chem-bla-ics

New paper: “ToxTempAssistant: using large language models to standardise cell-based toxicological test method descriptions”

Egon Willighagen

Published October 3, 2026

Citation
Willighagen, E. (n.d.). In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/v5ehk-42y77

Keywords
Fair, Llm, Vhp4safety, Openscience

Copyright
Copyright © Egon Willighagen 2026. Distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Jente Houweling published her first PhD thesis chapter earlier this year: "ToxTempAssistant: using large language models to standardise cell-based toxicological test method descriptions" (doi:10.1080/2833373X.2026.2638036). So far, I have been blogging about many of the articles on which I am (co-)author. To put it in context. To reflect on the work. I have been postponing writing about this paper because there is a lot to reflect on. I will pick out two things. First, I look at the use of AI. The second is the unique, innovatie publishing model of the journal where the article was published. If you want to just see it in action, ToxTempAssistant is running on the Virtual Human Platform for safety assessment.

Before we go there, just a quick note on what ToxTemps are and ToxTempAssistent actually is:

> *The ToxTemp template, based on OECD Guidance Document 211, standardises reporting for cell-based NAMs. However, completing its 77 questions constitutes a substantial bottleneck. The aim of this study is to introduce ToxTempAssistant, a Large Language Model (LLM)-assisted web tool that supports toxicologists in drafting ToxTemp documents based on user-supplied context documents. This study quantifies the tool's baseline performance under controlled conditions. ToxTempAssistant uses grounded, per-question prompting with mandatory source attribution.*

The LLM aspects

AI is very old. Arguably, my PhD thesis had this as key topic, though I prefered to use to term chemometric or machine learning. Critial thinking has been essential to this field for a long time, and much of the PhD thesis is actually about critically assessing the performance of the methods used in the thesis. There is decades of research how you do this. Sadly, when it comes to Large Language Models (LLMs), these are not routinely used.

LLMs are indeed something new. I have seen natural language processing research when I was Cambridge with Peter Murray-Rust. The current LLMs are different, less deterministic, more probabilistic. From a chemometrics perspective, that makes sense. Language is complex, and not so deterministic in itself. Moreover, when representing words, sentences as numbers, you can integrate any resource (think data tables, images, etc). Even more, digital representation was even more the central theme of my PhD thesis.

But because of the nature of the method, the nature of how the commercial, better known models are trained (and the impact on the notion of copyright), the impact on the climate emergency, the black box that these LLMs often are, there are so many technical, scientific, and ethical reasons to stay away from them.

You can write books about that. Literally (I did not read them yet):

- The Nerd Reich
- The AI Con

(I have the feeling I am missing one title I wanted to highlight. If I remember, I will add it.)

And more here and here. And I am looking forward to reading *Deep Unlearning: The Rise of AI and the Radicalization of a Tech Idealist*. Also, I recommend at least following Timnit Gebru and Emily Bender.

A year ago, I signed the Open Letter: Stop the Uncritical Adoption of AI Technologies in Academia, now signed by more than 2,000 people (it is not too late).

So, with all these things mind, I am happy that Jente did critically adopt LLMs in her work. The paper includes various experiments to explore the impact of various model parameters on generating nonsense. She also explored to use of alternative LLMs platforms, opening the option that some day it runs on other, more ethical platforms. ToxTempAssistant, moreover, limits the material it takes information from, from a limited set of sources, provided by the users. Of course, the model itself is still training on resources with questionable provenance.

Jente's paper describes the use of positive and negative controls to put the performance in perspective. In doing so, the paper formalizes how to evaluate the use of LLMs in situations where the LLM is used to summarize other reports, a common thing to do. This allows us to monitor the impact of, for example, new LLM releases.

The results are promising and various independent projects have shown interest in adoption. The ToxTemp reports are important for safety assessment: they provide essential context to experimental results and as such essential to the FAIR-ness of toxicology data.

Open Peer Review

The ToxTempAssistant paper is published in the relatively new journal Evidence-Based Toxicology (EBT):

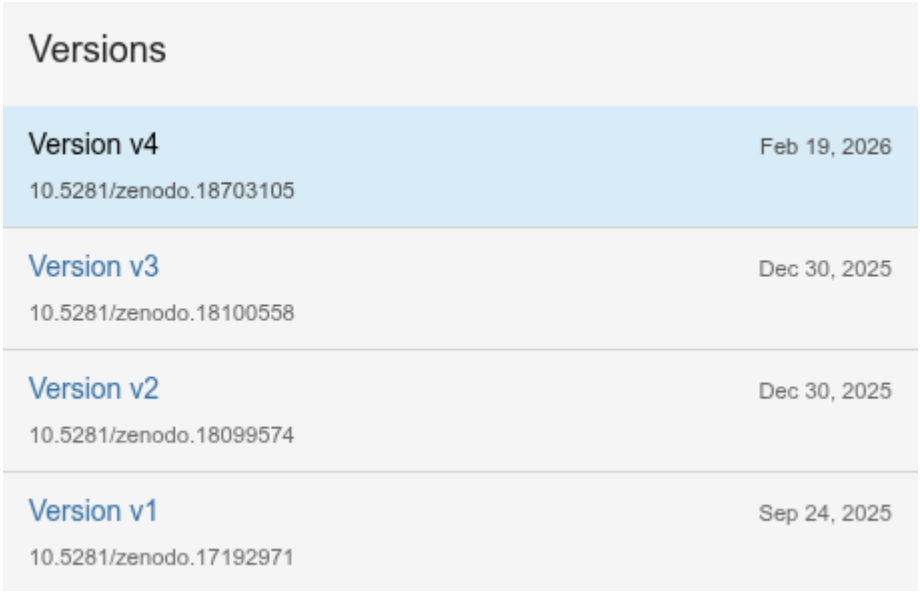
> *Evidence-Based Toxicology is a broad-focus, gold open-access journal, created to support the use of open science practices and evidence-based methods in toxicology and environmental health.*

So, CC-BY license (gold open access) and support for Open Science. And Open Peer Review, as we will see. Also, I undestand it is not a diamond open access journal, so expect APCs.

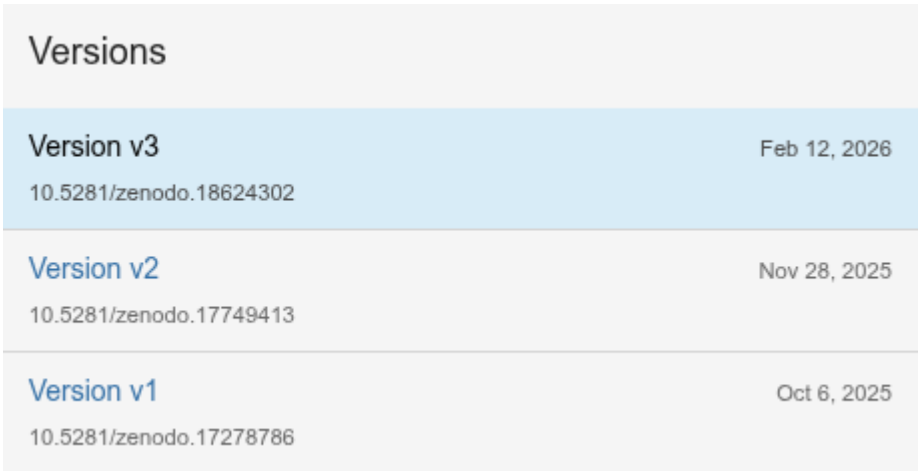
The journal has a community on Zenodo for preprints and peer-reviews. I think this is a really nice choice. Of course, a journal specific preprint server has downsides too. For example, if it gets rejected from EBT, do you use the EBT preprint server when submitting to another journal? Do you upload a new version (after all, you should address some of the comments why it was rejected) to another preprint server?

The EBT preprint gets a record and revisions are uploaded as new versions to the same Zenodo record (doi10.5281/zenodo.17192970):

chem-bla-ics



But because EBT uses open peer review, there is a parallel Zenodo entry with the reviews (doi10.5281/zenodo.17278785):



The last version here is the acceptance notice:

Acceptance Note for TEBT-2025-0014

Submission Title	**ToxTempAssistant: Using Large Language Models to Standardise Cell-Based Toxicity Test Method Descriptions**
Submission Reference	TEBT-2025-0014
Handling Editor	Paul Whaley | 0000-0003-4021-0785
Editor Declaration of Interests	The editor is not aware of any conflicts of interest that could compromise the handling of this submission.
Number of Reviewers	1 editor and 3 peer-reviewers
Submission Type	Research Article
Preregistration Level	N/A (Not Applicable) (explanation)
Preprint DOI	https://doi.org/10.5281/zenodo.17192970
Report Number	3
Evaluation Stage	Peer-Review
Evaluation Date	12 Feb 2026
Use of AI in evaluation	DataSeer.ai was used to perform data availability checks on the manuscript.
Recommendation	Accept

One of the reviewer suggestions was to use the TRIPOD-LLM template (doi:10.1038/s41591-024-03425-5), which was included in a later revision. That made a lot of sense and is a nice example how the journal actively works on applying open science ideas. The journal webpage writes:

> *We define "open science" as the set of practices aimed at improving the transparency, validity, reproducibility, and accessibility of scientific research, while promoting equality of opportunity to participate in and benefit from the products of said research.*

One comment here is that the template asks on what page that point of the checklist is discussed in the article. That is nice, but links it directly to the revision of the manuscript that form applies too, but that is not reported in the template. Not ideal. Then again, the point is the checking, so maybe not a big deal.

Another comment is that it seems the journal website's page for the article does not seem to actually link to the preprints nor the peer-review reports (or acceptance note). So, in time, these open science aspects of this article will likely get lost in time. Who will find those reports if the

article does not cite them? This will require Taylor&Francis to modernize the publishing model, and I sincerly doubt that that will ever happen.

Prof. Anne Kienhuis, one of the co-authors, suggested this journal lead by Prof. Paul Whaley. And I am happy to have seen this new approach in action and like to thank Paul for pushing for these innovations. I recommend trying it yourself for your next toxicology work.

References

Gallifant, J., Afshar, M., Ameen, S., Aphinyanaphongs, Y., Chen, S., Cacciamani, G., Demner-Fushman, D., Dligach, D., Daneshjou, R., Fernandes, C., Hansen, L. H., Landman, A., Lehmann, L., McCoy, L. G., Miller, T., Moreno, A., Munch, N., Restrepo, D., Savova, G., … Bitterman, D. S. (2025). The TRIPOD-LLM reporting guideline for studies using large language models. *Nature Medicine*, *31*(1), 60–69. https://doi.org/10.1038/s41591-024-03425-5 **[cito:discusses]** CrossRef Scholia

Houweling, J. (Johanne) M., Arras, M. M., Willighagen, E., Kienhuis, A., Jennen, D., & Evelo, C. (2026). ToxTempAssistant: Using Large Language Models to Standardise Cell-Based Toxicity Test Method Descriptions. *Zenodo*. https://doi.org/10.5281/ZENODO.17192970 **[cito:citesAsEvidence]** CrossRef Scholia

Houweling, J. M., Arras, M. M. L., Willighagen, E. L., Jennen, D. G. J., Evelo, C. T., & Kienhuis, A. S. (2026). *ToxTempAssistant* : using large language models to standardise cell-based toxicological test method descriptions. *Evidence-Based Toxicology*, *4*(1). https://doi.org/10.1080/2833373x.2026.2638036 **[cito:discusses]** CrossRef Scholia

Krebs, A., Waldmann, T., Wilks, M. F., van Vugt-Lussenburg, B. M. A., van der Burg, B., Terron, A., Steger-Hartmann, T., Ruegg, J., Rovida, C., Pedersen, E., Pallocca, G., Luijten, M., Leite, S. B., Kustermann, S., Kamp, H., Hoeng, J., Hewitt, P., Herzler, M., Hengstler, J. G., … Leist, M. (2019). https://www.altex.org/index.php/altex/article/view/1339. *ALTEX*, 682–699. https://doi.org/10.14573/altex.1909271 **[cito:citesForInformation]** CrossRef Scholia

Whaley, P. (2026). *Evaluation Reports for TEBT-2025-0014 | Houweling et al.* Zenodo. https://doi.org/10.5281/ZENODO.17278785 **[cito:citesAsEvidence]** CrossRef Scholia