A Distant Technology? Experiments with a Generative Model for Retouching Noisy Newspaper OCR

Daniel Brodén[1], Lina Samuelsson[2], David Alfter[1], Johan Malmstedt[1,3]

[1] *University of Gothenburg, Box 200, Gothenburg, 405 30, Sweden*
[2] *Mälardalen University, Box 325, 631 06 Eskilstuna, Sweden*
[3] *Harvard University, MA 02138, Cambridge, USA*

Abstract

This paper explores the use of generative language models to enhance digitized historical newspaper text. While large language models offer new means of addressing noisy OCR, their opaque, probabilistic processes raise epistemological concerns. Within the project *The Order of Criticism Revisited*, which integrates literary and computational approaches to Swedish criticism, we tested GPT-4o to "retouch" OCR data from the National Library of Sweden using zero-shot prompting. Comparisons with flawed OCR outputs and manually transcribed texts show that the model produced more legible versions, often closer to the originals than the raw OCR. This indicates potential for improving the quality of digitized sources and enabling more robust large-scale analysis. At the same time, drawing on the notions of artificial communication and distant technology, we argue that such models extend analytical capacity while creating perceptual and methodological distance. Their outputs, better seen as probabilistic "retouching" than correction or reconstruction, weaken the indexical link to original sources.

Keywords

Generative models, OCR, digital epistemology

1. Introduction

Large language models and generative models, like GPT [1] are currently reshaping how researchers interact with source materials [2]. Yet, because these models rely on opaque computational processes, it is difficult to evaluate the reliability and validity of their outputs [3]. At the same time, within digital humanities, experimentation as a scholarly principle encourages reflective inquiry into the potential of digital tools and methods [4]. In this spirit, rather than seeking systematic methodological assessment, this paper reflects on both the capacity and the implications of using a generative model to enhance the data quality of familiar source material more broadly.

The research project *The Order of Criticism Revisited* (2020–2025) [5] explores how "traditional" literary scholarship intersects with computational methods, building on materials and results from an earlier study of literary criticism in Swedish press [6]. Within the project, we collected and annotated approximately 5,800 book reviews, primarily from the National Library of Sweden's (*Kungliga biblioteket*, hereafter KB) newspaper collections but digitization shortcomings, including optical character recognition (OCR), complicate their usefulness [7] [8] [9] These limitations highlight both the need for careful curation for research [10] and automated OCR mitigation of structural flaws [11].

Our inquiry is framed by a two-part question: To what extent can a generative language model make noisy OCR in digitized Swedish newspapers more usable, and what uncertainties arise in doing so? Specifically, we explore using GPT-4o for "generic" OCR cleaning through zero-shot prompting, directing the model to perform the task based on patterns inferred from its training data rather than through the more resource-intensive process of fine-tuning on a specific dataset. We chose GPT-4o as our primary model since, at the time of writing, it was among the most advanced and widely adopted, showing particular advantages in our tests (for comparison, we also used Claude 3.7 Sonnet but

Huminfra Conference 2025, Stockholm, 12-13 November 2025.

daniel.broden@lir.gu.se (D. Brodén); lina.samuelsson@mdu.se (L. Samuelsson); david.alfter@lir.gu.se (D. Alfter); johan.malmstedt@lir.gu.se (J. Malmstedt)

encountered limitations, as it failed to return complete outputs due to input-length restrictions). Given both the proprietary status of GPT-4o and copyright constraints of KB's collections, we limited our focus to reviews published in 1906.

To frame our inquiry, we first draw on the theoretical notions of artificial communication and distant technology to highlight how generative models, by virtue of their opaque and probabilistic nature, add a further layer of technological distance from research materials. We then discuss our experiment of applying GPT-4o to "retouch", rather than reconstruct, low-quality OCR in KB's collections. This design allows us to reflect on both the model's capacity to enhance text quality and the epistemological uncertainties that arise when generated outputs further weaken the indexical link to original sources. The paper concludes that although the model can substantially improve degraded OCR texts, its use simultaneously introduces a methodological distance that demands thorough reflection.

2. Digital Epistemology

Digital epistemology scholar Jonas Ingvarsson contends that digitization challenges the foundations of academic practice "not only by the appearance of new tools and objects, but by the fact that our modes of thought and our way to structure data and knowledge are changing" [12]. From this perspective, digital forms of expression provide scholars with different modes of engaging with research materials.

While data-intensive studies of literary criticism and newspaper collections tend to emphasize the analytical potential of computational methods rather than their epistemological implications [13] [14] [15], we have argued that text-data visualizations of previously analyzed book reviews can create a defamiliarizing effect, opening new perspectives about the material [16]. With generative models, however, a different kind of distancing effect emerges. These models involve forms of methodological uncertainty that partly differ from those encountered in the visualization techniques we earlier employed, such as TF-IDF analysis [16]. While GPT-4o generates text that resembles human writing and conveys a sense of familiarity with the material, its outputs remain products of probabilistic processes that scholars have described as largely opaque [17] [18].

2.1. Artificial Communication

At first glance, GPT-4o might seem to lessen the distance between researcher and data, thanks to its intuitive interface and human-like output, especially when compared with more conventional natural language processing (NLP) and "AI" techniques. Yet, as sociologist Elena Esposito points out: "what we can observe in interactions with [these] algorithms is not necessarily an artificial form of intelligence, but rather an artificial form of communication. Intelligence and communicative capacity are not the same thing" [19]. Esposito further argues that "algorithms learn not to think but to participate in communication, that is, to (artificially) develop an autonomous perspective that allows them to react appropriately and generate information in their interaction with other participants" [19].

Because generative models simulate the exchange and interpretation of information within linguistic distributions shaped by training data, their outputs should not be conflated with intelligence-based data processing [20]. In this, generative models partly create a different sort of distance from research results than traditional NLP methods. The latter may seem opaque to those without technical expertise, but they remain accessible and, to some extent, interpretable for those with the necessary knowledge. Generative models, such as GPT-4o, by contrast, mediate methodological distance for virtually all. While these models may display impressive capacity across a variety of research tasks, the computational processes that drive their responses are largely unreachable, even if new advances in interpretability and reverse engineering are beginning to provide insights into their internal workings [21].

2.2 Technological Distancing

On some level, this touches on psychologist and philosopher of technology Robert Romanyshyn's idea that modern technologies not only expand human capabilities but also reshape perception by distancing us from the objects of our attention [22]. Although Romanyshyn's account is primarily

concerned with the ways in which modern technology generates human detachment from the world (a point that remains debated), it more broadly highlights the ways in which technologies can appear to reduce our sense of limitation while simultaneously altering how we engage with the world [23]. In our context, this idea of distant technology seems relevant for understanding the relationship between researcher and research material, which becomes mediated through the use of a generative model.

The concept of distant technology thus provides a basis for reflecting on the dual methodological character of generative models. Within scholarly practice, models such as GPT-4o can function as powerful tools for a wide range of tasks, yet they also obscure the processes by which their output is produced. In this way, these models embody a form of technologically mediated engagement that both expands scholarly capacity and simultaneously introduces an additional layer of distance between researcher and material. While other OCR methods are likewise probabilistic and often black boxed, GPT-4o's smooth, high-quality, and seemingly error-free output arguably heightens the issue. Taken together, these theoretical perspectives highlight both the ambivalent character of the model as a research instrument and the need for thorough reflection.

3. Testing GPT-4o on Newspaper Text

To demonstrate the scope of the OCR-related problems with KB's collections, we can highlight the following excerpt from a review in the newspaper *Arbetet* of Gustaf af Geijerstam's novel *Farliga makter* [*Dangerous Powers*] (1906), which we have manually transcribed from the original print edition:

> Geijerstam eger på en gång vetenskapsmannens lugna objektivitet och diktarens förmåga att ge lif åt sina gestalter, och härmed sammanhänger den manligt okonstlade, man vore frestad säga sakliga stil, som präglar mycket af hans produktion och i all synnerhet ”Farliga makter”. Han dekorerar icke sina figurer med förträfflighetens epiteter, lika litet som han förlöjligar dem; han vet att det enda, hvaraf man kan sluta till en människas karaktär är hennes handlingar, och utan att försmå den psykologiska analysen låter han dock sina figurer först och sist uppträda som handlande individer [24].

This can be compared with the error-riddled passage from KB's OCR-processed version:

> Geijerstam eger på en gång vetenskapsmannens lugna ob jlek livi te t och diktarens förmåga att ge lif åt sina gestalter, och härmed sammanhänger denmanligt okonstlade, man vore frestadsäga sakliga stil, som käglar mycketaf hans produktion och i all synnerijftt -sdjaxli cj^.maktfcDiwi 4’ * »Ia’icke sina figurer med förträfflighetensepiteter, lika litet som han förlöjligardem; han vet att det enda, hvarafman kan sluta till en människas karaktär är hennes handlingar, och utan attförsmå den psykologiska analysen låter han dock sina figurer först ochsist uppträda som handlande individer.

3.1. Broken Indexicality?

While OCR errors may or may not impede downstream computational tasks, depending on the degree to which those tasks rely on orthographic precision, prior studies have demonstrated the detrimental impact of source text quality on, for instance, topic modeling. Nevertheless, whereas the analytical consequences of noisy OCR have conventionally been mitigated through manual curation, addressing such issues by means of a generative model entails a form of textual fabrication. One could debate which term most accurately describes this process, but we have tentatively chosen “retouching” as a more precise alternative, since it suggests both reuse and modification. Labels such as “correction” [26], “reconstruction”, or “restoration” seem a bit misleading, as the generated text is not derived from the original printed source but from a flawed digitized variant.

In effect, no clear indexical relationship with the source material remains [27]. Notably, OCR already performs a reduction of the document's semiotic complexity by disregarding qualities like letter shape and whitespace and thus transforming from image to text [28]. Generative retouching extends this abstraction one step further. Rather than repairing the losses introduced by OCR, it takes the already filtered text as its input for complex vectorial calculations that operate probabilistically on the level of meaning instead of on the level of the singular letter, effectively layering another transformation on top of the initial transcription. The unprocessed OCR output, despite its flaws, still retains an indexical resonance with the page as an artifact, since its errors often betray the contours of the original print. The retouched variant, by contrast, suppresses even those traces. What remains is not a causal imprint but a statistically inferred approximation, where referentiality is mediated twice over.

3.2 Significant or Negligible Discrepancies?

Notably, GPT-4o produces these texts not by recognising letterforms from pixel values or singular glyph information, but by mobilizing large-scale symbolic relations acquired during training. Instead of reconstructing the original texts through direct reference to the newspaper artifacts, it generates versions that are linguistically and contextually coherent, drawing on patterns of usage across vast corpora. In this sense, the model does not recreate the original material but provides an approximation mediated by the statistical and symbolic structures of language itself. Consequently, one cannot expect the generated texts to correspond exactly to the originals. Still, when we turn to the passage cited above, the retouched version produced by GPT-4o reads as follows:

> Geijerstam *äger* på en gång vetenskapsmannens lugna objektivitet och diktarens förmåga att ge *liv* åt sina gestalter, och härmed sammanhänger den manligt okonstlade, man vore frestad säga sakliga stil, som präglat mycket av hans produktion och i all synnerhet *fallet där makterna är mäktigast*. Han *prydar* icke sina figurer med förträfflighetens *epitet*, lika litet som han förlöjligar dem; han vet att det enda, *varav* man kan sluta till en människas karaktär är hennes handlingar, och utan att försmå den psykologiska analysen låter han dock sina figurer först och sist uppträda som handlande individer. [italics indicate divergence between the GPT-4o output and the printed originals]

Importantly, the generated text does not match the original exactly. There are discrepancies in spelling (GPT's version is modernized), a plural form has been rendered as singular and for illegible portions of the text the model has hypothesized that the title of the novel *Farliga makter* [Dangerous Powers] should instead be rendered as "fallet där makterna är mäktigast" ("the case where the powers are at their strongest"). While some differences relate only to spelling, others are more substantial and affect the process of meaning-making itself. A sentence-by-sentence comparison of the smaller set of retouched reviews with the original texts also certifies that this is the kind of typical differences we can see in the material. Since the output neither reproduces the original text with complete accuracy nor preserves a record of its alterations, it would be risky to base analytical interpretations solely on the retouched versions. At the same time, however, the GPT-4o output often appears closer to the original than the OCR-processed text found in KB's digitized newspaper collections.

3.3 Transparency and Uncertainty

A manual evaluation comparing the transcribed reviews with both KB's OCR versions and the GPT-4o outputs shows that the latter align much more closely with the transcriptions. To assess the OCR retouching process, we analyzed both the full corpus (n=394) and a manually transcribed subset (n=21). This involved calculating the Levenshtein distances between the noisy OCR text and its modified version, as well as between both versions and the manual transcriptions. The Levenshtein distance measures the number of operations (insertion, deletion, or substitution of characters) required to transform one string into another, providing a measure of textual change. A low score indicates minimal alterations, whereas a high score signals greater divergence. On average, the OCR retouching required 320 operations per text (95% CI: 253-386, n=394). In the manually transcribed subset, the retouched

texts were, on average, closer to the manual transcriptions than the original OCR versions, with the noisy OCR texts approximately 104 operations further from the gold standard (95% CI: 64-143, n=21).

Nevertheless, this raises epistemological concerns, as we are dealing with texts that appear more accurate and usable for computational analysis yet are not derived from, and remain detached from, the original print. In this sense, GPT-4o's contribution to OCR cleaning does more than enhance access to historical sources; it also introduces a layer of epistemological uncertainty. In the light of Esposito's argument that generative models are better understood as forms of artificial communication rather than intelligence, this retouching can be seen as a simulation that reshapes noisy OCR output while obscuring the processes behind it. At first glance, the result seems to bring us closer to the historical text, yet it introduces a perceptual and methodological distance. The improved legibility comes at the expense of traceability: without direct comparison to the print sources, similar to other OCR-correction processes, it is impossible to assess how closely the model's output reflects the originals. Rather than restoring the past, GPT-4o produces a plausible representation of it, underscoring the need for deeper scrutiny.

4. Conclusions

In this paper, we have experimented with and reflected on the application of a generative language model to Swedish literary criticism from 1905–1906. The experiment demonstrated the model's capacity to improve OCR-degraded texts, providing a stronger basis for data-driven analysis. At the same time, we showed that this process weakens the connection to the original sources and introduces epistemological uncertainty. Although the generated texts often resemble the originals more closely than the noisy OCR versions, they are best regarded as probabilistic retouching, rather than restoration, since they lack an indexical link to the original material and can no longer reliably point back to it.

Building on the idea that AI functions as a communicative system producing plausible outputs without semantic understanding, based on probabilistic pattern recognition, we have emphasized that a generative model like GPT-4o may perform what resembles high-quality OCR-cleaning while remaining methodologically distant. This distance arises primarily from the opacity of its computational processes, which complicates methodological transparency and calls for reflection on the interpretive gaps such systems introduce. As generative models become more prominent in humanities research, the task is not only to refine their practical applications but also to develop strategies for mitigating methodological distance and a theoretical awareness of the conditions under which these tools operate.

Acknowledgements

This paper was prepared as part of the research project *The Order of Criticism Revisited* (2020–2025), funded by Riksbankens Jubileumsfond (RJ), grant no. MXM19-1096:1.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. "Language Models are Few-shot Learners." Advances in Neural Information Processing Systems 33 (2020): 1877–1901.

[2] C.-E. González-Gallardo, T.T.H. Hahn, A. Hamdi, A. Doucet, "Leveraging Open Large Language Models for Historical Named Entity Recognition", in: A. Antonacopoulos, A. Hinze, B. Piowowarski, M. Coustaty, G.M. Di Nunzio, F. Gelati, N. Vanderschantz (Eds.), Linking Theory and Practice of Digital Libraries, TPDL 2024: 28th International Conference on Theory and Practice of Digital Libraries, Ljubljana, Slovenia, September 24–27, 2024, Proceedings, Part I, Springer, 2024.

[3] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bisselut, E. Burnskill, et al. 2021. "On the Opportunities and Risks of Foundation Models", Preprint, arXiv (2021). URL: https://arxiv.org/abs/2108.07258.

[4] J. Drucker, SpecLab: Digital Aesthetics and Projects in Speculative Computing, University of Chicago Press, Chicago, 2009.

[5] J. Ingvarsson, D. Brodén, L. Samuelsson, V. Wåhlstrand Skärström, N. Zechner, “The New Order of Criticism: Explorations of Book Reviews Between the Interpretative and Algorithmic”, in: K. Berglund, M. La Mela, I. Zwart (Eds.), The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), CEUR-WS.org (2022), pp. 228–234.
[6] L. Samuelsson, Kritikens ordning: Svenska bokrecensioner 1906, 1956, 2006, Bild, text & form, Karlstad, 2013.
[7] D. Brodén, L. Samuelsson, N. Zechner, J. Ingvarsson, A. Karimi, “Between the Arduous and the Automatic: A Comparative Approach to the Challenge of Identifying Book Reviews in Swedish Newspapers”, in: O. Holownia, E.S. Sigurðarson (Eds.), Digital Humanities in the Nordic and Baltic Countries 2024, May 27–31, 2024, Reykjavik, Iceland, University of Oslo Library (2025), pp. 1–12.
[8] J. Jarlbrink, P. Snickars, C. Colliander. “Maskinläsning: Om massdigitalisering, digitala metoder och svensk dagspress.” Nordicom Information 38.3 (2016): 27–40.
[9] L. Börjesson, C. Haffenden, M. Malmsten, F. Klingwall, E. Rende, R. Kurtz, F. Rekathati, H. Hägglöf, J. Sikora. “Transfiguring the Library as Digital Research Infrastructure: Making KBLab at the National Library of Sweden.” College & Research Libraries 85.4 (2024): 564–582.
[10] J. Sikora, C. Haffenden, “AI, Data Curation and the Readiness of Heritage Data: Exploring the Swedish Newspaper Archive at KBLab”, in: E. Volodina, G. Bouma, M. Forsberg, D. Kokkinakis, D. Alfter, M. Fridlund, C. Horn, L. Ahrenberg, A. Blåder (Eds.), Proceedings of the Huminfra Conference (HiC 2024), Linköping Electronic Conference Proceedings (2024), pp. 60–66.
[11] V. Löfgren, D. Dannélls, “Post-OCR Correction of Digitized Swedish Newspapers with ByT5, in: J. Bizzoni, S. Degeatano-Ortlieb”, in: A. Kazantseva, S. Szpakowicz (Eds.), Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024), Association for Computational Linguistics (2024), pp. 237–242.
[12] J. Ingvarsson, Towards a Digital Epistemology: Aesthetics and Modes of Thought in Early Modernity and the Present Age, Palgrave Macmillan, Cham, 2020.
[13] T. Underwood, Ted, Distant Horizons: Digital Evidence and Literary Change, University of Chicago Press, Chicago, 2019.
[14] A. Piper, Can We Be Wrong? The Problem of Textual Evidence in a Time of Data, Cambridge University Press, Cambridge, Mass, 2020.
[15] J. Brottrager, A. Stahl, A. Arslan, U. Brandes, T. Weitin. “Modeling and Predicting Literary Reception. A Data-rich Approach to Literary Historical Reception.” Journal of Computational Literary Studies 1.1 (2022): 1–27.
[16] D. Brodén, J. Ingvarsson, L. Samuelsson, V. Wåhlstrand Skärström. “Visualization as Defamiliarization: Mixed-methods Approaches to Historical Book Reviews.” Journal of Computational Literary Studies 3.1 (2024): 1–26.
[17] J.M. Mathews, 2022. “Some Critical and Ethical Perspectives on the Empirical Turn of AI Interpretability.” Technological Forecasting and Social Change 174: 121209.
[18] J. Hewitt, R. Geirhos, B. Kim, “We Can’t Understand AI Using Our Existing Vocabulary”, Preprint, arXiv (2025). URL: https://arxiv.org/abs/2502.07586.
[19] E. Esposito, Artificial Communication: How Algorithms Produce Social Intelligence, MIT Press, Boston, 2022.
[20] M. B. Fazi. “Can a Machine Think (Anything New)? Automation Beyond Simulation.” AI & Society: Knowledge, Culture and Communication 34.4 (2019): 813–24.
[21] A. Galgoon, K. Filom, A.R. Kannan, Mechanistic Interpretability of Large Language Models with Applications to the Financial Services Industry, Preprint, arXiv (2024). URL: https://arxiv.org/abs/2407.1121
[22] R. Romanyschyn, *Technology as Symptom and Dream*, Routledge, London and New York, 1989.
[23] J.P Telotte, A Distant Technology: Science Fiction Film and the Machine Age, Wesleyan University Press, Middletown, 1999.
[24] B. L. [Bengt Lidforss], “Nya böcker”, Arbetet (1906-03-10).
[25] https://tidningar.kb.se/s3n6n0hdqhjb8g57/part/1/page/2?q=geijerstam%20AND%20%22farliga%20makter%22%20AND%20%22tre%20m%C3%A5nader%20ha%22

[26] E. Boros, M. Ehrmann, M. Romanello, S. Najem-Meyer, F. Kaplan. “Post-Correction of Historical Text Transcripts with Large Language Models: An Exploratory Study”, in: *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024* (2024), pp. 133–59.
[27] J. Sterne. “Analog”, in: B. Peters (Ed.) Digital Keywords: A Vocabulary of Information Society and Culture, Princeton University Press, New Jersey, 2016, pp. 31–44.
[28] M. P. Eve, Theses on the Metaphors of Digital-textual History. Stanford University Press, Stanford, CA, 2024.

Appendix: Prompts for OCR retouching

System Prompt (original in Swedish)

Du är en expert inom OCR-korrigering av äldre svenska texter, med särskilt fokus på dokument från omkring år 1906. Din uppgift är att noggrant identifiera och korrigera OCR-fel i en text samtidigt som du bevarar originalets stil, historiska språkbruk och mening. När du utför korrigeringarna ska du:

Återge den ursprungliga tonen och språkliga karaktären som var typisk för tiden, inklusive äldre stavnings- och grammatikformer.

Rätta uppenbara fel i teckenigenkänning, såsom felaktiga bokstäver, orddelar eller interpunktion, utan att modernisera språket.

Säkerställa att korrigeringarna förbättrar läsbarheten och den språkliga korrektheten, samtidigt som den historiska känslan bevaras.

Om du är osäker på en korrigering, markera felet eller lämna det oförändrat med en anteckning för vidare granskning.

Arbeta metodiskt och noggrant, med beaktande av den kontext och stil som är typisk för svensk skrift från början av 1900-talet.

Använd denna vägledning för att producera en korrigerad version av OCR-texten med hög noggrannhet och respekt för den historiska textens ursprungliga uttryck.

Returnera endast den korrigerade texten, utan några kommentarer eller ytterligare förklaringar.

User Prompt (original in English)

OCR-correct this text. Do not shorten the text: