We Don't Know How to Assess LLM Contributions in VIS/HCI

Anamaria Crisan *
University of Waterloo, CANADA

Abstract

Submissions of original research that use Large Language Models (LLMs) or that study their behavior, suddenly account for a sizable portion of works submitted and accepted to visualization (VIS) conferences and similar venues in human-computer interaction (HCI). In this brief position paper, I argue that reviewers are relatively unprepared to evaluate these submissions effectively. To support this conjecture I reflect on my experience serving on four program committees for VIS and HCI conferences over the past year. I will describe common reviewer critiques that I observed and highlight how these critiques influence the review process. I also raise some concerns about these critiques that could limit applied LLM research to all but the best-resourced labs. While I conclude with suggestions for evaluating research contributions that incorporate LLMs, the ultimate goal of this position paper is to simulate a discussion on the review process and its challenges.

Index Terms: Large Language Models, Visualization, Review Process

1 Introduction

The publication of AlexNet [7] in 2012 was a watershed moment for deep neural networks (DNNS). In a matter of a few years, DNN came to be the predominant technique for a variety of tasks, rendering earlier methods as obsolete or just less viable. Even skeptics of DNNs eventually came to appreciate their versatility and shifted aspects of their research programs. The pace of innovations in DNNs accelerated over the coming years. Alongside these innovations, visualization research also began to explore DNNs, both their use for various tasks and to interrogate their inner workings [6]. Another significant watershed moment was the invention of the Transformer Model in 2017 [11], which eventually paved the way for the Generative Pre-trained transformer model (GPT, [1]) architecture that now dominates the landscape of Large Language Models (LLMs). While it took several iterations for GPT models to demonstrate their capabilities, the eventual release of ChatGPT in November of 2022 generated significant public and research interest.

A consequence of this AI Summer [5, 2] is a proliferation of submissions to visualization (VIS) and human-computer interaction conferences (HCI) that use LLMs or that study their behavior. While some of this proliferation is reflected in the proceedings of these venues, what was absent was a sense of what research did not make the cut. Over the past year, I served on the program committee of several VIS and HCI conferences. A sizeable majority of papers I was assigned concerned LLMs in some form. As a result, I observed first-hand what made the cut and, more commonly, what did not. Here I summarize some of the common critiques that I observed and include my subjective reflection on their influence toward the review process. **The goal of this brief position paper is to raise awareness of these common critiques and elevate them for wider discussion among the visualization research community as we strive to navigate the increased use of LLMs.**

*e-mail: ana.crisan@uwaterloo.ca

2 Summarizing Reviewer Critiques

I present anecdotal observations from my experience serving on program committees of four conferences. As reviewing is a confidential process, I do not present evidence or take any particular methodological approach to summarize the observations I discuss here. However, do include a positionality statement that clarifies my own feelings toward LLM use in visualization and HCI research.

2.1 Context and positionality

I volunteered to serve as an associate chair or program committee member for four conferences, CHI, FAccT, CSCW, and VIS. One reason I did this was because I was genuinely curious about how new technology is reviewed. I also had, what I consider to be, a reasonable tenure acting as an AC/PC member and have also submitted my research to these venues (both successfully and unsuccessfully). Given that expert reviewers are in relatively short supply, I also felt my efforts doubled as good service contributions – to that end, I did my best to try to ensure good quality reviews.

While I do not have visibility into the total scope of submissions across all of these venues, I bid on research works concerning LLMs for data analysis and visualizations. The total sample of papers I was responsible for was roughly 45. Accounting for the total diversity of my bids, my areas of expertise, and conflicts, approximately 60% of those papers concerned LLMs. Across two venues, my role was only to provide a summary review when serving as primary (CHI, CSCW), otherwise, I had to provide a full review for all papers. Except for FAccT, I was also responsible for assigning external reviewers.

Despite our best intentions, subjective research perspectives influence how any paper is appraised. For this reason, I will briefly summarize my views on LLMs. I generally find LLMs to be a surprisingly capable if still limited, technology. Some of their flaws are serious, perpetuating harmful biases in ways that are difficult to fully detect or mitigate. Their environmental impacts should not be ignored. At the same time, improvements to these models are also occurring at a surprising pace. In certain tasks, they exhibit a flexibility and capability that is a genuine and significant improvement over prior methods. A good, if obvious, example is enabling conversational interactions (via text or voice) with data; LLMs are an improvement over domain-specific languages, pares trees, heuristics, or other machine learning approaches. If and when LLM capabilities will plateau remains to be seen and it is difficult to know how these capabilities will evolve (or how quickly). *Overall, I currently have a generally positive view toward the use of LLM technology, if the research motivates a fairly reasonable rationale of its use.* Relative to the range of perspectives I have encountered — from LLMs are useless toys to LLMs are the harbinger of Artificial General Intelligence that will replace all humans — I would rate my perspective as somewhere more toward the middle of this spectrum, and potentially more toward the former (toys) than the latter (p(doom) = 100%).

2.2 Critiques

What follows is a non-exhaustive list of common critiques I observed while acting as a program committee member. As I present these critiques, I add my subjective commentary summarizing the pros and cons.

Critique 1: LLMs have non-deterministic behavior. This is the low-hanging fruit of critiques, even if it is a valid one. Language models will produce different, even if highly similar, responses to the same inputs. While some of this behavior can be mitigated slightly, for example by modifying the temperature, the overall sensitivity of LLMs to prompts makes it difficult to fully appraise their behavior and performance. For contributions using LLMs as part of a system or technique implementation, this means that evaluations are especially vulnerable. Are the evaluation results simply lucky and do they reflect what long-term or ongoing use might look like? Contributions studying LLM behaviors face similar challenges.

One possible approach to begin to address this issue could be to include a sensitivity (or multiverse) analysis as part of any submission that uses LLMs. If authors can show that they have at least considered how sensitive models are to prompting strategies, and present some evidence for the variability in the model's results, this could be at least some reasonable evidence for reviewers to consider. However, these kinds of analyses are not common in VIS/HCI research (they arguably should be) and it remains a roll of dice as to whether this could be considered acceptable evidence.

Critique 2: You didn't evaluate against enough LLMs. While there are many similarities among the architectures, training data, training methods, and even fine-tuning approaches, among all LLMs currently available, their behaviors and outputs are still different. At a high level, there is a distinction between open-source (e.g. LLaMA![10]) and closed-source (e.g., ChatGPT) models. Some reviewers take issue with research that only uses closed-sourced models because it does impact the reproducibility of the research. At the time of this writing, some of the best-performing models are closed-sourced. Effectively comparing against multiple LLMs means comparing against both types of models, and in fact, many research papers do this, especially in AI/ML and NLP literature where assessing model performance on benchmark datasets is more common.

I observed that it was less common for VIS/HCI papers to conduct such comprehensive examinations against multiple LLMs, which was also a common source of reviewer critiques. Many of the concerns around comparing to multiple LLMs were extensions of those concerning the non-deterministic behavior of LLMs – that is, that the reported performance was lucky based upon the choice of model. Interestingly, this critique would also be valid for any VIS/HCI research that incorporated ML models. Why did so many topic model papers just use one specific implementation of LDA? There exist other implementations and other methods, but it has not been a requirement for authors to exhaustively consider them.

Generally, this is one critique where I feel considerable sympathy toward authors. Not all reviewers requested comparisons to multiple LLMs, and indeed papers were accepted having only considered ChatGPT. However, it is a valid and relatively simple critique that can be difficult to address. Closed-source LLMs have costs associated with API use, but these can still be cheaper relative to the infrastructure required to host multiple open-source LLMs. The costs can quickly add up if authors are required to compare against multiple models and also conduct some analysis around the sensitivity of model outputs. Thus, this critique may limit research into LLM use and applications to groups with financial and infrastructure resources in place. This is also regrettably true for a lot of research, but especially acute for LLMs because of the significant interest around them.

Critique 3: You didn't evaluate against the latest LLMs. Similar to the previous critique, language models are also being constantly updated and new models with state-of-the-art performance are being released. The timing of these updates and releases is unpredictable. Being able to quickly add this model to your research paper, even if extremely close to the submission deadline, may be required. I was somewhat concerned by the number of CHI'24 papers including GPT-4, which was released July 2023 or approximately 2.5 months before the submission deadline. More concerned still see work rejected because it did not compare to GPT-4. Some research groups, by the privilege of their associations, get early access to these models and so have genuine opportunities to robustly assess them. More likely, others are adding these models at the last minute and conducting rushed evaluations ahead of submitting their research. It may be possible to conduct robust assessments in a short period of time, but, I express some skepticism, especially in light of Critiques 1 and 2. Once again, the ability to conduct robust and meaningful research may thus be limited to a small number of groups with access and resources.

Critique 4: You didn't discuss the latest LLM paper (and it obviates your results). New research concerning LLMs in various application contexts is constantly appearing on pre-print sites like ArXiv. A subset of this research is high-quality work that will eventually be published at well-recognized research venues. However, the majority of this research is more oriented toward 'flag-planting' – the authors of such pre-prints wish to claim some stake in a quickly moving field. An argument could be made this requires some prudence from researchers to focus their work on areas that might be more future-proof to being prematurely scooped [2]. However, I would push back on this idea a little bit. Just because some research exists on a topic, it does not mean that it is good. It is up to the reviewer to consider the merits of existing pre-prints relative to the manuscript under revision. Is the pre-print making solid and valid claims? Does the current manuscript under review perhaps do a better job, or take a more innovative lens or perspective on the problem? Unfortunately, sifting through the glut of pre-prints would also increase the overhead of the review process. Some conferences have adopted a reasonable limit on pre-prints and even new models, indicating that *it is not reasonable to ask authors to conduct in-depth comparison to pre-prints, or even new models, that were released fewer than three months prior to the submission deadline* (e.g., see the ACL Policies for Review and Citation).

However, with the rapid pace of LLM research, review cycles, from initial submission to eventual publication, may simply be too long to ever seriously address this critique. Reviewers, and perhaps the wider research community, may be significantly underestimating the challenges of scoping a truly robust multi-year research plan against the backdrop of LLM advancements. I speculate that even if researchers avoid LLMs altogether, the critique of "couldn't an LLM do this?" can still crop up.

Critique 5: You didn't cite ANY relevant research. This was perhaps one of the more surprising, and yet also very valid, reviewer concerns – and one that I raised myself too many times. Some papers did not consider *any* prior research on a particular topic area for the submission venue or related venues. As an example, this would constitute a paper on LLMs for visualizing topics in text documents without citing CHI, VIS, (etc.) prior research in this area. Often these papers would entirely cite from the machine learning and natural language processing literature. It may be the case that the research does present some brand new topic that VIS/HCI research venues have not already examined, but I have yet to find an example of such a case. It is possible, as others have speculated [3], that this reflects attempts of AI/ML/NLP researchers to publish their works across a greater variety of venues. If true, and if this trend continues, it can serve to overburden an already overworked reviewer pool. Moreover, authors of such submissions would be inappropriate reviewers for other works appearing at the VIS/HCI venues they are submitting to because they lack expertise.

Critique 6: The LLM Wrapper Paper. In a recent blog post, Ian Arawjo defined the phenomenon of LLM wrapper papers to essentially be "we apply LLMs to X problem" [3]. I have observed that there is generally negative sentiment toward LLM wrapper papers

and I would agree in instances where these submissions truly do not cite any relative prior work (Critique 5).The issue is that sometimes, examining or applying LLMs to X problem is both useful and pertinent if done well. This is true if for no other reason than companies seeking to commercialize this technology ARE trying to sell LLMs as a solution for X problem. It is pertinent to have independent critical voices that can objectively assess, within some reasonable limit, the extent that these commercial claims are true.

Critique 7: I don't like LLMs/I am so over them. Yes, this is a critique that comes up. This perspective may be stated explicitly or implicitly in the review or articulated during the discussion period amongst reviewers. Sometimes a solid framing and motivation can address this critique, but, it is regrettably a roll of the dice what its impact is on the overall review process.

3 What is to be done?

In the long term, I have concerns about whether the myopic focus on one technology will ultimately stifle research in VIS and HCI. I have heard similar concerns expressed about a narrowing focus in AI/ML/NLP research. However, there is not only internal pressure to keep up with the latest technology, but also external pressure from funding agencies, industry partners, governments, and even academic hiring committees to invest in research, and researchers, with so-called 'AI expertise'. So how do we manage the 'AI summer' [5, 2] and particularly its impacts on the review cycle? I have a few suggestions.

Make a reasonable attempt to understand the variability of LLM outputs. Supplemental Materials are a useful place to include a basic analysis of how sensitive language models are to specific prompts. There is some active research [9] in this area too that would be helpful for visualization researchers to be aware of and consider incorporating. While it can be difficult to define what constitutes a basic or reasonable analysis, I would initially set the bar quite low – authors should demonstrate that they at least tried to think about it and have some sense of how much their results are influenced by prompt choice. I believe that many published papers would not meet this standard. In general, AI/ML models are sensitive to their initialization, parameter choices, training data, etc., and it is good practice to conduct sensitivity or multiverse analysis.

Make a reasonable attempt to justify the use of an LLM. Does your research really require the use of an LLM? The answer can be yes. Is it really worthwhile to study the behavior of an LLM for X problem? The answer can also be yes. It is useful to articulate a rationale for using or studying an LLM that provides better justification than "everyone is using LLMs these days". As an example, initial research suggests that LLMs can outperform prior approaches for generating visualizations from natural language utterances [8]. It's reasonable to conduct research that builds on those findings or tests their validity. However, there may also be some instances where an LLM adds limited value while being more costly. Again, what constitutes a reasonable justification is highly subjective. Somewhere in between the spectrum of "everyone is using an LLM" and "exhaustively compare to all prior approaches to justify using an LLM", there is a reasonable balance.

Limit requirements to incorporate pre-prints and new models released close to the submission deadline. As already indicated, some venues have set a threshold of three months before the submission deadline. It may otherwise be deemed unreasonable for reviewers to request comprehensive comparisons to new released work or models. It may be up to individual communities to discuss how pre-prints emerging close to submission deadlines, or that arise during the review process, should be incorporated (if at all).

Consider requiring a budget and access statement. I was delighted when I recently came across a pre-print [4] that stated the costs of reproducing the results of research using an LLM. Similarly, it is worthwhile for authors to disclose if they had early access to a model before a general public release. Requiring budget and access statements is not meant to diminish the contributions groups that are well-resourced or have privileged relationships. Instead, this level of transparency can be useful for the VIS/HCI research community to reflect on the accessibility of LLM research. If it becomes prohibitively expensive to conduct research using LLMs, or consequential results rely on early access, then it may also be the case that only certain perspectives appear in the accepted literature.

Make better use of the desk reject. In HCI conferences where I have served, the decision to desk reject is not left solely to subcommittees or area chairs. The decision is part of the tasks of the program committee members. Area chairs can still screen out submissions that are incomplete or inappropriate. However, it would be the responsibility of the primary reviewers to flag manuscripts for desk rejection based on content or alignment with the submission venue. The primary also provides a brief justification for the choice to reject. Secondaries must decide if they agree with the decision to desk reject and should also provide a brief review. The final decision could still rest with the area chairs. Quickly removing research that is clearly not well aligned with the submission venue reduces overhead in the review process. LLM wrapper papers that do not cite *any* prior research in VIS/HCI are candidates for desk rejection.

Actively engage in the discussion period. Once the initial round of reviews is completed, everyone who reviewed the paper has about a week to consider other reviews and engage in an anonymous and asynchronous discussion. The discussion period is often underutilized, with some reviewers opting not to engage at all. In cases where work is clearly rejected by all, the discussion period can be quite short. The same may be true of papers that everyone chooses to clearly accept. For all other cases, which are the majority, the discussion period is important for evaluating common critiques of the manuscript, ironing out the differences, and providing meaningful feedback to the authors. Given the challenges of reviewing LLM papers, discussion periods are an important and underappreciated forum for giving good work a chance.

4 Conclusion

Large Language Models have come to constitute a sizable portion of papers submitted to VIS and HCI conferences. From my service on several program committees, I noted several recurring and valid concerns about research papers that incorporate LLMs. I have summarized these as a set of seven critiques and discussed my perception of their pros and cons. I suggest some concrete actions that can be taken to address some of these critiques and to ideally lower the burden to reviewers. However, the goal of this position paper is ultimately to stimulate conversation around the review process and, if it is possible, to generalize this reflection towards how our community responds to significant technological shifts.

References

[1] Improving Language Understanding by Generative Pre-Training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf, 2018. 1

[2] E. Adar. Applied AI Grants in the AI Summer. https://medium.com/@eytanadar/applied-ai-grants-in-the-ai-summer-62bd7414be74, 2024. 1, 2, 3

[3] I. Arawjo. LLM Wrapper Papers are Hurting HCI Research. https://ianarawjo.medium.com/llm-wrapper-papers-are-hurting-hci-research-8ad416a5d59a, 2024. 2

[4] S. Bordt, H. Nori, V. Rodrigues, B. Nushi, and R. Caruana. Elephants never forget: Memorization and learning of tabular data in large language models, 2024. 3

[5] A. Bruckman. Surviving the AI Summer. https://asbruckman.medium.com/surviving-the-ai-summer-64626e5547e3, 2024. 1, 3

[6] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2674–2693, 2019. doi: 10.1109/TVCG.2018.2843369 1

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, vol. 25, 2012. 1

[8] P. Maddigan and T. Susnjak. Chat2vis: Generating data visualizations via natural language using chatgpt, codex and gpt-3 large language models. *IEEE Access*, 11:45181–45193, 2023. doi: 10.1109/ACCESS.2023.3274199 3

[9] M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting, 2024. 3

[10] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023. 2

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, vol. 30, 2017. 1