The Author was Killed by an AI - or Was It?

Tereza Matějková[†], Pavel Ircing[*,†]

University of West Bohemia, Univerzitní 8, 301 00 Plzeň, Czech Republic

Abstract

This short opinion paper looks at the usage of large language models as writing assistants through the lenses of Roland Barthes's work. It discusses the advantages and disadvantages of the different approaches to authorship in the coming era of massive usage of AI models, especially with respect to the human responsibilities and merits in the authorship process.

Keywords

large language models, authorship, Roland Barthes

1. Introduction

Roland Barthes in his famous essay *The Death of the Author* [1] promotes the idea that that the interpretation of a text should not be limited by the intentions, creativity or even a biography of the author. Instead, the meaning of a text is generated solely through the interaction between the reader and the text itself. In other words, the author is no longer important, the importance is shifted almost entirely to the reader. Barthes also replaces the person of the author with a *scriptor*, a kind of mediator through which language itself acts. In the *scriptor*, pre-existing texts are mixed, and he only imitates previous gestures of writing, but he is never original. From these texts, the *scriptor* forms an immense vocabulary from which he draws when writing, and no longer carries passions, moods, feelings, or impressions. If we look at one of the often quoted passages:

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

we can be tempted to see the very principle of the way the large language models (LLMs) work in these sentences and directly attribute the role of the *scriptor* to the individual LLM.

The possible use of these models for the automatic generation of literary works also seems to confirm the overall tone of Barthes's essay - namely that the person of the author should at least be considerably marginalized when reading the text, perhaps even completely eliminated, since the text generated by LLMs is really an extremely sophisticated but still "just" a permutation of the texts that were originally produced by many human authors. On top of that, people also usually don't really attribute the authorship of the passages of text to the AI systems, no matter how sophisticated or substantial for the final work they are [2].

The relation between the decades old essay and the most modern text generation technology seems so apparent the even one of the first literary works experimentally created with extensive use of LLM is named *Death of an Author: A Novella* [3].

In our work, we abstract from the fact that it is not certain whether Barthes's famous text should be understood literally or whether it should considered to be - at least partly - an intellectual provocation,

LaCATODA 2024 – The 9th Linguistic and Cognitive Approaches to Dialog Agents Workshop, November 19, 2024

[*]Corresponding author.

[†]These authors contributed equally.

matejkot@ff.zcu.cz (T. Matějková); ircing@kky.zcu.cz (P. Ircing)

0009-0005-8875-6295 (T. Matějková); 0000-0001-6967-1687 (P. Ircing)

and we will try to offer our view on how much are the ideas presented in the essay relevant to the various aspect of authorship in the light of the recent advances of the large generative models.

2. The Text as a Dialogue

First, we want to address the issue of literary communication. The author and the reader are two of the three most important elements in it. However, this communication is specific in that it is usually one-sided, either on the part of the author or on the part of the reader. To use Umberto Eco's theory and terminology from his book Six Walks in the Fictional Woods [4], the author creates a model reader in his/her mind when writing a text. The model reader is not a concrete physical person, but only a theoretical concept. In the same way, then, the real reader of the text creates a model author, but again this is only a theoretical concept and not the physical person of the actual author. Because literary communication is not immediate and takes place through the text, it is necessary to distinguish model readers and authors from real ones. We agree with saying that the author tries to put his own intention into his text, he wants to communicate something, but there is nothing that guarantees a connection between what the author wanted to say and the meaning of the text [5].

The same is true on the side of the reader. This is also the reason why texts have more than one interpretation. We agree with Barthes that although the person of the author/scriptor is important for the creation of the work, the author's intention could be completely irrelevant to determining the meaning - i.e., the interpretation - of the work. This interpretation is dependent entirely on the reader. However, it is obvious – and yet important to point out – that if the meaning that is found in the text by the majority of readers is completely different from the meaning that the author *intended* to put there, the communication between the author and the reader failed. It is because of the difference between the code of the author and the code of the reader. The term „code" refers to the system or framework of symbols, meanings, and interpretations the author and the reader use to communicate. The author's/reader's code is always influenced by personal experiences, knowledge, and cultural context. Because these can never be identical, the encoding key of the author and the decoding key of the reader also can never be identical. Simply put, sometimes we understand the other person's code easier and sometimes harder.

In the vein of the previous paragraphs, we feel that the Emily Bender and her co-authors in their often cited paper [6] are pushing at an open door when they state that "*coherence (of the text is) in the eye of the beholder*" since it is a well-established fact. On the other hand, we would like to challenge their other claim that the text generated by LLM is not grounded in the communication intent but we will postpone this discussion to the Section 4.

3. The Author-God

The concept of authorship as seen today was born in the Enlightenment period [7] and is closely related with the overall tendency for individualization that was very strong in that era [8]. Ever since then, the fiction writers and the academic authors in the humanities (especially philosophers) write predominantly as individuals, often strongly presenting their personal opinion. Over the years, this naturally led to a certain "celebritization" or even "deification" of the author and his/her persona became almost as important as the actual text he or she has produced. But thanks to Barthes (and others of course) and the popularity of his concept of the "death of the author", we can distinguish the personality and opinions of the authors of fictional texts from the meanings of the texts themselves. A current example of this is the large number of fans of the Harry Potter book series who still enjoy the stories, but separate them from the author of the books, J.K. Rowling, because of her views on gender and changes in her own interpretation of Harry Potter. So it's clear that the personality of the author may no longer be "the God" that will dictate the one and only correct meaning.

However, it is important to point out that this "author-centrism" of course has not spread to all disciplines and also seems to be in regress anyway. In the STEM academic writing, the collaborative

authorship was standard in the recent decades and in the extreme cases, the papers can have up to tens or even hundreds of authors. The persona of a single author can hardly make its way into the resulting text.

If we now consider non-fiction texts and scholarly texts, we have to acknowledge that in the contemporary sciences, regardless of whether they are humanities or STEM, great care is taken to ensure that any conclusion is supported by good arguments, sources, or measurements. Readers of such texts then do not ask the authors how they thought about their conclusions, but evaluate and critique the arguments that support the conclusions. They look for counter-arguments, verify the measurements, and publish their own new conclusions in turn. Regardless of whether a paper has one or hundreds of authors, it is common practice in peer-reviewed journals and at many conferences (such as the one we are submitting our paper to) to hide the identity of the authors in the double-blind review so that the reviewers are not influenced by the past achievements and/or the community status of the authors and judge the quality of the paper just by the quality of the text itself.

The paragraphs above may seem to state the obvious but the concept of the (solo) author is so central in the Barthes's essay that we felt it important to explicitly deal with it in our text.

4. The LLM: a Co-Author or a Writing Assistant?

Co-authorship or even collective [9] writing is a standard practice in at least some of the domains and it is quite natural to accept that LLM can be used (and perceived) as yet another co-author. However, it turns out that it is better to use the term *writing assistant* instead. This subtle and seemingly only formal (the LLM of course operates in the same way regardless of the name you ascribe to it) change of framing substantially eases the debate about the role of the LLM in the writing process.

First of all, it is generally accepted that the author has not only the *skills* in putting together words and sentences that result into a good text but that he/she also has the *reason* why has has decided to create the text in the first place. In other words, there are *author's intentions* behind the text. It is quite understandable that many people have a problem attributing *intentions* to the language model. Actually, the author of [10] provides both the theoretical insight and the experimental evidence that LLMs are able to efficiently *encode* the multitude of communicative intentions of the authors of texts that were used for the LLM training. Here we go back to the claim made by Bender et al. that

> Text generated by an LM is not grounded in communicative intent, any model of the world, or any model of the reader's state of mind. [6]

and – supported by [10] - we state that this claim is not valid as even though the LLM cannot be thought as actually *having* a communicative intent, the communication is grounded in the communicative intents of the human writers encoded in the LLM. Surely also the LLM itself *is* the model of the world (at least the fragments of the world as captured by the training texts) and it also somehow encodes the fictional model readers (cf. Section 2) that the training texts' writers had in mind.

On the other hand, it's readily acceptable that an assistant – as opposed to an author – does not have much intentions by itself. If we take it to the extreme, its main purpose is to please the person it assists. And this – the human satisfaction – is exactly the reward function the ChatGPT models are optimized for [11].

When employing the LLM as a writing assistant, the net of participants in the literary communication (again cf. Section 2) broadens – there is a "short loop" where the human author engages in the dialogue with an LLM with a purpose of creating a better text. It is not a peer-to-peer dialogue – the human is actually using prompts and gradually refined instructions to push the LLM into the appropriate location in the "multidimensional space" (cf. Section 1) where the style of the LLM outputs is most suitable for the authorial intent. The big advantage of the LLM as the writing aid is that it is actually not a single writing assistant but a composite of many "writers" with varying perspectives, styles, and expertise. And with appropriate prompting, you can "summon" a *scriptor* that is best for the occasion.

The "long loop" of (indirect) dialogue between the writer and the reader remains the same as before.

5. The Pros and Cons of Denying the LLM Authorship

The positive aspects of "degrading" LLMs to mere assistive tools are actually discussed in the previous section. To sum them up – this shift relieves us from the necessity of attributing real or simulated intentions to the non-human agents and also at least partially puts aside the issue of judging the creativity such as in [12]. Viewing LLMs as tools also simplifies the process of understanding their role in content creation, maintaining the traditional boundaries between human and machine agency.

Now on the negative side. In Section 4, we wrote that the true author needs to have the writing skills and the intentions in order to write a text piece. What we did not mention is that authorship in our society also includes the *responsibility* for the published text and the *eligibility for rewards* potentially stemming from the work.

If the LLM is just a writing assistant as argued above, the responsibility part is easy – there is still a human author who is responsible for any impact that the text might have. But if we are to credit (co)-authorship to the LLMs, we open a Pandora's box of ethical issues akin to the ones that are being dealt with in connection with the autonomous cars (see e.g. [13] and many others). For example, how do we handle scenarios where AI-generated content causes harm, spreads misinformations, or violates ethical standards? How could LLMs be held accountable for such outcomes? They probably could not, so we would need to find a way how to balance the ethical responsibility between human operators, users, developers and any other stakeholders.

What we see as more pressing is the issue of copyright and connected financial reward and academic (or any other) merits for the authorial work. It is again convenient to dismiss these concerns with reference to a "just-an-assistive-tool" status of the LLM but it this case it is very unfair to the authors of the original texts used for the LLM training. The paper [14] provides a good argumentation for granting a *ghostwriter* status to the LLM but this approach still misses an important point – while the human ghostwriter is not granted the copyright and usually also does receive any credit for their work, they are of course compensated financially. But given the way LLMs are trained, it is of course impossible to trace exactly whose text – or writing style – had the greatest influence on the final output and who should therefore receive the largest portion of the reward. It is also crucial to point out that the majority of the human authors were not even asked for permission to use their work for the LLM training – and some of the are now trying to actually forbid this exploitation.

The European AI Act [1] that has recently come into force provides the authors a legislative help in this battle - it clearly states that:

> In order to increase transparency on the data that is used in the pre-training and training of general-purpose AI models, including text and data protected by copyright law, it is adequate that providers of such models draw up and make publicly available a sufficiently detailed summary of the content used for training the general-purpose AI model. While taking into due account the need to protect trade secrets and confidential business information, this summary should be generally comprehensive in its scope instead of technically detailed to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law, for example by listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used.

Thus the AI Act – if correctly and effectively implemented –should enable the authors at least to find out that their work was used for training the models and hopefully also give them the means to stop this usage. However, we are not sure whether there is an established framework for compensating authors financially in the case when they are willing to grant permission for the work to be used in training but just don't want to grant it for free.

If the LLM is actually regarded as a co-author, traditional models of copyright and authorship may be upended totally. In this case, how would royalties or recognition (not just ethical responsibility as

[1] https://eur-lex.europa.eu/eli/reg/2024/1689/oj

mentioned above) be divided between the machine, its developers, and the human user? Because of this ambiguity, it is hard to fit AI-generated works into existing intellectual property frameworks.

The practical consequences of granting or denying LLMs authorship stretch beyond just copyright or responsibility concerns. Denying LLM authorship reinforces the human-centric model of creativity and responsibility, preserving the integrity of authorship as a distinctly human activity. But if we start treating LLMs as authors or co-authors, we could end up undervaluing human creativity. Content and art creation might shift more toward automation and it might leave human writers and artists on the sidelines. This could shake up academic and professional fields where originality and human effort really matter.

On the other hand, LLMs are here and we can't shut down or ban them completely. So we need to find a way to learn how to work with them to use them as effectively as possible, and we will have to acknowledge the possibility that this may be accompanied by a change in the way these areas operate. This raises some critical questions: Will society continue to value human creativity in the same way? Will human authors struggle to maintain relevance in a world flooded with AI-generated works?

6. Conclusion

We would like to conclude that in many aspects the LLMs fit very accurately into the concept of the Barthesian *scriptor*. Yet we argue that both *scriptors* – the original theoretical one and the recent one actual constructed through the means of deep learning – need an *author* that, presently, still has a form of human being.

However, we argue that this concept of authorship should be at least partially redefined in the new era of generative models. We can actually again refer to Roland Barthes in this aspect since he sees the authorship as a *social construct* and as such, it can of course be re-constructed to fit the actual needs. Such reconstruction must of course also (and maybe foremostly) include the relevant intellectual property laws.

Acknowledgments

The work has been supported by the grant of the University of West Bohemia, project No. SGS-2024-023, and by the project No. LM2023062 LINDAT/CLARIAH-CZ from the Ministry of Education, Youth and Sports of the Czech Republic.

References

[1] R. Barthes, The death of the author, in: Image, Music, Text, Fontana, London, 1977, pp. 142–148.
[2] F. Draxler, A. Werner, F. Lehmann, M. Hoppe, A. Schmidt, D. Buschek, R. Welsch, The AI ghostwriter effect: When users do not perceive ownership of AI-generated text but self-declare as authors, ACM Trans. Comput.-Hum. Interact. 31 (2024). doi:10.1145/3637875.
[3] A. Marchine, S. Marche, Death of an Author: A Novella, Pushkin Industries, 2023.
[4] U. Eco, Six Walks in the Fictional Woods, Harvard University Press, 1994. doi:10.2307/j.ctvjhzps3.
[5] A. Compagnon, Le Démon de la Théorie: Littérature Et Sens Commun, Seuil, 1998. URL: https://books.google.cz/books?id=NexYAAAAMAAJ.
[6] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, 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, Association for Computing Machinery, New York, NY, USA, 2021, p. 610–623. doi:10.1145/3442188.3445922.
[7] C. Hesse, Enlightenment epistemology and the laws of authorship in revolutionary france, 1777-1793, Represetations 30 (1990) 109–137. doi:10.2307/2928448.

[8] M. Foucault, Authorship: What is an Author?, Screen 20 (1979) 13–34. doi:10.1093/screen/20.1.13.

[9] M. A. Peters, T. Besley, S. Arndt, Experimenting with academic subjectivity: collective writing, peer production and collective intelligence, Open Review of Educational Research 6 (2019) 26–40. doi:10.1080/23265507.2018.1557072.

[10] J. Andreas, Language models as agent models, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, 2022, pp. 5769–5779. doi:10.18653/v1/2022.findings-emnlp.423.

[11] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. F. Christiano, G. Irving, Fine-tuning language models from human preferences, CoRR (2019). doi:10.48550/arXiv.1909.08593.

[12] M. A. Runco, AI can only produce artificial creativity, Journal of Creativity 33 (2023). doi:10.1016/j.yjoc.2023.100063.

[13] C. J. Copp, J. J. Cabell, M. Kemmelmeier, Plenty of blame to go around: Attributions of responsibility in a fatal autonomous vehicle accident, Current Psychology 42 (2023) 6752–6767. doi:10.1007/s12144-021-01956-5.

[14] A. J. Nowak-Gruca, Could an artificial intelligence be a ghostwriter?, Journal of Intellectual Property Rights 27 (2022) 25–37. doi:42/jipr.v27i1.51259.