaining an open repository of web crawl data" (Common Crawl n.d.). As of April 2023, CommonCrawl contains 3.1 billion web pages (Nagel 2023). An analysis by Luccioni and Viviano (2021) found that around 5% of the web pages included in CommonCrawl contain hate speech and slurs. There have been many efforts to filter CommonCrawl (most notably, C4; Raffel et al. 2020), including by Brown et al. (2020) in their creation of GPT-3. However, it is virtually impossible to comprehensively filter or audit a data set on the scale of CommonCrawl. Additionally, as Bender and Gebru et al. (2021) point out, the nature of Internet data means that datasets like CommonCrawl necessarily overrepresent the voices of young, male Internet users in developed countries at the expense of marginalized people.

With language models as large and complex as GPT-3, it can be difficult to conceptualize the links between the data it was trained on and the output it produces. However, an understanding of the training data used to train an LLM should be in the foreground of any attempts to determine the source of its output.

2.3. Where we are now: ChatGPT

Most of OpenAI's current state-of-the-art models are direct descendants of GPT-3,[2] or more specifically, of InstructGPT (Ouyang et al. 2022). What differentiates InstructGPT from the initial version of GPT-3 is mainly two new phases of training called instruction tuning and reinforcement learning from human feedback.

Instruction tuning is an extension of pre-training in which the model is trained on a dataset consisting of a set of instructions (such as writing prompts or math problems) and answers that satisfy those instructions. The motivation behind this training is that it aligns the model to its likely downstream usage: users will prompt the model with a description of the output that they want, with the expectation that the model will provide a response conforming to their specifications (ibid). To train InstructGPT, OpenAI collected a set of instructions and answers from human labelers, including users of GPT-3 and paid contractors (ibid). The resulting dataset has not been released. Reinforcement Learning from Human Feedback (RLHF) is a stage of training following pre-training and instruction tuning, which may even continue once the model is deployed (as is the case with ChatGPT (OpenAI 2022)). In this phase of training, human judges are presented with multiple model outputs for the same prompt and asked to rank them in order of quality (ibid). After being trained on this feedback, the model is more likely to produce output similar to the higher rated examples.

The motivation for using RLHF is what OpenAI has described as their aim "to make artificial general intelligence (AGI) aligned with human values and follow human intent" (Leike

[2] The exception to this is GPT-4, which is generally assumed to have far more parameters than than the GPT-3 class. The exact number of parameters has not been released, but OpenAI CEO Sam Altman has strongly suggested that it has fewer than 100 trillion (Vincent 2023).

2022). In practice, Ouyang et al. accomplish this by "[having] labelers evaluate whether an output is inappropriate in the context of a customer assistant, denigrates a protected class, or contains sexual or violent content" (2022, 10).

2.4. Differences Between Human Language Production and LLM Text Generation

ChatGPT and its ilk are undoubtedly impressive technological feats. After a brief interaction with OpenAI's chatbot, many users are surprised by its apparent mastery of the English language. However, Bender and Gebru et al. (2021) provide a cautionary reminder for interpreting LLM-generated text: "coherence [is] in the eye of the beholder" (616). LLMs might *appear* to understand human language and produce meaningful output in return, but that meaning is not created by the LLMs themselves—it is created by their human interlocutors (Bender and Koller 2020).

Human communication "takes place between individuals who share common ground and are mutually aware of that sharing (and its extent), who have communicative intents which they use language to convey, and who model each others' mental states as they communicate" (Bender and Gebru et al., 2021, 616). Language models, having no experience of the world beyond the tokens in their training data, do not share common ground with their human users, nor do they have communicative intent or mental states. When a person reads text generated by an LLM, it may seem as though there is thought and intent behind the response. This is not the case, as it simply isn't possible for an LLM to have thought or intent, and the illusion of communication comes from our own human linguistic capabilities: "our perception of natural language text, regardless of how it was generated, is mediated by our own linguistic competence and our predisposition to interpret communicative acts as conveying coherent meaning and intent, whether or not they do" (ibid, 616).

3. Authorship Theory in KO

LLMs process trillions of forms and learn to recognize statistically significant patterns in their usage, but this is not the same thing as understanding their meaning (Saussure 1959). Just as a copy of *Wuthering Heights* is a representation of Emily Brontë's work and not the work itself, the forms used to pre-train an LLM are not intrinsically meaningful.

3.1. The Conceptual Structure of A Work

Smiraglia explains that "works are core narratives in every part of human experience—from sacred texts to legal foundations to iconic structures to iconic novels" (2019, 311). While we tend to engage with these "mentefacts" (Gnoli 2018), or mental constructs, through physical artifacts, "a work is abstract at every level, from its creator's conception of it, to its reception and inherence by its consumers" (Smiraglia 2019, 310). From an information retrieval perspective, these conceptual problems are typically circumvented through a fixation on the item-level object.

Hypothetically speaking, identifying the title of a specific bibliographic object is a straightforward activity. From there, assigning the author should be similarly easy.

While nice in theory, there are at least two factors that complicate the description of linear author-work association.

1. Different manifestations of a particular work can exhibit significant deviations from the original expression.
2. Since works are abstract concepts, determining where the boundary of one ends and another begins is a complex perceptual activity.

To the first point, take *Wuthering Heights*. According to *Resource Description and Access* (RDA) guidelines, all versions of the novel are to be collocated under the same nominal authorized access point (AAP) associated with the original manuscript: Emily Brontë. Editions published in 1848 and 1948 will likely have different covers and exhibit cosmetic editorial differences, but, by and large, few would deny both are versions of the same work. But how much can be changed before the item is no longer *Wuthering Heights*? For example, under the entity-relationship model at the core of the *Functional Requirements for Bibliographic Records* (FRBR), translations of a work should be primarily associated with the original author. This means a Hebrew edition of the text will be attributed to Brontë even though, at the time she penned her novel, the language had yet to be revitalized and adapted for general use.

On the one hand, this Hebrew translation will hopefully preserve the abstract work concept intended by Brontë; as such, her creative labor deserves recognition. On the other hand, using her name as the primary AAP "inevitably devalues the role of the translator and ignores the creative license and labor required in the translation process" (Soos and Leazer 2020, 486). Translating is not a one-to-one process in which individual words are simply swapped for identical ones of another language. A talented translator will exhibit fluency in the source and target languages, possess a deep knowledge of the particular work, and utilize various linguistic tools to articulate its essence. So while the goal is the maintain both the semantic and affective qualities evoked by Brontë, a translator's unique choices can severely alter a reader's experience of her work.

To the second point, as abstract concepts, works are subject to the same influences and factors that impact all perceptive activities. While writing her book, Brontë built on her own experiences—the things she read, the people she knew, the social context in which she lived—to create something new. Similarly, when a person is reading *Wuthering Heights*, their understanding of her text is inevitably influenced by their own unique experiences, and, having now interacted with Brontë's work, it is difficult to know how her ideas might impact their own creative production. In some kind of authorial butterfly effect, composer Jim Steinman might not have written "It's All Coming Back to Me Now" (popularized by Celine Dion) having not been inspired

by his own reading of *Wuthering Heights.* Still, his song is unanimously viewed as a distinct work external to the original text.

Within the FRBR model, the concept of a "super work" (Svenonius 2009, 38) seeks to situate derivative works, like Steinman's, as "ideational nodes within the set" (Smiraglia 2019, 313). An influential "progenitor work" (Smiraglia 2007, 182), such as *Wuthering Heights,* can be viewed as a primary connective node within an "instantiation" (ibid) or "textual identity" (Leazer and Furner 1999) network, but it is intentionally positioned adjacent to the works it inspired. Yet even within these more robust webs of relationships, there exists the problem of determining where one work ends and another begins. Smiraglia arguably resolves this issue with his definition of a work, which he defines as "a deliberately created informing entity intended for communication" (2019, 308). "Deliberately" is the key term here, as the creator's intention to produce something distinct from the progenitor marks the beginning of a new work entity. This is an view supported in other disciplines, where creative genres like the readymade and the parody use a person's ideation and intentionality to distinguish influence from theft.

3.2. Influence and Intention

Quests for originality and authenticity can be equally liberatory as they are oppressive. While there is undeniable value in personal expression, pressures that tie a person's worth—be it economic, professional, or social—to the originality of their creative output forces them to view their peers as competition rather than collaborators. In *The Anxiety of Influence*, Bloom argues that writers are both limited and motivated by this desire to distinguish themselves from their predecessors,

> For the poet is condemned to learn his profoundest yearnings through an awareness of other selves. The poem is within him, yet he experiences the shame and splendor of being found by poems—great poems—outside him. To lose freedom in this center is never to forgive, and to learn the dread of threatened autonomy forever. (Bloom 1997, 26)

In an act of *kenosis*, the author seeks "discontinuity with the precursor" (ibid, 14), a response that paradoxically concedes power to the other's influence. Moving away from something is as much a response as moving towards, and in rejecting the progenitor work a writer simply reaffirms their place within the creative continuum.

Although Bloom constructed his theory around poetic networks, the anxiety of influence transcends genre and medium to gesture towards a broader humanistic desire for self-actualization. While this tendency is not inherently bad, the judgment of a work based on its intellectual purity

sets a standard of originality almost impossible to achieve. Authors think and create surrounded by the works of others, not within sterile incubators free from outside influence. So when the ultimate test of intellectual autonomy rests upon someone's ability to produce innovative work—poetic or otherwise—completely detached from the works of others, anxiety is an entirely reasonable response to this unachievable expectation.

Building from Bloom and Foucault (1977), Soos and Leazer suggest that the author "as a lone and entirely detached figure simply does not exist," arguing instead that "the complex nature of intellectual and creative production makes it impossible to draw a clear and distinct boundary around a particular work and attribute it to one unique individual" (2020, 487). Rather than viewing authors as "owners" of an idea, they suggest that an "author-as-node" approach better preserves the inherently collaborative nature of creative production. Just as work-based instantiation networks connect individual items through a unifying progenitor node, this model positions an author as a singular entity within a sea of influential relationships.

That being said, even Bloom rejected the claim that "no one ever had or ever will have a self of his or her own" as nothing more than an "unamiable fiction" (1997, xlvi). Yes, works are created within complex intellectual ecosystems, but, as individuals, the people that produce them have unique perspectives and talents worthy of recognition. To borrow Smiraglia's word, they have intentionality.

Any KO theory of authorship inevitably reaches a seemingly contradictory impasse: people are unique individuals with unique ideas and unique intentions—and, at the same time, they are products of their environments. Authors are influenced by those who came before them, the people who inspire them, and the community that cares for them, but they also offer an essential quality that only they can provide. While nuanced discourse can simultaneously hold the importance of relationality (Littletree, Belarde-Lewis, and Duarte 2020) and individuality, notions of authorship conveyed through standard ontological frameworks generally fail to capture this duality. FRBR extends authorship beyond individual persons to include families and corporate bodies, and the replacement of "author" with "contributor" in RDA perhaps better gestures to the expansive nature of the work creation. However, the use of standardized AAPs in author attribution still removes a person, family, or corporate body from their broader context. In doing so, we are essentially suggesting that influence is secondary to the intention it yields.

Although epistemically valuable, these influence networks are often too complex and messy to visually represent through a basic KOS. At the end of the day, a student probably just wants to find *Wuthering Heights* in the stacks and finish their assignment, and they will likely do so by searching for "Emily Brontë," not "Jim Steinman." These authorial networks might help the user contextualize Brontë's work, but this is not typically the primary goal of most catalogs.

Yet while presenting authors as "owners" of a work is the reasonable choice given user-warranted practices, doing so defends particular ontological commitments that hide the social, cultural, economic, and professional "complexities that affect the production of new objects and

ideas" (Soos and Leazer 2020, 486). And the consequences of these decisions extend far beyond any one user's search query.

4. The Authorship of LLM Content

Most universities have some kind of academic integrity policy. Cheating and other forms of academic dishonesty are of primary concern, with plagiarism being one of the most vehemently condemned. Learning to find, interpret, and cite sources are core skills needed for academic success, and plagiarism—a spectrum of actions that ranges from an uncited paraphrase to the wholesale appropriation of another student's writing—is largely viewed as antithetical to the ethos of the academy.

Plagiarism occurs when "somebody presents the work of others (data, words or theories) as if they were his/her own and without proper acknowledgment" (Wager and Kleinert 2012, 167). Under the authorship concepts defended by RDA and FRBR, avoiding accusations of plagiarism appears to be a straightforward task: you only need to indicate when you are referring to another person's work and never suggest their ideas are your own. Simple enough. We can debate the conceptual boundaries of works and authors, but, using the attribution protocols generally accepted across higher education, plagiarism is framed as an avoidable issue.

The broader adoption of generative AI has revealed the limitations of this approach. Following the relatively quick adoption of ChatGPT by students and staff, many institutions formally declared the use of pre-trained language models to produce or enhance one's work to be a violation of academic integrity. Based on the above definition, asking ChatGPT to write your *Wuthering Heights* essay seems to be a clear-cut case of plagiarism; the student did not produce the content and is presenting it "as if they were his/her own and without proper acknowledgment." But who, or what, is being plagiarized?

4.1. Communicative Intent and Work Creation

OpenAI has done a wonderful job of developing an application that appears to possess so-called "general intelligence." But, as previously noted, while ChatGPT's "human-like" responses can be quite convincing, the chatbot does not understand what it is saying—at least not in the typical sense in which people use those words. It also does not answer user queries in an intentional act of communication—again, at least not in the way implied by such a claim.

That this lack of "communicative intent" (Bender and Gebru et al. 2021) marks the fundamental distinction between the way humans and LLMs utilize language. Within the context of Smiraglia's definition, this inability to experience or express intention essentially disqualifies ChatGPT from being able to produce a work. So, although a language model is capable of producing information, it cannot make a work. Absent a work, it cannot be a victim of plagiarism.

4.2. User Queries & Feedback

The model itself may be incapable of intentional action, but there are myriad other associated parties who are. The most obvious is the accused student.

For all intent and purpose, there is nothing technically preventing this person from being named the creator of the *Wuthering Heights* essay. Entering a query into ChatGPT, copying the text into a new document, adding their name, and submitting the file are all intentional acts focused on recording and expressing a particular viewpoint. Sure, the student did not fabricate a majority of the text, but the essay was deliberately created using their actions, knowledge, and capabilities.

4.3. Training Data

This appears to be a victimless crime until one considers the broader context. The plethora of data used to train an LLM directly supports the parameters it uses to generate new responses. ChatGPT may be incapable of "understanding," but the millions of authors responsible for its immense training set probably are. Although they did not personally write the exact words used in this exact essay, the collective can be viewed as a "family or corporate body" responsible for this immense network of data. Following this logic, one could argue that the generated essay paraphrases this corpus of material, making the members of this family/corporate body targets of plagiarism.

Well, it's an answer. But, as Dehouche argues, an accusation of plagiarism "appears rather inadequate when the 'others' in question consist in an astronomical number of authors, whose work was combined and reformulated in unique ways" (2021, 21). While those individuals intentionally created the material that was used to train ChatGPT, and while they offer a wonderful metaphor for how textual identity networks function, OpenAI was actually the one that developed the GPT model that made the *Wuthering Heights* essay possible.

In an interesting turn, OpenAI can now either be viewed as the victim of plagiarism (by the student) or a perpetrator (towards the family/corporate body). Both the code used to create ChatGPT and the parametric memory defined during its training are proprietary works intentionally created by those at OpenAI[3]. As the student failed to cite either, that can be viewed as an act of plagiarism. At the same time, ChatGPT's parametric memory was constructed from billions of other works that cannot be cited. Whether that memory constitutes a work on its own is another matter entirely.

5. Plagiarism Revisited

All creative acts are forms of collaboration. New ideas and works develop within a broader social context that directly and indirectly contributes to their production, and any single author is but one

[3] Though, as we describe in Section 2.3, the datasets used to train ChatGPT–the sources of its parametric memory–largely are not the sole property of OpenAI.

node in a vast network of influence. The ambiguous boundaries between specific authors and works are further eroded by the innately diffused nature of LLMs.

Our failure to accommodate this generative content within preexisting notions of plagiarism reveals the conceptual limitations of an author-as-owner approach and highlights the importance of networked attribution. "Plagiarism" is a semantic category that allows for varying degrees of membership. Its prototypical examples—for example, paying another person to write your college assignment—generally support the validity of linear work-author relationships and, therefore, reaffirm the validity of the class. However, the "internal structure" of this category (Rosch 1975) is much more stratified than standard use of the term suggests. The ambiguous nature of LLM-generated works just presents a more obvious challenge to the seemingly stable concept.

While fooder for an interesting philosophical discussion, we think debating whether ChatGPT's *Wuthering Heights* essay is an example of plagiarism—or a component of a bigger plagiarism racket—largely circumvents and obscures the actual issue. When real humans are being obviously plagiarized, holding the culprit responsible is often viewed as a way of minimizing the harm caused to this other party. But when the "other" is unidentifiable, what harm is being caused? Why are so many people upset by the thought of a student getting an "A" on an essay produced by an LLM?

Plagiarism is perhaps best viewed as an attempt to standardize a prescriptive claim about intellectual morality. In higher education, "plagiarism evokes deeply held emotions related to deviance, credibility, and what it means to be outside the norm" (Rooksby quoted in McCarthy 2023, 4). So even when a particular victim may be difficult to identify, submitting an essay you did not write undermines the core tenets of an academic meritocracy: you should be assessed based on what you know and how well you can articulate that knowledge. Yet this protective barrier around "what you know" is deceptively precarious. Removed from the author-as-owner paradigm, the concept is nearly impossible to enforce.

To be clear, this claim is not intended to defend violations of student conduct codes or refute the importance of intellectual honesty. Quite the contrary. Simply asking people to not "steal" the "property" of others (i.e. their works and ideas) is a low bar that prevents us from having deeper discussions about what it means to think and live in relation to others. We should demand more of those within a learning community, and reconsidering our views of solitary authors with wholly distinct ideas provides an opportunity to explicitly acknowledge our reliance on one another. When work production is reframed as a community activity rather than the mark of independent genius, the harm of plagiarism is no longer reduced to a localized interpersonal event.

6. Conclusion: Reframing Accountability

The moralization of technology is rarely beneficial. ChatGPT is a powerful tool with a number of promising pedagogical uses—at the same time, it can also facilitate non-learning and perpetuate harmful educational practices. The matriculation of LLM tools into different domains will continue to reveal possible benefits and risks, and our affective responses to these mis/applications should be viewed as indicators of unfulfilled values.

Upset over the "plagiarizing" of ChatGPT's content suggests a yearning for individual responsibility and community accountability. Authorship is used to "confer credit" for a job well done, but the connection between individuals and ideas additionally ensures "authors understand their role in taking responsibility and being accountable for what is published" (ICMJE 2023). The opaque webs of influence central to LLM writing tools complicate our ability to assign responsibility and, consequently, challenge what it means to be held accountable to both oneself and one's community. As the discussions prompted by these new technologies lead us to reflect on our values, we are provided with a meaningful opportunity to reaffirm those congruent with our views and replace the ones that no longer serve us.

References

Bandy, John, and Nicholas Vincent. 2021. "Addressing 'Documentation Debt' in Machine Learning: A Retrospective Datasheet for BookCorpus." In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* 1.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*: 610-623. Online: ACM.

Bender, Emily M. and Alexander Koller. 2020. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–98. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.463.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. "A Neural Probabilistic Language Model." *Journal of Machine Learning Research* 3: 1137-1155.

Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe. "Multimodal datasets: misogyny, pornography, and malignant stereotypes." Preprint, submitted October 2021. https://arxiv.org/abs/2110.01963.

Bloom, Harold. 1997. *The Anxiety of Influence*. 2nd ed. Oxford: Oxford University Press.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal,

Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. "Language Models are Few-Shot Learners." Preprint, submitted July 2020. https://arxiv.org/abs/2005.14165.

Common Crawl. N.d. "About." https://commoncrawl.org/about/.

Dehouche, Nassim. 2021. "Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3)." *Ethics In Science And Environmental Politics* 21: 17-23. doi: 10.3354/esep00195.

Dey, Sneha. 2021. "Reports Of Cheating At Colleges Soar During The Pandemic." NPR. https://www.npr.org/2021/08/27/1031255390/reports-of-cheating-at-colleges-soar-during-the-pandemic.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–86. Minneapolis, Minnesota: Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423.

Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. "Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 1286–1305. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021. doi:10.18653/v1/2021.emnlp-main.98.

Dolan, Jill. 2023. "Guidance on AI/ChatGPT: Memo to All Teaching Faculty - January 25, 2023." The McGraw Center for Teaching and Learning, Princeton University. https://mcgraw.princeton.edu/guidance-aichatgpt.

Elias, Jennifer. 2023. "Google execs tell employees in testy all-hands meeting that Bard A.I. isn't just about search." CNBC. https://www.cnbc.com/2023/03/03/google-execs-say-in-all-hands-meeting-bard-ai-isnt-all-for-search-.html.

Firth, John Rupert. "A Synopsis of Linguistic Theory, 1930-1955." In *Studies in Linguistic Analysis*, 1-31. Oxford: Blackwell, 1957.

Foucault, Michel. 1977. "What Is an Author?" In *Language, Counter-Memory, Practice: Selected Essays and Interviews*, ed. Donald F. Bouchard. Ithaca: Cornell University Press, 113-38.

Gehman, Samuel, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 'RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models'. 2020. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 3356–

69. Online: Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.301.

Gnoli, Claudio. 2018. "Mentefacts as a missing level in theory of information science." *Journal of Documentation* 74(6): 1226-1242. doi: 10.1108/JD-04-2018-0054.

Harris, Zellig. 1954. "Distributional Structure." In *WORD* 10(2-3): 146-162. doi: 10.1080/00437956.1954.11659520.

International Committee of Medical Journal Editors (ICMJE). 2023. "Defining the Role of Authors and Contributors: Why Authorship Matters." https://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html.

Jenkins, Baylee D., Jonathan M. Golding, Alexis M. Le Grand, Mary M. Levi, and Andrea M. Pals. 2022. When Opportunity Knocks: College Students' Cheating Amid the COVID-19 Pandemic. *Teaching of Psychology* 0(0). doi: 10.1177/00986283211059067.

Jurafsky, Dan and James H. Martin. 2023. "Neural Networks and Neural Language Models." In *Speech and Language Processing*, 3rd edition. Preprint, submitted January 2023. https://web.stanford.edu/~jurafsky/slp3/.

Kasneci, Enkelejda, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stepha Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. "ChatGPT for good? On opportunities and challenges of large language models for education." *Learning and Individual Differences* 103. doi: 10.1016/j.lindif.2023.102274.

Leazer, Gregory H. and Jonathan Furner. 1999. "Topological Indices of Textual Identity Networks." In *Proceedings of the 62nd Annual Meeting of the American Society for Information Science*, 1999, ed. L. Woods. Medford, NJ: Information Today, 345-58.

Leike, Jan, John Schulman, and Jeffrey Wu. 2022. "Our approach to alignment research." OpenAi. https://openai.com/blog/our-approach-to-alignment-research.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. "BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–80. Online: Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.703.

Littletree, Sandra, Miranda Belarde-Lewis, and Marisa Duarte. 2020. "Centering Relationality: A Conceptual Model to Advance Indigenous Knowledge Organization Practices." *Knowledge Organization* 47(5): 410-426. doi:10.5771/0943-7444-2020-5-410.

Lock, Samantha. 2022. “What is AI chatbot phenomenon ChatGPT and could it replace humans?” *The Guardian*. https://www.theguardian.com/technology/2022/dec/05/what-is-ai-chatbot-phenomenon-chatgpt-and-could-it-replace-humans.

McCarthy, Claudine. 2023. “ChatGPT use could change views on academic misconduct.” *Dean & Provost* 24(10): 1-4. doi: 10.1002/dap.31202.

Mikolov, Tomas, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. “Efficient Estimation of Word Representations in Vector Space.” *International Conference on Learning Representations,* 2013.

Nagel, Sebastian. 2023. “March/April 2023 crawl archive now available.” Common Crawl. https://commoncrawl.org/2023/04/mar-apr-2023-crawl-archive-now-available/.

OpenAI. 2022. “Introducing ChatGPT.” https://openai.com/blog/chatgpt.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. “Training language models to follow instructions with human feedback.” Preprint, submitted March 2022. https://arxiv.org/abs/2203.02155.

Paullada, Amandalynne, Inioluwa Deborah Raji, Emily M. Bender, Emily L. Denton and A. Hanna. 2020. “Data and its (dis)contents: A survey of dataset development and use in machine learning research.” *Patterns* 2(11). doi: 10.1016/j.patter.2021.100336.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. “Language Models are Unsupervised Multitask Learners.” Technical report, OpenAI.

Raffel, Colin, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.” Preprint, submitted July 2020. https://arxiv.org/abs/1910.10683.

Rosch, Eleanor. 1975. “Cognitive Representations of Semantic Categories.” *Journal of Experimental Psychology: General* 104(3): 192-233.

de Saussure, Ferdinand. 1959. *Course in General Linguistics*. Translated by Wade Baskin. New York: The Philosophical Society.

Smiraglia, Richard P. 2007. “The ‘Works’ Phenomenon and Best Selling Books.” *Cataloging & Classification Quarterly* 44(3-4): 179-195. doi: 10.1300/J104v44n03_02.

Smiraglia, Richard P. 2019. “Work.” *Knowledge Organization* 46(4): 308-319. doi: 10.5771/0943-7444-2019-4-308.

Soos, Carlin and Gregory H. Leazer. 2020. “Presentations of Authorship in Knowledge Organization” *Knowledge Organization* 47(6): 486-500. doi: 10.5771/0943-7444-2020-6-486.

Svenonius, Elaine. 2009. *The Intellectual Foundations of Information Organization*. Cambridge, MA: MIT Press.

University of California, Los Angeles (UCLA). 2023. "ChatGPT and AI: Starting Points for Discussion." Online Teaching & Learning. https://online.ucla.edu/chatgpt-ai/.

University of Washington. 2023. "ChatGPT and other AI-based tools." Center for Teaching and Learning. https://teaching.washington.edu/topics/preparing-to-teach/academic-integrity/chatgpt/.

University of Wisconsin-Madison. 2023. "Considerations for Using AI in the Classroom." L&S Instructional Design Collaborative. https://idc.ls.wisc.edu/guides/using-artificial-intelligence-in-the-classroom/.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Ł. Ukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." In *Advances in Neural Information Processing Systems* 30.

Vincent, James. 2023. "OpenAI CEO Sam Altman on GPT-4: 'people are begging to be disappointed and they will be.'" The Verge. https://www.theverge.com/23560328/openai-gpt-4-rumor-release-date-sam-altman-interview.

Wager E, Kleinert S. "Cooperation Between Research Institutions And Journals On Research Integrity Cases: Guidance From The Committee On Publication Ethics (COPE)." *ACTA Informatica Medica* 20(3):136-40. doi: 10.5455/aim.2012.20.136-140.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. "XLNet: Generalized Autoregressive Pretraining for Language Understanding." In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY: 2019.

Zhang, Zhengyan, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. "ERNIE: Enhanced Language Representation with Informative Entities." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1441–51. Florence, Italy: Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1139.

Zhu, Yukun, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba and Sanja Fidler. 2015. "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books." *2015 IEEE International Conference on Computer Vision (ICCV)*: 19-27.