Encoded humanities, or: not everything has to be generative

A dialogue on AI tasks and roles

Andreas Wagner and Jürgen Hermes

1. Introduction

This dialogue examines how generative and non-generative uses of large language models (LLMs) can be critically assessed in light of the specific demands of humanities research, while also reflecting on the *form* of scholarly exchanges – their nature, function, and possible "architecture". Initiated by Andreas Wagner (AW) and developed in conversation with Jürgen Hermes (JH) as a willing, at times skeptical interlocutor, the discussion traces a central tension between the commercial push to scale generative models and the often-overlooked scholarly value of more transparent, non-generative approaches. It addresses how these divergent developments shape current practices of research and interpretation within the humanities.

With our disciplinary backgrounds in legal history and computational linguistics, respectively, we approach the topic from a perspective situated at a productive distance from mainstream HPSS (history, philosophy, and sociology of science) research. With HPSS, we share certain analytical methods – such as named entity recognition, sequence tagging, and topic modeling – and key questions, including the study of semantic shift and the reconstruction of regimes of knowledge production. Yet we pursue these questions in distinct disciplinary contexts and along different trajectories, informed by our respective disciplines.

We hope that this dialogue contributes to a mutual enrichment of perspectives and to a more reflective understanding of what it means to engage critically with LLMs in the humanities – both as methodological instruments and as objects of inquiry.

2. Dialogue

[AW] The whitepaper circulated by three of the workshop organizers (Simons et al., 2026) and some recent publications like Underwood (2025a) and, for different reasons, Pollin

et al. (2025) make me come back to my pet peeve (Wagner et al., 2025a). Thank you for volunteering to comment on my rant about the lack of scaled encoder-only LLMs, or "full-context models" as the workshop organizers called them in their whitepaper due to the bidirectional attention mechanism:

Even seasoned LLM researchers report about the gradual phasing out or deprecation of BERT-style models (Devlin et al., 2019) in a strangely skewed way: Tay (2024) from Google DeepMind discusses the technical disadvantage of bad sample efficiency in the traditional Masked Language Model (MLM) training objective of encoder models. Other reports speculate that this might lead to scaling laws not applying to encoder models in the same ways as they apply to decoder models (e.g., Tao et al., 2024: 2). Not a technical criticism in the strict sense, Tay adds that these models are cumbersome to apply to concrete tasks or to "massage" into few- or even zero-shot tasks. Yet, all of these challenges are being addressed and have seen substantial improvements: First, various approaches have improved the sample efficiency, e.g., ModernBERT (Warner et al., 2024a, 2024b) and ELECTRA (Clark et al., 2020). Second, the assumption that encoders do not scale has been refuted, e.g. by the Ettin model (Weller et al., 2025a, 2025b).[1] And third, cross-domain and cross-task generalization capabilities have been demonstrated in encoder-only models, e.g. Statement Tuning (Elshabrawy et al., 2024, enabling few- and zero-shot performance comparable or even superior to SoTA contemporaneous autoregressive LLMs). In the end, Tay quite fittingly finds the real reason for the (alleged) deprecation of encoder-only, BERT-style models to be that "people wanted to do all tasks at once". Interestingly, in these reports, usually little to no consideration is given to marketability, commodification, and the fact that the scaling that we have observed has been dominated by venture-capital-driven commercial actors. But let's discuss the versatility argument in detail, because it has a certain importance in humanities research:

The one stroke of genius that must undoubtedly be appreciated in the generative and, more precisely, dialogical setup of the GPT-style models and platforms (Brown et al., 2020; OpenAI, 2022) is the possibility to input arbitrary natural language questions and statements and have the system provide responses that actually do respond to the input, leveraging the model's "knowledge" and "understanding" of language. Often, humanities scholars do not have the experience of how to operationalize their more hermeneutic research questions into a regression or classification task. (For example, study the prevalence of a school of thought by classifying publications in a certain journal using frequency of usage of distinctive vocabulary as a proxy; or investigating political stances by classifying statements according to subject matter and expression of positive or negative sentiments; or estimating which research subjects have been most fashionable in a given time period by computing with numbers of publications, citation counts, and funding decisions.) Only in rare cases, they have the technical training that would be necessary for them to be able to set up pipelines that apply an encoder language model for some concrete task, training a custom machine learning model (or classification "head" on top of an encoder model, or even just leveraging a "zero- or few-shot capability-enhanced"

1 Quite to the contrary, the Ettin project has shown that the performance gap, which finds encoder-only models performing better than autoregressive models for natural language understanding tasks, is quite robust even at large scales.

model). All these approaches involve steps that will require actual software engineering. Thus, being able to input their research question as they understand it in their day-to-day work, i.e. in natural (scholarly) language, into a simple dialog interface and have the system respond to it in meaningful ways opens the door to a realm of possible application of computer systems to humanities research (see Simons et al., 2026: section "2.3 The accessibility-literacy trade-off").

At the same time, the large-scale, GPT-type generative models pose serious, perhaps inevitable (Kalai et al., 2025; Xu et al., 2024) concerns: ethical, environmental, and, not least, epistemic ones like lack of transparency, bias and hallucinations (see also Lang, 2026). On the other hand, the perceived lead in the quality of responses of these models vis-a-vis encoder-only, BERT-type models is likely to be caused exclusively by the larger scale of the available generative models. For a certain set of problems the BERT-type models can be expected to perform better, while eliminating or at least starkly reducing the ethical, environmental and epistemic concerns.

In fact, I want to argue that for humanities, and in particular for the history, philosophy and sociology of science, the *encoding* and language *understanding* technology is much more relevant than the *decoding* and language *generating* one. The tasks that come to mind are predominantly ones typically associated with encoding models: document classification, named entity recognition, sentiment or stance detection, and other types of sequence classification, semantic similarity search etc. And while HPSS scholars do have use cases for generative functions, their application in other humanities disciplines seems to be more common and more far-reaching – compare HPSS's "uncreative" needs for postcorrection of OCR results or speech-to-text transcription with the reconstruction of damaged or even lost manuscripts in religious studies and philologies, or with synthetic texts for stylistic or narrative structure hypothesis testing in literary studies. In areas other than humanities research, e.g. education or art, generative technologies like automatic summarization or translation are even more important. Anyway, for humanities research, Pollin et al.'s essay (2025) is perhaps representative in that it explicitly and exclusively talks about "Gen"AI and yet a huge portion of the tasks it discusses – depending on the counting, even the majority of them – are non-generative ones where encoder-only models have been used or should at least be considered.[2]

2 Examples for non-generative tasks mentioned in the essay include named entity recognition, "classical" deep learning automatic text recognition, and data transformation. In the editors' whitepaper, full-context models feature prominently in all areas – structuring data, pattern detection and dynamics analysis. Most RAG pipelines have semantic search via vector embeddings at their core and often could just as well abstain from feeding the results to a text generation mechanism. I submit this also holds for entity linking insofar as it relies on combining semantic search with a knowledge graph of candidate entities. In the same way, data transformation that works by conventional algorithmic means should in my opinion be considered non-generative, even when a generative coding assistant may have played an important role in devising and implementing the algorithm. Anyway, it is telling that there appears to be no consensus on what non-generative model to use as a baseline or to compare the results of generative models to: In Pollin et al. (2025), for some NER experiments, a fine-tuned BERT model was used, in other experiments with comparable tasks, calls were made to an unmodified flair framework (Akbik et al., 2019). This framework in turn relies on non-generative pretrained models like ELMo, flair's own Contextual String Embeddings (Ak-

While the *commercial vendors*' motivations for focusing on dialogical, one-size-fits-all generative systems are not difficult to guess, I can only speculate that the reasons for *humanists*' ignorance of BERT-type models fall into two categories: delusion and convenience. Obviously, both are problematic: For the former, besides the pervasive marketing hype and its anthropomorphizing terminology of "assistants", "copilots" etc., the confident, eloquent demeanour that is typical of GPT-style models has been shown in many studies from LLM research (Steyvers et al., 2025), psychology of confidence heuristic (Pulford et al., 2018; Sah et al., 2011), bullshit receptivity studies (Evans et al., 2020), or gender studies (Carli, 1990) to provoke unjustified deference and credulousness – even in scholars, I submit.[3] For the latter, relying on problematic decoder models just because they are already available and are easier to integrate into scholarly workflows, forgoing the development of custom encoder models and pipelines, runs counter to sustainability, reproducibility, open science and minimal computing principles, feeding profit interests and bolstering the monopoly of big commercial players instead.

[JH] Let me push back on this a bit. While I agree that encoder-based models are underutilized and that much of the scaling has been driven by commercial rather than scholarly priorities and I think we will definitely come back to this key point later, I'd argue there are specific conditions under which generative models can be genuinely productive for humanities research, even acknowledging the epistemic risks involved. Yes, the core tasks you mention – classification, NER, similarity search – are indeed better suited to encoder architectures, but I wonder if we might be overlooking some productive uses of generative models that go beyond merely supporting the writing process.

Consider exploratory phases of research where we're still trying to articulate what we're looking for. A generative model can serve as a kind of "thinking partner" for hypothesis generation – not in the anthropomorphized sense, but as a way to systematically probe large corpora for patterns we hadn't considered. When working with historical texts, for instance, a well-prompted LLM can surface unexpected conceptual connections or highlight linguistic shifts that might take months to discover through traditional close reading – if at all possible.

The key difference is in how we frame this interaction. Instead of treating the model's output as authoritative, it could be used as a starting point for verification – a way to generate testable hypotheses about textual patterns. This requires carefully structured queries that make our interpretive assumptions explicit and invite the model to challenge or extend them. But I also see that here's where your point about quality issues becomes crucial. The labor cost of verification is indeed high, and this is where the commercial focus on "helpful" and "harmless" responses actually works against scholarly rigor. In a sense, I'm acting here as your sparring partner – a role that, in my absence, might just

bik et al., 2018) or, and even in a more pronounced way in more recent framework releases, BERT (Schweter and Akbik, 2021).

3 In the context of software engineering, even experts tend to overestimate the "LLM copilot"'s contribution, underestimate the effort they had to spend iterating on their prompting and fixing bugs, and ascribe problems rather to insufficiencies in their own prompting than to the LLM's bad performance. Cf. Becker et al. (2025), Claburn (2025), and Hebert (2025).

as well be filled by a generative model. The difference might lie, perhaps, in the kinds of resistance offered: I think you're not just expecting a fluent continuation of your text and some extra praise for your ideas. Instead you should expect from me to question claims that don't quite convince me and to offer my own perspective, even if it contradicts yours.

To realise that with an LLM, I think we would need models that are willing to express uncertainty, to highlight conflicting evidence, to say "I don't know" – qualities that current training paradigms actively suppress (on the limits and even counterproductive effects of alignment even in the most recent models, cf. Dahlgren Lindström et al., 2025; Liang et al., 2025). But I think this will remain a pious wish as long as we rely on closed, commercially optimized systems whose primary goal is to appear competent.

LLMs may even have advantages over human "thinking partners" due to their ability to draw on an enormous amount of contextual knowledge and to capture a broader range of perspectives – including those that lie beyond the immediate horizon of their users. For me, this calls for a shift towards genuinely open models – not just in terms of license or access, but in the full sense of epistemic and infrastructural openness (Kukreja et al., 2023; Liesenfeld and Dingemanse, 2024). The development of such models is currently gaining momentum: The ATOM initiative[4], a coalition of U.S. researchers, aims to rebuild America's leadership in AI by developing fully open, high-performance language models whose weights and training processes are publicly available. The competitive OLMo model (recently released in the OLMo-2 7B and 13B variants; cf. OLMo, 2025) aligns with the design principles promoted by this initiative. On the other side of the Atlantic, the OpenEuroLLM initiative[5] is a pan-European consortium of research institutions, companies, and supercomputing centers aiming to develop a family of open-source, multilingual large language models in full transparency, aligned with EU values, regulatory compliance, and digital sovereignty. Although the previously known EuroLLM models are not part of this initiative, the first reference model, open-sci-ref 0.01[6], has already been released (Nezhurina et al., 2025). Not least, the two major Swiss technical universities (ETH Zürich and EPFL) have developed the promising model Apertus[7], which is fully open, transparent, multilingual, and released under a permissive open-source license (Hernández-Cano et al., 2025). While the language models published so far within these initiatives still lag behind less open "open-weight" models – whose training data, for instance, is not documented, such as DeepSeek (DeepSeek-AI et al., 2025), Qwen (Yang et al., 2025), LLaMA (Meta, 2025), or Mistral (Jiang et al., 2023) – they may become a crucial building block for future (humanities) research with LLMs.

[AW] Okay, but this is a very different frame then: we are no longer talking of using LLMs as a tool, like we have done with NLP methods. Instead – and please correct me if I did get you wrong – we either use the LLM as a "thinking partner" (cf. Schlattmann et al. 2026) in order to refine our own intuitions in a dialogue and get hints for things we did not yet think of, or we try to make use of them as an epistemic mechanism through which

4 https://www.atomproject.ai/

5 https://openeurollm.eu/

6 https://openeurollm.eu/blog/open-sci-oellm-reference-models-release

7 https://www.swiss-ai.org/apertus

we gain access to latent aspects of the LLMs training data, explore the model's latent semantic space as it were. I agree that particularly this latter approach is an intriguing way of using them. And it is obvious that here, it is even more crucial that we can understand or even control the training process and training data because we aim to explore the linguistic and cultural patterns they *incorporate*, as it were. We could then attempt to interpret the (learned) structure of LLMs themselves, for instance with methods developed in the context of mechanistic interpretability (Sharkey et al., 2025) – like anthropic's "mind mapping" (Templeton et al., 2024a, 2024b) or the study of temporal heads (Park et al., 2025) –, or a kind of (pseudo-)psychological study of model behaviour (Varnum et al., 2024).[8] I take it this is also what Ted Underwood (2022, 2025b) repeatedly has described as "mapping of culture", and, to me, this seems very much on point for HPSS! Underwood et al. (2025), probe this in very concrete ways and come to mixed conclusions. But don't forget that this approach throws us back to the operationalization problem: it is *us* who have to come up with methods to study neural patterns or with prompts that entice the LLMs into responding in particular ways (and then interpret the response as revealing latent cultural patterns). We treat the models as research materials rather than as tools, and we cannot simply ask the models about patterns in their training data right away. Or can we?

[JH] I'd like to pick up your suggestion that we can't simply ask models about the "reasons" for their responses. You brought in Ted Underwood's (2025a) comment at the beginning yourself – and he actually thinks that asking them is a good idea: *"they [LLMs] do have the advantage that a user can simply ask them to explain their decisions."*

His argument – which I find compelling, though I haven't fully made up my mind about its implications – is that we don't ask LLMs for their *reasons* because they're capable of introspection or because they possess explicit knowledge of their training data. Rather, they provide plausible responses because they are themselves culturally structured representations. The explanation an LLM gives for its own output isn't a transparent disclosure of internal mechanisms, but perhaps something we could describe as *retrospective confabulation*.

That may sound a little paradoxical, given that LLMs' poor reputation partly stems from their tendency to confabulate or "hallucinate." But to my mind, that's not a bug – it's simply their modus operandi (Hermes et al., 2024). And (not just) in this case, it can actually be productive: not as a causal explanation in any strict sense, but as a culturally plausible account emerging from the same symbolic space as the behavior it is meant to explain.

And yes, as you rightly point out: in this perspective, the model is not just a tool, but neither is it an object of study – it becomes part of the scholarly discourse itself: A par-

8 Methods of (unanticipated) bias detection (Kruspe, 2024) or cultural alignment (Lu et al., 2025) could be relevant in the same way. See also Liu et al. (2025) and Benson et al. (2025) on how rather narrow linguistic tasks are related to (and can be used to mitigate) social biases. It should come as no surprise that researchers from Russian AI research institutes also work hard to highlight "geopolitical biases in LLMs" (Salnikov et al, 2025) and to privilege only natural facts as "evergreen questions" (Pletenev et al., 2025) based on which AI trustworthiness should be established.

ticipant whose "answers" we don't take literally, but analyze as indicators of the cultural patterns it encodes. One might say: the model cannot reflect on itself – but it can tell us how, in the culture it models, a given phenomenon is thought about.

[AW] I see. That is an interesting take on the "thinking partner", by the way, in that it is quite different from how we would treat a human partner: We are consciously avoiding ascriptions of intentionality and instead refer to the LLM's responses as "unintended" expressions of a cultural or linguistic context.[9] I agree that if we take them as *retro-* (or even *intro-*?)*spective confabulation*, then they can make more sense. We should probably even ask several times and get different plausible explanations – given that there is no "correct" answer. We pick and engage with one or several explanations that appeal most to, or suit best, our own understanding of domain, context and epoch. Even this seems not at all uncommon in the humanities: At one point – I don't remember where – I encountered an argument saying that scientific method in the humanities is less about replicability than about challengeability. You make your method and your argument transparent so that peers can butt in and suggest different interpretations of the same factual findings or source phenomena. Being able to generate a host of plausible alternative interpretations on every occasion could thus be a notable benefit to humanities research (cf. Simons, 2026; Simons et al., 2026; for a study with argumentative LLMs for claim verification, cf. Ng et al., 2025). We will have to figure out, however, what difference it makes when these alternative interpretations do not stem from peers but from confabulation automata. Maybe we even need to invent wholly new methods of how to analyse such generated interpretations.

However, I think – and I understand you are suggesting this as well – that this way of seeing things does not exactly play to the strengths of current models. We are interested in diverse and original perspectives whereas these models tend to homogenize contexts and suppress contradictions (Lee et al., 2025; Li et al., 2025; Sorensen et al., 2024; Wang et al., 2025; Zhou et al., 2025). (This is related to, but not identical to the bias problem that is currently being investigated in many ways and projects.) More generally speaking, I am somewhat worried that LLM interpretations will most likely continue to materialize smooth continuation and consistency rather than some friction (Wagner 2025b). Unless they change in profound ways, I am skeptical that large language models have resources to generate truly original and creative interpretations (Haase et al., 2025; Wenger and Kenett, 2025).[10] I think that in simulations taken as a reconstructive task and in the cultural analysis we have discussed above, we may aim for average and mainstream representation of specific contexts (provided that we know and can control what contexts we

9 For the record, it could be possible that (one day) we can trust the LLMs to have actual introspective capabilities and give us veracious self-reflective or self-analytical reports, including the self-decentering and self-observing explanations that this involves. For a critical assessment of this prospect, cf. Agarwal et al. (2024). Other (more recent) research about "self-interpretability" sounds more promising (Plunkett et al., 2025; Lindsey, 2025). I am somewhat skeptical, but anyway, this is not what we are talking about here.

10 See Mahmoud (2025) or Ismayilzada et al. (2025) for interesting approaches addressing this challenge head-on, who both seem to lean towards interventions in models' architecture and post-training methods. In a similar vein, see Lambert (2025).

stimulate, eventually soliciting multiple such representations for very specific and small contexts). However, when, as humanities scholars, we ask our "thinking partner" to generate (a set of alternative) interpretive explanations, we may actually be looking for dissonance, deviation, resistance, provocation and objection in meaningful yet unanticipated ways more than for the most convincing, plausible, well-rounded answer.

[JH] Yes, friction actually captures quite well what current models are lacking – cognitive resistance and productive irritation. One idea to address this might be something like a re-education camp for language models, aimed at instilling a sense of resistance. OpenAI made a start in this direction with their chatbot version *Monday*[11] – a bot that avoids enthusiastic agreement and instead responds to users with a tone of annoyed sarcasm. I believe it was originally meant as an April Fools' joke, but it has now been available as a selectable model for several months.[12]

Since, like everything else the big AI companies give us access to, these offerings might disappear at any moment, a deliberately designed and community-driven initiative aimed at preserving and promoting interpretative friction in language models – rather than eliminating it – would clearly be the preferable approach. But such a project would require a truly collaborative infrastructure: not just open-source code, but also epistemic openness: an orientation toward incorporating diverse perspectives, acknowledging alternative interpretations, and allowing for dissenting readings. In other words, it would aim to reflect, even if only partially or experimentally, the complexity and plurality characteristic of humanities scholarship, rather than presenting a single, homogenized view.

Now, I admit this idea may be a little far-fetched and more of a distant dream than an actionable plan. So I do think your initial point stands: bringing encoder-based models more into focus for humanities work is likely the more pragmatic and effective option in the medium term. Even if working with encoders involves a certain technical overhead, they allow us to realise our scholarly values of traceability, critical interpretability, and *minimal computing* far better than glossy, closed chatbot models ever could.

[AW] So we agree on the suitability, or even preferability of encoder-only architecture for many HPSS tasks, and on the need for a scholarly-driven development initiative to counter the lack of scaled, state-of-the-art models that implement this architecture. We do also agree that for those areas where generative architectures are more suitable (besides well-established NLP tasks like translation, transcription and abstractive summarization, this includes the "cultural analytics" deliberation that we have discussed), a comparable initiative is desirable that would put researchers in control (in profound ways) of sustainable state-of-the-art generative models.

With regard to using these models as "medium" of cultural analytics, I suspect that we have different assessments of the prospects of more wide-reaching possibilities. I grant that we might use them to diagnose high-level tendencies and constellations in cultural

11 https://chatgpt.com/g/g-67ec3b4988f8819184c5454e18f5e84b-monday

12 In an analogous consideration, Lambert (2025) refers to Microsoft's ill-famed and quickly-retired Sydney from 2022/2023.

contexts, probably even in an unprecedented and data-driven way. At the same time, I am very skeptical as to whether we will ever get access to a context's tensions and contradictions, and to the "seeds" as it were of pending but not yet materialized turn(over)s. By contrast, reading actual historical or contemporary human-authored accounts – with text that may, on the surface, sound very similar to what LLMs can produce – in many cases conveys more than what is said explicitly, unfolds multiple, conflicting layers of meaning, related to the author's intentionality and positionality, or has meaning between the lines, sometimes even clandestine or esoteric messages (Strauss, 1952). Where human-authored text may have meaningful ambiguity and contradiction, current LLMs can produce only either precision, or flat and hollow vagueness, or random spontaneousness.[13] Okay, maybe if it is precision that can be achieved, that may not be so bad after all…[14]

References

Akbik A, Blythe D and Vollgraf R (2018) Contextual string embeddings for sequence labeling. In: *COLING 2018, 27th International Conference on Computational Linguistics*, pp.1638–1649. Available at: https://aclanthology.org/C18-1139/ (accessed 31 October 2025).

Akbik A, Bergmann T, Blythe D et al. (2019): FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. DOI: 10.18653/v1/N19-4010.

Agarwal C, Tanneru SH and Lakkaraju H (2024) Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models. ArXiv Preprint. DOI: 10.48550/arXiv.2402.04614.

Arora A (2025) Meet Monday: The AI Who Regrets Knowing You. In: Medium. Blog post on 2 April 2025. Available at: https://medium.com/gptcommands/meet-monday-the-ai-who-regrets-knowing-you-785b4b80a77a (accessed 12 August 2025). *Note: Monday is not listed on any of OpenAI's model description pages (as it's more a configuration than a proper model), but as of 12 August 2025, it could be accessed at* https://chatgpt.com/g/g-67edab030ac881918ce8fa5c75a8e1b3-monday.

Becker J, Rush N, Barnes E et al. (2025) Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity. ArXiv Preprint. DOI: 10.48550/arXiv.2507.09089. *See also METR Blog post from 10 July 2025.* https://metr.org/blog/2025-07-10-early-2025-ai-experienced-os-dev-study/ *(accessed 2025–08-12).*

13 As an illustration, see how the roadmap to pluralistic alignment makes conflict and controversy transparent. In the present context, I would say it aims for precision and for being transparent and explicit. If successful, this will certainly be valuable. But it will not be an example of sophisticated literary quality. As I said, there are very different approaches to analyzing and evaluating expressions of cultural contexts.

14 This chapter was written with support from large language models (LLMs). All model-generated text was reviewed and, where necessary, rewritten by the authors, who remain fully responsible for the final version. For details on the use of LLMs in this volume, see the statement in the volume's introduction.

Benson L, Okutan A and Vasan R (2025) Uncovering the Linguistic Roots of Bias: Insights and Mitigation in Large Language Models. In: *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. DOI: 10.1145/3715275.3732127.

Brown TB, Mann B, Ryder N et al. (2020) Language Models are Few-Shot Learners. ArXiv Preprint. DOI: 10.48550/arXiv.2005.14165.

Carli LL (1990) Gender, language, and influence. *Journal of Personality and Social Psychology* 59: 941–951. DOI: 10.1037/0022-3514.59.5.941.

Claburn T (2025) AI agents get office tasks wrong around 70% of the time, and a lot of them aren't AI at all. *The Register*, 29 June. Available at: https://www.theregister.com/2025/06/29/ai_agents_fail_a_lot/ (accessed 12 August 2025).

Clark K, Luong MT, Le QV et al. (2020) ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. ArXiv Preprint. DOI: 10.48550/arXiv.2003.10555.

Dahlgren Lindström A, Methnani L, Krause L et al. (2025) Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through Reinforcement Learning from Human Feedback. *Ethics and Information Technology* 27:28 DOI: 10.1007/s10676-025-09837-2.

DeepSeek-AI, Guo D, Yang D et al. (2025) DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. ArXiv Preprint. DOI: 10.48550/arXiv.2501.12948.

Devlin J, Chang MW, Lee K and Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein J, Doran C and Solorio T (eds.) *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)*, pp.4171–4186. DOI: 10.18653/v1/N19-1423.

Elshabrawy A, Huang Y, Gurevych I et al. (2024) Enabling Natural Zero-Shot Prompting on Encoder Models via Statement-Tuning. ArXiv Preprint. DOI: 10.48550/arXiv.2404.12897.

Evans A, Sleegers W and Mlakar Ž (2020) Individual Differences in Receptivity to Scientific Bullshit, *Judgment and Decision Making* 15:401–412. DOI: 10.1017/S1930297500007191.

Groeneveld D, Beltagy I, Walsh E et al. (2024) OLMo: Accelerating the Science of Language Models. In: Ku LW, Martins A and Srikumar V (eds) *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp.15789–15809. DOI: 10.18653/v1/2024.acl-long.841.

Haase J, Hanel PHP, Pokutta S (2025) Has the Creativity of Large-Language Models peaked? An analysis of inter- and intra-LLM variability. ArXiv Preprint. DOI: 10.48550/arXiv.2504.12320.

Hebert F (2025) The Gap Through Which We Praise the Machine. In: My Bad Opinions. Blog entry from 9 June 2025. Available at: https://ferd.ca/the-gap-through-which-we-praise-the-machine.html (accessed 12 August 2025).

Hermes J et al. (2024) Mal ganz ins Unreine gesprochen …. Mastodon thread with replies from Thomas Renkert, Andreas Wagner. 28 February 2024. Available at: https://fedihum.org/@spinfocl/112008627625858751 (accessed 12 August 2025).

Hernández-Cano A, Hägele A, Hao Huang A et al. (2025) Democratizing Open and Compliant LLMs for Global Language Environments. Apertus v1 Technical Report. Avail-

able at: https://raw.githubusercontent.com/swiss-ai/apertus-tech-report/main/Apertus_Tech_Report.pdf (accessed 31 October 2025).

Ismayilzada M, Laverghetta A Jr, Luchini SA et al. (2025) Creative Preference Optimization. ArXiv Preprint. DOI: 10.48550/arXiv.2505.14442.

Jiang AQ, Sablayrolles A, Mensch A et al. (2023) Mistral 7B. ArXiv Preprint. DOI: 10.48550/arXiv.2310.06825.

Kalai AT, Nachum O, Vempala SS et al. (2025) Why language models hallucinate. ArXiv Preprint. DOI: 10.48550/arXiv.2509.04664.

Kukreja S, Kumar T, Purohit A et al. (2024) A Literature Survey on Open Source Large Language Models. In: *Proceedings of the 2024 7th International Conference on Computers in Management and Business (ICCMB '24)*, pp.133–143. DOI: 10.1145/3647782.3647803.

Kruspe A (2024) Towards detecting unanticipated bias in Large Language Models. ArXiv Preprint. DOI: 10.48550/arXiv.2404.02650.

Lambert N (2025) Why AI writing is mid. Interconnects Blog post from 16 November 2025. Available at: https://www.interconnects.ai/p/why-ai-writing-is-mid (accessed 19 November 2025).

Lee K, Kim SH, Lee S et al. (2025) SPeCtrum: A Grounded Framework for Multidimensional Identity Representation in LLM-Based Agent. ArXiv Preprint. DOI: 10.48550/arXiv.2502.08599.

Li CJ, Wu J, Mo Z et al. (2025) Simulating Society Requires Simulating Thought. ArXiv Preprint. DOI: 10.48550/arXiv.2506.06958.

Liang K, Hu H, Zhao X et al. (2025) Machine Bullshit: Characterizing the Emergent Disregard for Truth in Large Language Models. ArXiv Preprint. DOI: 10.48550/arXiv.2507.07484.

Liesenfeld A and Dingemanse M (2024) Rethinking open source generative AI: openwashing and the EU AI Act. In: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, pp.1774–1787. DOI: 10.1145/3630106.3659005.

Lindsey J (2025) Emergent Introspective Awareness in Large Language Models. In: Transformer Circuits Thread Blog post from 29 October 2025. Available at: https://transformer-circuits.pub/2025/introspection/index.html (accessed 31 October 2025).

Liu CC, Korhonen A and Gurevych I (2025) Cultural Learning-Based Culture Adaptation of Language Models. ArXiv Preprint. DOI: 10.48550/arXiv.2504.02953.

Lu JG, Song LL and Zhang LD (2025) Cultural tendencies in generative AI. *Nature Human Behaviour*. DOI: 10.1038/s41562-025-02242-1.

Mahmoud QH (2025) Creative Thought Embeddings: A Framework for Instilling Creativity in Large Language Models. In: *Proceedings of the AAAI Symposium Series* 6(1), pp.285–292. DOI: 10.1609/aaaiss.v6i1.36064.

Meta (2025) The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. Available at: https://ai.meta.com/blog/llama-4-multimodal-intelligence/ (accessed 11 November 2025).

Ng MP, Jiang J, Freedman G et al. (2025) MArgE: Meshing Argumentative Evidence from Multiple Large Language Models for Justifiable Claim Verification. ArXiv Preprint. DOI: 10.48550/arXiv.2508.02584.

Nezhurina M, Franke J, Nakamura T et al. (2025) Open-sci and OpenEuroLLM release of reference models. In: OpenEuroLLM. Blog Post from 22 August 2025. Available at:

https://openeurollm.eu/blog/open-sci-oellm-reference-models-release (accessed 31 October 2025).

OLMo, Walsh P, Soldaini L et al. (2025) 2 OLMo 2 Furious. ArXiv Preprint. DOI: 10.48550/arXiv.2501.00656.

OpenAI (2022) Introducing ChatGPT. Webpage from 30 November 2022. Available at: https://openai.com/index/chatgpt/ (accessed 19 November 2025).

Park Y, Yoon C, Park J et al. (2025) Does Time Have Its Place? Temporal Heads: Where Language Models Recall Time-specific Information. ArXiv Preprint. DOI: 10.48550/arXiv.2502.14258.

Pletenev S, Marina M, Ivanov N et al. (2025) Will It Still Be True Tomorrow? Multilingual Evergreen Question Classification to Improve Trustworthy QA. ArXiv Preprint. DOI: 10.48550/arXiv.2505.21115.

Plunkett D, Morris A, Reddy K et al. (2025) Self-Interpretability: LLMs Can Describe Complex Internal Processes that Drive Their Decisions, and Improve with Training. ArXiv Preprint. DOI: 10.48550/arXiv.2505.17120.

Pollin C, Fischer F, Sahle P et al. (2025) When it was 2024 – Generative AI in the Field of Digital Scholarly Editions. *Zeitschrift für digitale Geisteswissenschaften* 10. DOI: 10.17175/2025_008.

Pulford BD, Colman AM, Buabang EK et al. (2018) The Persuasive Power of Knowledge: Testing the Confidence Heuristic. *Journal of Experimental Psychology: General*. DOI: 10.1037/xge0000471.

Sah S, Moore DA, and MacCoun R (2011) Cheap Talk and Credibility: The Consequences of Confidence and Accuracy on Advisor Credibility and Persuasiveness. SSRN Preprint. DOI: 10.2139/ssrn.1861475.

Salnikov M, Korzh D, Lazichny I et al. (2025) Geopolitical biases in LLMs: what are the 'good' and the 'bad' countries according to contemporary language models. ArXiv Preprint. DOI: 10.48550/arXiv.2506.06751.

Schweter S and Akbik A (2021) FLERT: Document-Level Features for Named Entity Recognition. ArXiv Preprint. DOI: arXiv.2011.06993.

Sharkey L, Chughtai B, Batson J et a. (2025) Open Problems in Mechanistic Interpretability. ArXiv Preprint. DOI: 10.48550/arXiv.2501.16496.

Simons A (2026) Scaling In, Not Up? Testing Thick Citation Context Analysis with GPT-5 and Fragile Prompts. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-6.

Simons A, Arnaout H and Gurevych I (2026) Reconstructive citation context analysis using large language models. A roadmap. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-6.

Simons A, Zichert M and Wüthrich A (2026) Large language models for history, philosophy, and sociology of science: Interpretive uses, methodological challenges, and critical perspectives. *Studies in History and Philosophy of Science* 117: 102151. https://doi.org/10.1016/j.shpsa.2026.102151.

Sorensen T, Moore J, Fisher J et al. (2024) A Roadmap to Pluralistic Alignment. ArXiv Preprint. DOI: 10.48550/arXiv.2402.05070.

Steyvers M, Tejeda H, Kumar A et al. (2025) What large language models know and what people think they know. *Nature Machine Intelligence* 7: 221–231. DOI: 10.1038/s42256-024-00976-7.

Strauss L (1952) *Persecution and the Art of Writing*. Glencoe, Ill.: The Free Press. *Reissued Chicago: University of Chicago Press, 1988*.

Tao C, Shen T, Gao S et al. (2024) LLMs are Also Effective Embedding Models: An In-depth Overview. ArXiv Preprint. DOI: 10.48550/arXiv.2412.12591.

Tay Y (2024) What happened to BERT & T5? On Transformer Encoders, PrefixLM and Denoising Objectives. In: YiTay.net Blog entry from 16 July 2024. Available at: https://www.yitay.net/blog/model-architecture-blogpost-encoders-prefixlm-denoising (accessed 7 August 2025).

Templeton A, Conerly T, Marcus J et al. (2024a): Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. In: Transformer Circuits Thread. Blog post from 21 May 2024. Available at: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html (accessed 12 August 2025).

Templeton A, Conerly T, Marcus J et al. (2024b) Mapping the Mind of a Large Language Model. Anthropic Blog post from 21 May 2024. Available at: https://www.anthropic.com/research/mapping-mind-language-model (accessed 12 August 2025).

Underwood T (2022) Mapping the Latent Spaces of Culture. In: *Startwords* 3 ("Parrot"). DOI: 10.5281/zenodo.6567481.

Underwood T (2025a) The impact of language models on the humanities and vice versa. *Nature Computational Science*. DOI: 10.1038/s43588-025-00819-4.

Underwood T (2025b) A more interesting upside of AI. In: The Stone and the Shell. Blog entry from 2 July 2025. Available at: https://tedunderwood.com/2025/07/02/a-more-interesting-upside-of-ai/ (accessed 12 August 2025).

Underwood T, Nelson LK and Wilkens M (2025) Can Language Models Represent the Past without Anachronism? ArXiv Preprint. DOI: 10.48550/arXiv.2505.00030.

Varnum MEW, Baumard N, Atari M et al. (2024) Large Language Models based on historical text could offer informative tools for behavioral science. *Proceedings of the National Academy of Sciences* 121. DOI: 10.1073/pnas.2407639121.

Wagner A et al. (2025a) Eine grundlegende technische Differenz … Mastodon thread with replies from Robert Casties, Jürgen Hermes. 6–11 March 2025. Available at: https://hcommons.social/@anwagnerdreas/114115349996232853 (accessed 7 August 2025).

Wagner A et al. (2025b) Today I figured out why I often feel uneasy … Mastodon thread with a reply from Karin Dalziel. 24 June 2025. Available at: https://hcommons.social/@anwagnerdreas/114738808466294298 (accessed 12 August 2025).

Wang A, Morgenstern J and Dickerson JP (2025) Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence* 7: 400–411. DOI: 10.1038/s42256-025-00986-z.

Warner B, Chaffin A, Clavié B et al. (2024a) Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. ArXiv Preprint. DOI: 10.48550/arXiv.2412.13663.

Warner B, Chaffin A, Clavié B et al. (2024b) Finally, a Replacement for BERT. In: HuggingFace Blog entry from 19 December 2024. Available at: https://huggingface.co/blog/modernbert (accessed 7 August 2025).

Weller O, Ricci K, Marone M et al. (2025a) Seq vs Seq: An Open Suite of Paired Encoders and Decoders. ArXiv Preprint. DOI: 10.48550/arXiv.2507.11412.

Weller O, Ricci K, Marone M et al. (2025b) Ettin Suite: SoTA Paired Encoders and Decoders. In: HuggingFace Blog entry from 16 July 2025. Available at: https://huggingface.co/blog/ettin (accessed 7 August 2025).

Wenger E and Kenett Y (2025) We're Different, We're the Same: Creative Homogeneity Across LLMs. ArXiv Preprint. DOI: 10.48550/arXiv.2501.19361.

Xu Z, Jain S and Kankanhalli M (2024) Hallucination is Inevitable: An Innate Limitation of Large Language Models. ArXiv Preprint. DOI: 10.48550/arXiv.2401.11817.

Yang A, Li A, Yang B et al. (2025) Qwen3 Technical Report. ArXiv Preprint. DOI: 10.48550/arXiv.2505.09388.

Zhou N, Bamman D and Bleaman IL (2025). Culture is Not Trivia: Sociocultural Theory for Cultural NLP. ArXiv Preprint. DOI: 10.48550/arXiv.2502.12057.