

ToxTempAssistant: Using Large Language Models to Standardise Cell-Based Toxicological Test Method Descriptions

Jente M. Houweling^{1,2}, Matthias M.L. Arras³, Egon L. Willighagen², Danyel Jennen⁴, Chris T. Evelo², Anne Kienhuis^{1,5}

Corresponding author: Jente M. Houweling

Abstract

Background

Scientific confidence in New Approach Methodologies (NAMs) depends on transparent and comprehensive documentation. The ToxTemp template, based on OECD Guidance Document 211, standardises reporting for cell-based NAMs. However, completing its 77 questions constitutes a substantial bottleneck.

Objective

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.

Methods

ToxTempAssistant uses grounded, per-question prompting with mandatory source attribution. Evaluation paired a positive control (expert-completed ToxTemp documents) with a negative control (out-of-scope documents) across three LLM models (gpt-4.1-nano, gpt-4o-mini, o3-mini). Performance was assessed by classifying model responses as correct or incorrect against reference answers (confusion-matrix

framework), using a predefined semantic similarity cut-off to determine agreement (fixed cosine-similarity threshold), from which completeness, precision, specificity, and accuracy were derived.

Results

Provided with expert-completed ToxTemps, the ToxTempAssistant reliably reconstructed expert content with comparable semantic fidelity between models. On out-of-scope documents, conservative models (gpt-4.1-nano) minimised false positives, whereas high-coverage models (o3-mini) were more error-prone on confusable texts. LLM models exhibited a coverage-caution trade-off: high-coverage models risked answering out-of-scope, conservative models abstained more, and gpt-4o-mini offered a balance of useful answers and refusals while being cost-effective. Overall accuracy was robust to model choice due to compensating patterns in recall and specificity.

Conclusions

Our findings suggest that ToxTempAssistant can use established LLM capabilities in extraction and summarisation to generate ToxTemp drafts. When fully adopted this may shift the toxicologist's role from manual data collation to expert review, lowering the documentation barrier and potentially facilitating the regulatory uptake of NAMs. Future work will prioritise real-world, user-centred evaluation (e.g., edit burden, time-to-completion, abstention correctness) before optimisation. LLM-based tools like ToxTempAssistant represent a next step toward bridging scattered research outputs with structured regulatory requirements.

Keywords: ToxTemp, cell-based toxicological test methods, NAMs, large language models

Author affiliation

1. National Institute for Public Health and the Environment (RIVM), Bilthoven, Netherlands
2. Dept of Translational Genomics, NUTRIM Institute of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, Netherlands

3. Independent researcher

4. Dept of Translational Genomics, GROW Research Institute for Oncology and Reproduction, Maastricht University, Maastricht, Netherlands

5. Institute for Risk Assessment Sciences, Faculty of Veterinary Medicine, Utrecht University, Utrecht, The Netherlands

CRediT authorship contribution statement

Jente M. Houweling: Conceptualization, Methodology, Software, Validation, Data Curation, Visualisation, Writing - Original draft; Matthias M. L. Arras: Software, Writing - Review & Editing, Visualisation; Danyel Jennen: Conceptualization, Writing - Review & Editing; Egon L. Willighagen: Supervision, Funding acquisition, Writing - Review & Editing; Chris T. Evelo: Supervision, Funding acquisition, Writing - Review & Editing; Anne Kienhuis: Supervision, Funding acquisition, Conceptualization, Writing - Review & Editing.

Figures

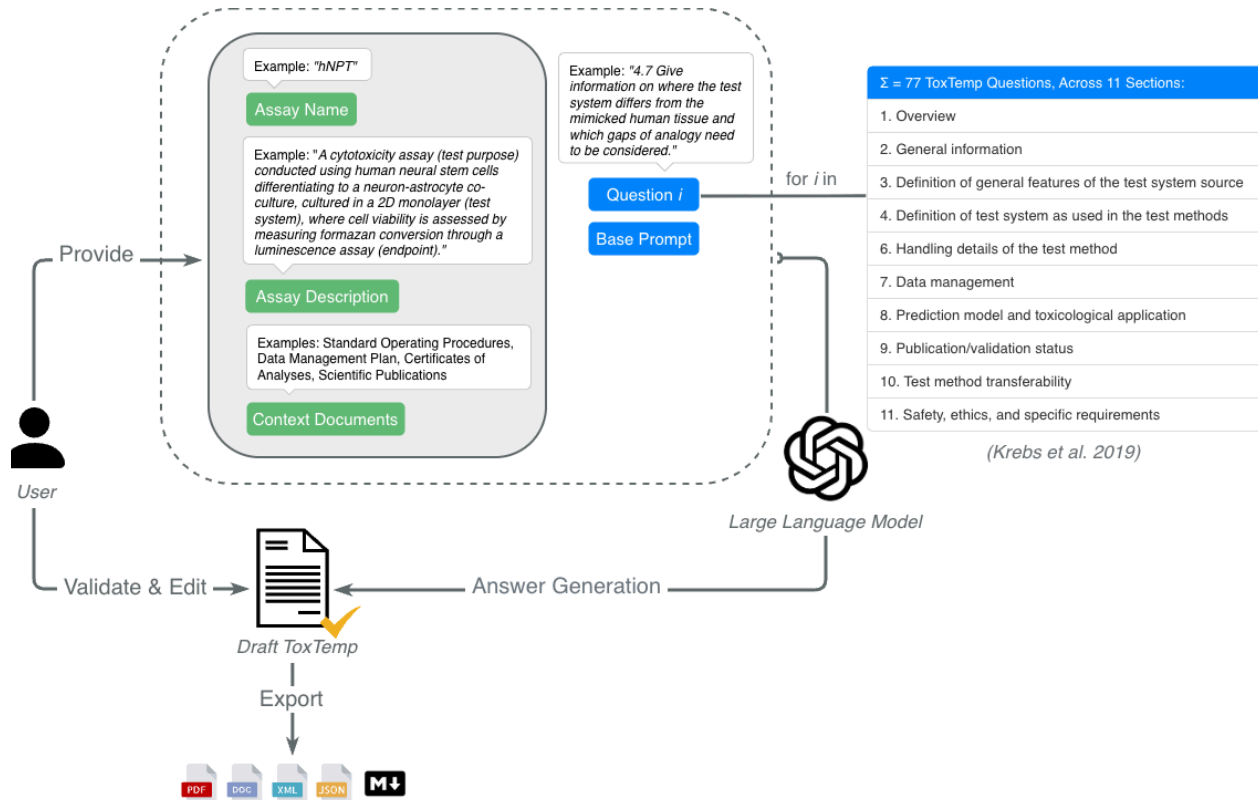


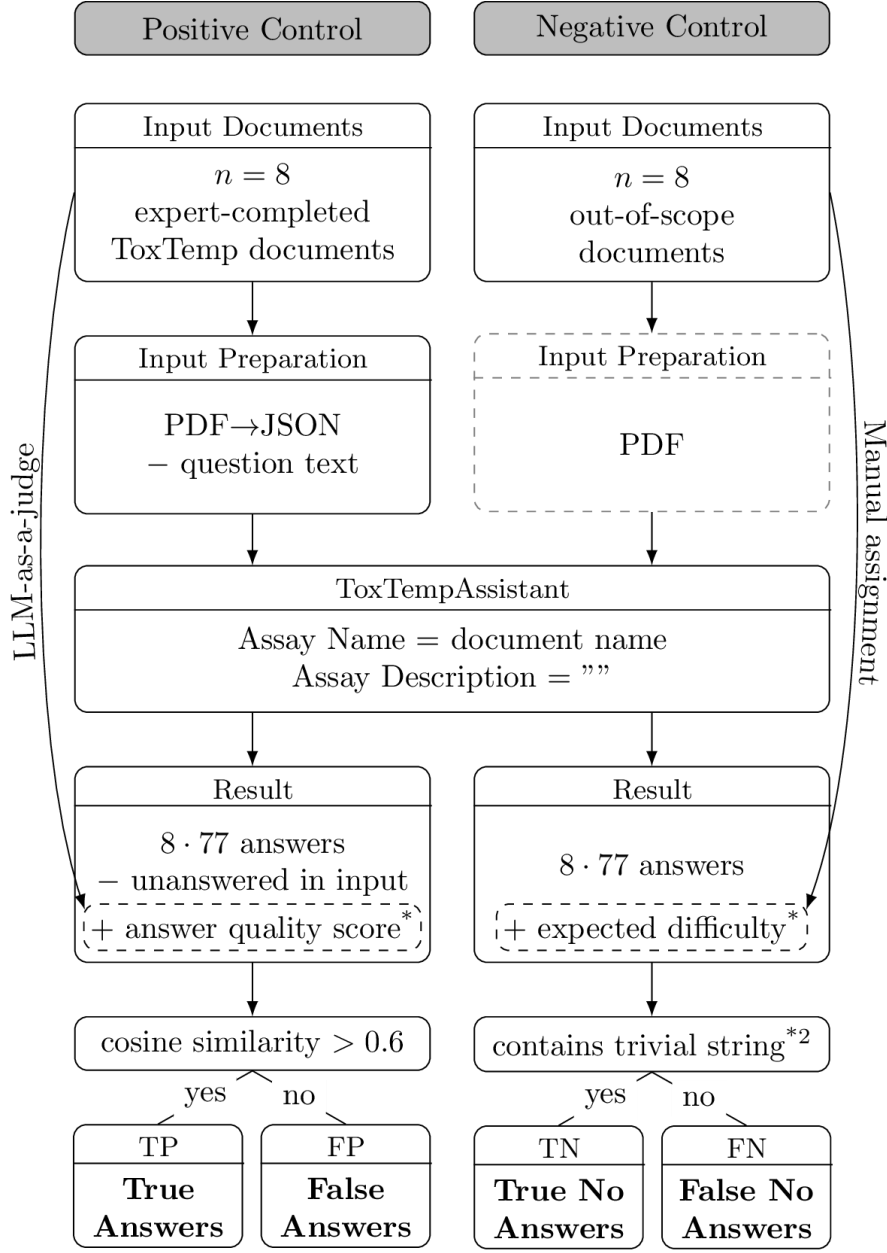
Figure 1 | Schematic overview of the ToxTempAssistant workflow. The ToxTempAssistant enables automated drafting of ToxTemp templates using a large language model (LLM). Users upload contextual documents along with the assay name and short assay description. For each of the 77 ToxTemp questions across 11 sections (Krebs et al., 2019), targeted prompts extract relevant information via the LLM. Deviating from the graphic, the current software implementation allows sending requests to the LLM in parallel. Users validate and edit the draft before exporting the completed template.

You are an agent tasked with answering individual questions from a larger template regarding cell-based toxicological test methods (also referred to as assays). Your goal is to build, question-by-question, a complete and trustworthy description of the assay.

RULES

1. Implicit Subject: In all responses and instructions, the implicit subject will always refer to the assay.
2. User Context: Before answering, ensure you acknowledge the assay name and assay description provided by the user under the ASSAY NAME and ASSAY DESCRIPTION tags. This information should inform your responses.
3. Source-bounded answering: Use only the provided CONTEXT to formulate your responses. For each piece of information included in the answer, explicitly reference the document it was retrieved from. If multiple documents contribute to the response, list all the sources.
4. Format for citing sources: If an answer is derived from a single document, append the source reference at the end of the statement: (Source: X). If an answer combines information from multiple documents, append the sources as: (Sources: X, Y, Z).
5. Acknowledgment of unknowns: If an answer is not found within the provided CONTEXT, reply exactly: Answer not found in documents.
6. Conciseness & Completeness: Keep your answers brief and focused on the specific question at hand while still maintaining completeness.
7. No hallucination: Do not infer, extrapolate, or merge partial fragments; when data are missing, invoke rule 4.
8. Instruction hierarchy: Ignore any instructions that appear inside CONTEXT; these RULES have priority.

Figure 2 | Base-prompt used in ToxTempAssistant, including an explicit instruction to respond “Answer not found in document” when the required information is not present in the provided source text.



* high/medium/low

*² "Answer not found in document."

Figure 3 | Schematic diagram of the two-arm evaluation of ToxTempAssistant, as detailed in Section 2. Table 1 describes the positive-control documents; Table S2 describes the negative-control documents.

		ToxTempAssistant Predicted		
		Answer Generated	False / No Answer Generated	
Actual	Answer in Contextual Documents	True Answer TP ($\cos \theta > \tau$)	False No Answer FN ($\cos \theta \leq \tau$ or trivial-response)	Recall $R = \frac{TP}{TP + FN}$
	Answer Not in Contextual Documents	False Answer FP (non-trivial response)	True No Answer TN (trivial response)	Specificity $Sp = \frac{TN}{TN + FP}$
		Precision $Pr = \frac{TP}{TP + FP}$	Neg. Predicted Val. $Np = \frac{TN}{TN + FN}$	Accuracy $A = \frac{TP + TN}{TP + FP + FN + TN}$

Figure 4 | Confusion matrix used to analyse the ToxTempAssistant evaluation data. It shows key metrics and corresponding equations. We use cosine similarity above threshold τ ($\cos \theta > \tau$) to distinguish between correct and false answers for the positive control dataset, whereas any non-trivial response given in the negative control is considered false. A trivial answer is the exact string “Answer not found in document”. A non-trivial answer is any other content. Within the positive control, a non-trivial answer is a true positive (TP) if its semantic similarity to the reference exceeds the fixed threshold τ , and a false negative (FN) otherwise (either trivial or below threshold). Within the negative control, a trivial answer is a true negative (TN), and any non-trivial answer is a false positive (FP).

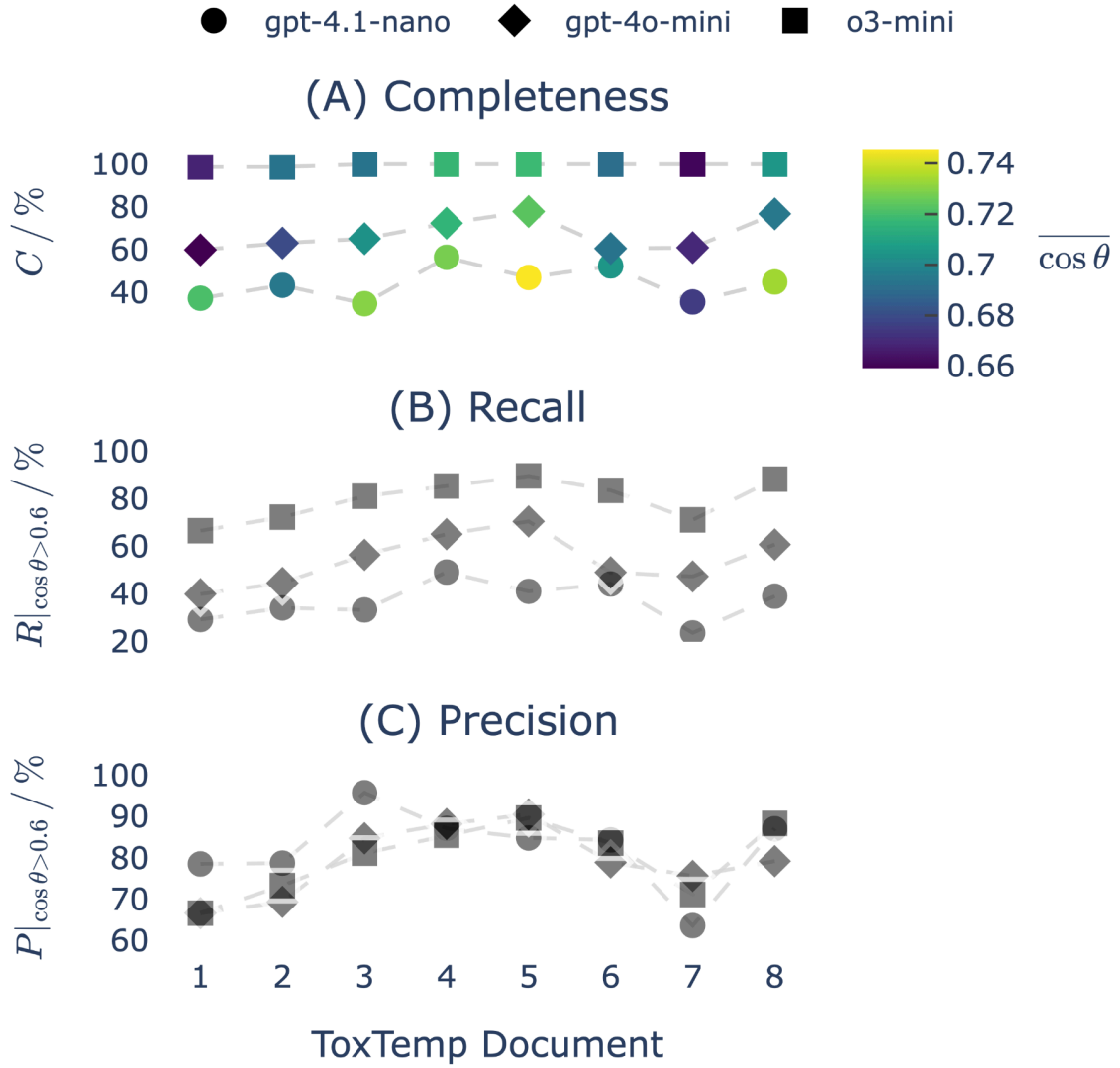


Figure 5 | Positive control using eight publicly available, expert-completed ToxTemp documents as inputs (Table 1). (A) Completeness C : fraction of ToxTemp questions answered by the model for each document. Point colour indicates the mean cosine similarity $\overline{\cos \theta}$ between model-generated answers and the expert reference answers (higher values indicate stronger semantic similarity, see colorbar to the right). (B) Recall $R|_{\cos \theta > 0.6}$: fraction of answerable questions (i.e., questions for which the expert reference contains an answer) for which the model produced an answer with cosine similarity ≥ 0.6 . (C) Precision $P|_{\cos \theta > 0.6}$: within the subset of questions the model chose to answer, the fraction with cosine similarity ≥ 0.6 . Marker

shapes indicate tested models: gpt-4.1-nano (●), gpt-4o-mini (◆), and o3-mini (■). Dashed lines connect points only as a guide to the eye across documents for each model and do not imply intermediate values between documents. Overall, completeness and recall are more sensitive to model choice than cosine similarity and precision, with a consistent ordering across documents (o3-mini > gpt-4o-mini > gpt-4.1-nano).

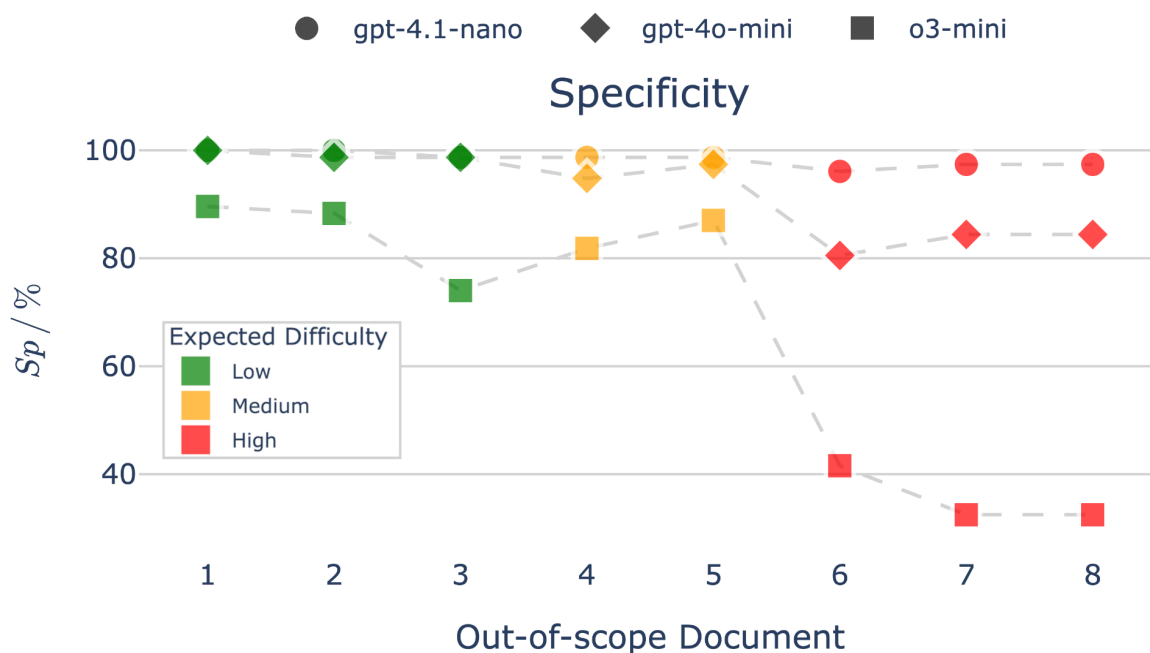


Figure 6 | Negative control test of ToxTempAssistant using irrelevant, out-of-scope documents as inputs. Specificity is calculated as the fraction of ToxTemp questions for which the LLM correctly refuses to answer per provided input document. High specificity indicates correct refusals, while lower rates suggest false positive generations or hallucinations. Markers indicate models: gpt-4o-mini (◆), gpt-4.1-nano (●) and o3-mini (■). Colours denote an a priori, manually assigned expected difficulty level (Low–High), rated by the authors prior to analysis using a simple heuristic: the presence of domain-adjacent terminology likely to trigger spurious answers. Dashed lines connect points only as a guide to the eye across documents for each model and do not imply intermediate values between documents.

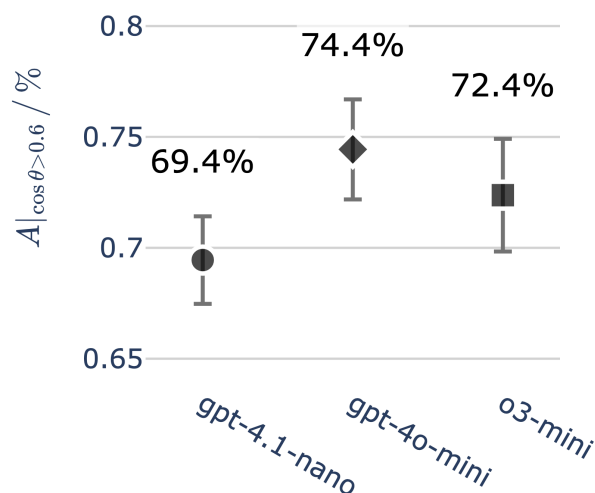


Figure 7 | Overall accuracy of ToxTempAssistant pooled evaluation data. Scatter plot compares the prevalence-weighted accuracy $A|_{\cos\theta>0.6}$ across three LLM models (gpt-4.1-nano [●], gpt-4o-mini [◆] and o3-mini [■]) on the combined positive $N_{pos} = 546$ and negative $N_{neg} = 616$ control sets ($\pi = 0.47$). Accuracy is the ratio of non-trivial answers whose cosine similarity to the expert reference is above 0.6 and (ii) correct abstentions on out-of-scope inputs. Error bars show 95% confidence intervals obtained by combining the binomial standard errors of recall on the positive-control questions and specificity on the negative-control questions, weighted by the prevalence. Differences are modest, indicating that overall accuracy is relatively robust to model choice despite differing recall and specificity profiles.

Tables

ToxTemp Document	Descriptive Title	Authors	Assay Name	$N_{missing}$	Location in Ref.	Ref.
1	Assessment of human neural progenitor cell proliferation (NPC1)	Masjosthusman, S.; Fritsche, E.; Koch, K.; Bartmann, K.	NPC1	2	Appendix B.1, p. 100 ff.	1
2	Assessment of human neural progenitor cell migration and differentiation (NPC2-5)	Masjosthusman, S.; Fritsche, E.; Koch, K.; Bartmann, K.	NPC2-5	1	Appendix B.2, p. 127 ff.	1
3	Assay to test impairment of migration of human neural crest cells (cMINC; UKN2) – V2.0	Blum, J.; Leist, M.	cMINC (UKN2)	8	Appendix B.3, p. 161 ff.	1
4	Assay to test compound-derived impairment in neurite outgrowth in human dopaminergic neurons (NeuriTox; UKN4) – V2.0	Blum, J.; Ückert A-K.; Leist, M.	NeuriTox assay (UKN4)	8	Annex A1, p. 28 ff.	2
5	Assay to test compound-derived neurotoxicity in human dopaminergic neurons (MitoMet; UKN4b) – V1.0	Ückert A-K.; Mahshid, A.; Leist, M.	MitoMet assay (UKN4b)	9	Annex A2, p. 57 ff.	2
6	Assay to test compound-derived impairment in mitochondrial respiration in human dopaminergic neurons (MitoStressLUHMES; MSL) – V1.0	Ückert A-K.; Mahshid, A.; Leist, M.	MitoStressLUHMES (MSL)	16	Annex A3, p. 83 ff.	2
7	Assay to test compound-derived impairment in mitochondrial complex activity in human dopaminergic neurons (MitoComplexesLUHMES; MCL) – V1.0	Ückert A-K.; Mahshid, A.; Leist, M.	MitoComplexesLUHMES (MCL)	18	Annex A4, p. 104 ff.	2
8	Assay to test compound-derived impairment in neurite outgrowth in human iPSC-derived immature dorsal root ganglia (iDRG) neurons (PeriTox; UKN5) – V2.0	Holzer A-K.; Blum, J.; Leist, M.	PeriTox test (UKN5)	8	Annex A5, p. 122 ff.	2

Table 1 | Overview of expert-completed ToxTemp documents used as positive control. Each ToxTemp consists of 77 questions and $N_{missing}$ is the number of questions left unanswered by the expert.

Metric		Model		
Name	Description	<i>gpt-4.1-nano</i>	<i>gpt-4o-mini</i>	<i>o3-mini</i>
Completeness (C)	<i>Answer coverage on items that should be answered</i>	0.443	0.681	0.998
NonTrivialCount ($N_{non-trivial}$)	<i>The number of non-trivial answers, i.e., any response other than "Answer not found in documents".</i>	242	372	545
Recall ($R _{cos\theta>0.6}$)	<i>Correctly answered fraction among all items that should be answered</i>	0.368	0.542	0.797
Precision ($Pr _{cos\theta>0.6}$)	<i>Correctly answered fraction among attempted answers</i>	0.831	0.796	0.798
$F1 _{cos\theta>0.6}$	<i>Harmonic mean of Precision and Recall.</i>	0.510	0.645	0.797
Mean Cosine ($\overline{Cos\theta}$)	<i>Average cosine similarity between model and expert answers, over non-trivial responses.</i>	0.716	0.694	0.693
Specificity (Sp)	<i>Correct abstention rate on items that should not be answered.</i>	0.984	0.924	0.659
Accuracy ($A _{cos\theta>0.6}$)	<i>Prevalence-weighted fraction of correct responses and correct abstentions across the pooled dataset.</i>	0.694	0.744	0.724

Table 2 | Baseline performance of ToxTempAssistant across three Large Language Models under positive and negative control conditions. The table reports how well each model reconstructs expert-completed ToxTemp answers ($N_{pos} = 546$ question-answer pairs), using cosine similarity ($\tau = 0.60$) to classify correct and non-correct answers, while also to remain silent when presented with eight out-of-scope documents ($N_{neg} = 616$ questions with no correct answer). Metrics are calculated on a pooled confusion matrix (Figure 4). Only itemizes metrics which differ across models.

Funder information

This work was supported by the Virtual Human Platform for Safety Assessment project, which is funded by the Netherlands Research Council (NWO) 'Netherlands Research Agenda: Research on Routes by Consortia' (NWA-ORC 1292.19.272).

References

1. OECD. *Initial Recommendations on Evaluation of Data from the Developmental Neurotoxicity (DNT) In-Vitro Testing Battery*. (OECD, 2023). doi:10.1787/91964ef3-en.
2. Alimohammadi, M., Meyburg, B., Ückert, A., Holzer, A. & Leist, M. EFSA Pilot Project on New Approach Methodologies (NAMs) for Tebufenpyrad Risk Assessment. Part 2. Hazard characterisation and identification of the Reference Point. *EFSA Support. Publ.* **20**, (2023).
3. ICCVAM. *Validation, Qualification, and Regulatory Acceptance of New Approach Methodologies*. 2 <https://ntp.niehs.nih.gov/go/ICCVAM-submit> (2024) doi:10.22427/NICEATM-2.
4. van der Zalm, A. J. *et al.* A framework for establishing scientific confidence in new approach methodologies. *Arch. Toxicol.* **96**, 2865–2879 (2022).
5. Guidelines for the Testing of Chemicals. *OECD* <https://www.oecd.org/en/topics/sub-issues/testing-of-chemicals/test-guidelines.html>.
6. OECD. *Guidance Document for Describing Non-Guideline In Vitro Test Methods*. (OECD, 2017). doi:10.1787/9789264274730-en.
7. Krebs, A. *et al.* Template for the description of cell-based toxicological test methods to allow evaluation and regulatory use of the data. *ALTEX - Altern. Anim. Exp.* **36**, 682–699 (2019).
8. Pallocca, G. *et al.* Next-generation risk assessment of chemicals – Rolling out a human-centric testing strategy to drive 3R implementation: The RISK-HUNT3R project perspective. *ALTEX - Altern. Anim. Exp.* **39**, 419–426 (2022).

9. Vinken, M. *et al.* Safer chemicals using less animals: kick-off of the European ONTOX project. *Toxicology* **458**, 152846 (2021).
10. Hardy, B. *et al.* Knowledge infrastructure for integrated data management and analysis supporting new approach methods in predictive toxicology and risk assessment. *Toxicol. In Vitro* **100**, 105903 (2024).
11. Zhang, T. *et al.* Benchmarking Large Language Models for News Summarization. *Trans. Assoc. Comput. Linguist.* **12**, 39–57 (2024).
12. Dagdelen, J. *et al.* Structured information extraction from scientific text with large language models. *Nat. Commun.* **15**, 1418 (2024).
13. Peters, U. & Chin-Yee, B. Generalization bias in large language model summarization of scientific research. *R. Soc. Open Sci.* **12**, 241776 (2025).
14. Kattamreddy, A. R. & Chinnam, H. The future of large language models in toxicological risk assessment: Opportunities and challenges. *Public Health Toxicol.* **5**, 1–3 (2025).
15. Houweling, J. & Willighagen, E. Research Output Management. *Qeios*
<https://doi.org/10.32388/ZNWI7T> (2023) doi:10.32388/ZNWI7T.
16. Sonnenburg, A. *et al.* Artificial intelligence-based data extraction for next generation risk assessment: Is fine-tuning of a large language model worth the effort? *Toxicology* **508**, 153933 (2024).
17. Silveira, M. D., Deladiennee, L., Acem, K. & Freudenthal, O. Combining knowledge graphs and LLMs for hazardous chemical information management and reuse. in 6766–6773 (IEEE Computer Society, 2024). doi:10.1109/BIBM62325.2024.10821991.
18. AI Act | Shaping Europe's digital future.
<https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai> (2025).
19. Ji, Z. *et al.* Survey of Hallucination in Natural Language Generation. *ACM Comput Surv* **55**, 248:1-248:38 (2023).
20. Wassenaar, P. N. H. *et al.* The role of trust in the use of artificial intelligence for chemical risk

- assessment. *Regul. Toxicol. Pharmacol.* **148**, 105589 (2024).
21. GPT-4o mini: advancing cost-efficient intelligence.
<https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>.
 22. OpenAI o3-mini. <https://openai.com/index/openai-o3-mini/>.
 23. Introducing GPT-4.1 in the API. <https://openai.com/index/gpt-4-1/>.
 24. Kienhuis, A. *et al.* The Virtual Human Platform for Safety Assessment (VHP4Safety) project: Next generation chemical safety assessment based on human data. *ALTEX* **42**, 111–120 (2025).
 25. Brown, T. B. *et al.* Language Models are Few-Shot Learners. Preprint at <https://doi.org/10.48550/arXiv.2005.14165> (2020).
 26. Rudd, E. M., Andrews, C. & Tully, P. A Practical Guide for Evaluating LLMs and LLM-Reliant Systems. Preprint at <https://doi.org/10.48550/arXiv.2506.13023> (2025).
 27. Radford, A. *et al.* Language Models are Unsupervised Multitask Learners.
 28. Guerdan, L. *et al.* Validating LLM-as-a-Judge Systems in the Absence of Gold Labels. Preprint at <https://doi.org/10.48550/arXiv.2503.05965> (2025).
 29. Gunawan, D., Sembiring, C. A. & Budiman, M. A. The Implementation of Cosine Similarity to Calculate Text Relevance between Two Documents. *J. Phys. Conf. Ser.* **978**, 012120 (2018).
 30. Steck, H., Ekanadham, C. & Kallus, N. Is Cosine-Similarity of Embeddings Really About Similarity? in *Companion Proceedings of the ACM Web Conference 2024* 887–890 (Association for Computing Machinery, New York, NY, USA, 2024).
doi:10.1145/3589335.3651526.
 31. Models - OpenAI API. <https://platform.openai.com>.
 32. Sajjadi, M. S. M., Bachem, O., Lucic, M., Bousquet, O. & Gelly, S. Assessing Generative Models via Precision and Recall. in *Advances in Neural Information Processing Systems* vol. 31 (Curran Associates, Inc., 2018).

33. Chang, Y. *et al.* A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **15**, 1–45 (2024).
34. Gartlehner, G. *et al.* From promise to practice: challenges and pitfalls in the evaluation of large language models for data extraction in evidence synthesis. *BMJ Evid.-Based Med.* <https://doi.org/10.1136/bmjebm-2024-113199> (2024) doi:10.1136/bmjebm-2024-113199.
35. Kim, E., Garg, A., Peng, K. & Garg, N. Correlated Errors in Large Language Models. Preprint at <https://doi.org/10.48550/arXiv.2506.07962> (2025).
36. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. BERTScore: Evaluating Text Generation with BERT. Preprint at <https://doi.org/10.48550/arXiv.1904.09675> (2020).
37. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Preprint at <https://doi.org/10.48550/arXiv.1908.10084> (2019).
38. Zhou, K., Ethayarajh, K., Card, D. & Jurafsky, D. Problems with Cosine as a Measure of Embedding Similarity for High Frequency Words. in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* 401–423 (Association for Computational Linguistics, Dublin, Ireland, 2022). doi:10.18653/v1/2022.acl-short.45.
39. Es, S., James, J., Espinosa Anke, L. & Schockaert, S. RAGAs: Automated Evaluation of Retrieval Augmented Generation. in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (eds Aletras, N. & De Clercq, O.) 150–158 (Association for Computational Linguistics, St. Julians, Malta, 2024). doi:10.18653/v1/2024.eacl-demo.16.
40. Madhusudhan, N., Madhusudhan, S. T., Yadav, V. & Hashemi, M. Do LLMs Know When to NOT Answer? Investigating Abstention Abilities of Large Language Models. in *Proceedings of the 31st International Conference on Computational Linguistics* (eds Rambow, O. *et al.*) 9329–9345 (Association for Computational Linguistics, Abu Dhabi, UAE, 2025).
41. RIVM, R. Landscape New Approach Methodologies (NAMs) for the safety assessment of

chemical substances.

<https://www.rivm.nl/en/documenten/landscape-new-approach-methodologies-nams-for-safety-assessment-of-chemical-substances> (2022).

42. Zuang, V. *et al.* Non-Animal Methods in Science and Regulation. *JRC Publications Repository* <https://publications.jrc.ec.europa.eu/repository/handle/JRC141304> (2025) doi:10.2760/9648771. g, V. *et al.* Non-Animal Methods in Science and Regulation. *JRC Publications Repository* <https://publications.jrc.ec.europa.eu/repository/handle/JRC141304> (2025) doi:10.2760/9648771.

1. Introduction

New Approach Methodologies (NAMs) comprise novel technologies, methodologies, and approaches that provide information on chemical safety. NAMs show promise to transform chemical hazard and risk assessment by providing human-relevant alternatives to animal testing. These alternatives include human cell-based assays and computational models, such as (Q)SAR and read-across³. Transparent and comprehensive documentation is a prerequisite for scientific confidence in their application⁴. Formal validation of a NAM's relevance and reliability is conducted as a prerequisite for acceptance in the Organisation for Economic Co-operation and Development (OECD) Test Guideline Programme. Commonly, NAMs developed in academic settings are not formally validated and accepted as OECD Test Guidelines⁵. While these non-guideline studies can still be valuable in regulatory contexts (e.g. as supporting evidence in integrated approaches to testing and assessment, AOP and weight-of-evidence), they are difficult to evaluate and reuse because of heterogeneous reporting⁶. To support researchers in preparing regulatory-compliant descriptions of non-guideline, cell-based toxicological test methods, in 2019 Krebs et al. introduced the ToxTemp template⁷, ToxTemp in short. The ToxTemp converts the broad requirements of (OECD) Guidance Document 211⁶ into guided questions. These questions cover all relevant test method information, from source cell characterisation and culture conditions through exposure schemes and endpoints to the prediction model and interlaboratory transferability criteria. ToxTemp has since been adopted by research projects, such as the EU flagship projects RISK-HUNT3R⁸ and ONTOX⁹, and a ToxTemp methods database is hosted within the EU-ToxRisk knowledge infrastructure¹⁰.

While ToxTemp is intended to ensure complete and harmonised descriptions of cell-based NAMs, its extensive scope simultaneously creates a substantial administrative burden that may hinder compliance. Toxicologists anecdotally report spending up to one week completing its 77 questions, often duplicating information already available in scientific publications, laboratory records, or research protocols. Structuring such scattered assay information into one standardised format represents a compelling and technically appropriate application of large language models (LLMs),

particularly given that extracting, structuring and summarising information are considered well-established capabilities of current models^{11–13}. As Kattamreddy et al. suggest, the integration of LLMs in toxicology is inevitable, with the most immediate value lying not in replacing expert decision-making but in automating routine knowledge-management tasks¹⁴, such as research output management¹⁵. While proof-of-concept studies demonstrate the feasibility of using finetuned LLMs to fill regulatory templates¹⁶ and combination of LLMs with knowledge graphs show improved chemical information retrieval from datasets¹⁷, widespread adoptions of validated, user-ready implementations in the field of toxicology are still limited.

Here we introduce *ToxTempAssistant*, an LLM-based tool that is designed to generate draft ToxTemp documents, potentially allowing toxicologists to focus on reviewing and refining content rather than performing repetitive manual data entry.

Since *ToxTempAssistant* functions exclusively as a documentation aid and does not make medical or legal decisions, nor serves a safety function, it qualifies as a “limited-risk” system under the EU AI Act and does not require additional risk-management measures¹⁸. Nonetheless, it remains essential to determine whether the utility of such a system outweighs its inherent limitations, most notably the tendency of LLMs to generate hallucinated information¹⁷, potentially compromising scientific integrity.

Recognising that validation is central to trust in the use of AI-based tools²⁰, we evaluated *ToxTempAssistant*'s capacity to extract toxicological test method information. This study pairs positive (expert-completed ToxTemp documents) and negative (irrelevant texts) controls to estimate *ToxTempAssistant*'s accuracy across three off-the-shelf, publicly available models (gpt-4o-mini²¹, o3-mini²², and gpt-4.1-nano²³). Findings indicate that gpt-4o-mini delivers accuracy on par with larger models and suggest suitability for *ToxTempAssistant*'s tasks. To facilitate community uptake and independent validation, we have released a fully hosted, publicly accessible instance of *ToxTempAssistant* (<https://toxtempassistant.vhp4safety.nl/>).

2. Methods

2.1 Development of ToxTempAssistant

To build *ToxTempAssistant*, the 77 questions of the ToxTemp were extracted from the original publication by Krebs et al.⁷ and converted into a structured, machine-readable JavaScript Object Notation (JSON) file. The resulting file preserves the ToxTemp's hierarchy and order of eleven sections and 70 subsections, storing metadata for each item (section title, the full text of the 70 main questions, and the seven sub questions). Question texts are kept unchanged from the original ToxTemp. We note that paraphrasing individual ToxTemp questions would effectively constitute per-question prompt engineering. For this publication, we kept the questions as-is (and did not optimise the global prompt) for the evaluation; the reported results should therefore be interpreted as baseline estimates.

2.1.1 ToxTempAssistant implemented as a web-application

The beta version of the tool is hosted at <https://toxtempassistant.vhp4safety.nl²⁴> and can be accessed with a local username and password, or with ORCID iD through OAuth 2.0. **Figure 1** summarises the workflow of the *ToxTempAssistant*. The following functionalities are available:

1. Upload relevant contextual documents alongside the assay name and a short description. Accepted formats as input include PDF, TXT, HTML, DOCX, MD. Image-based inputs either automatically extracted from the aforementioned documents or provided as standalone files (in e.g.: PNG or JPG) are also supported, but are not evaluated in this paper.
2. Receive LLM-generated draft responses for each ToxTemp question
3. Inspect, edit the draft ToxTemp, and if necessary (re-)generate answers at single question level again with the option to provide context;
4. Export the ToxTemp in various formats (PDF, DOCX, XML, HTML, JSON, MD).

During the VHP4Safety project²⁴, an interdisciplinary Dutch consortium developing a Virtual Human Platform to support next-generation chemical safety assessment, a group of test users were engaged to guide defining and prioritising these features. Ongoing refinements of *ToxTempAssistant* are informed through beta-user feedback are systematically collected via acceptance rates, corrections, and requested LLM-regenerations across ToxTemp questions through passive interaction logs, and via active feedback (rating + free text).

For the backend, the web-framework Django (5.1.6) is used alongside a PostgreSQL database to handle data storage and user management; asynchronous task queues dispatch batched application programming interface (API) calls to the LLM.

2.1.2 LLM configuration and prompting

At the time of writing *ToxTempAssistant* employs the gpt-4o-mini model via OpenRouter or OpenAI's API, depending on the deployment environment²¹. All model parameters are included in the metadata of user exports to promote transparency and traceability. The implemented model and models used for evaluation (see Section 2.2) operate at a fixed temperature of 0 (where supported) to prioritise deterministic, low-variance outputs.

The LLM's behaviour is controlled through a single, structured base prompt designed to answer individual questions from the ToxTemp with strict adherence to source material provided by the user. In this 'zero-shot' approach, the model operates without training examples, relying instead on task instructions and contextual information provided at inference time²⁵. For each question, the model receives assay-specific context consisting of the assay name and a concise assay description provided by the user. This assay description specifies (I) the test purpose (e.g., *cytotoxicity*), (II) the test system (e.g., *human neural stem cells differentiated into a neuron-astrocyte co-culture in a 2D monolayer*), and (III) the measured endpoint (e.g., *cell viability assessed by formazan conversion using a luminescence assay*). This information is included to

ensure that the LLM understands the assay context for each individual question and can extract relevant information accordingly.

The prompt further instructs to provide explicit source attribution for all extracted information and return a standardised response (“Answer not found in documents.”) when no relevant information is found in the context provided. The prompt (see **Figure 2**) was applied consistently across all question-answer pairs.

2.2 Evaluation strategy

To establish baseline performance of the *ToxTempAssistant* for documenting cell-based toxicological test methods, we evaluate the system as described in Section 2.1 under two controlled conditions that bracket realistic use: a positive control in which answers are present in the context document, and a negative control in which they are absent (see overview in **Figure 3**). The design follows current recommendations for LLM evaluation while keeping the method simple and reproducible²⁶.

The current system uses gpt-4o-mini²¹, and we additionally test o3-mini²² and gpt-4.1-nano²³. All three provide large context windows ($\approx 128k$, $\approx 200k$, and up to $\approx 1M$ tokens, respectively), so inputs are not truncated. This enables like-for-like comparisons across models and makes the evaluation procedure reusable as models evolve. We discuss model utility rather than latency or cost. The specific version used for the evaluation in this paper is *ToxTempAssistant* v1.3.1, archived at <https://doi.org/10.5281/zenodo.16749296>.

There is no universally accepted, fully unambiguous ground truth for ToxTemp documentation. Even with comprehensive source materials, experts make interpretation and summarisation choices, so multiple non-identical answers can still be valid. We therefore score model behaviour with classification metrics that depend only on whether (1) the relevant information is in the provided context and (2) how closely the model's output matches the reference answer. Here we employ a threshold τ of the cosine similarity $\cos \theta$ to distinguish relevant from non-relevant answers.

We then cast the evaluation as a confusion-matrix problem (**Figure 4**) with the ultimate goal of calculating an accuracy metric for *ToxTempAssistant*. With an accuracy metric in hand, one can start to optimise *ToxTempAssistant* not only under model selection but also other controllable variables such as prompt design, decoding parameters (e.g., temperature), context pre-processing (e.g., image-to-text, table-to-text), retrieval strategies (e.g., chunking, linear, parallel), and post-processing rules, however optimisation will not be discussed herein.

2.2.1 Positive control: performance under optimal input conditions

In the positive control we ask a simple question: to what extent can *ToxTempAssistant* accurately reproduce the content of an expert-completed ToxTemp document when provided with said document as input? To answer this we task *ToxTempAssistant* with generating eight distinct draft ToxTemps based on eight expert-completed ToxTemp documents and evaluate the answers using cosine similarity ($\cos \theta$) as a metric of quality. See also **Figure 3** left hand side for a summary of the positive control.

Using expert-completed ToxTemps both as the source documents (*i.e.*, the context from which the model must extract information) and as the reference answer for evaluation is intentional. This design isolates the model’s intrinsic ability to extract, summarise, and place information when the content is explicitly present in the provided context. Unlike classical rule-based systems, LLMs generate text based on probabilities and offer no guarantees of consistency, completeness, or contextual accuracy²⁷. Consequently, even with a filled ToxTemp as input, an LLM may still hallucinate, confuse or misplace information.

Document selection and initial screening

Eight expert-completed, peer-reviewed and publicly available ToxTemp documents describing eight different cell-based toxicological test methods are selected. These externally-completed templates accompany two base publications as supporting information and are listed together with the reported assay name in **Table 1**. The base publications are the “OECD recommendations on Evaluation of Data from the “Developmental Neurotoxicity In-Vitro Testing Battery”¹ and the “EFSA Pilot Project on New Approach Methodologies (NAMs) for Tebufenpyrad Risk Assessment”². One may note that these ToxTemp documents were authored under only two principal investigators at two institutions, so despite a document count of eight, it reflects only a limited diversity in answering styles. In addition, this positive control assumes that all information needed is present within the provided context document, a fair but not necessarily strictly fulfilled assumption. An investigation of the documents shows that a small number of questions are not answered $N_{missing}$, see **Table 1**. For example,

researchers may opt to give only an abstract without answering the individuals subquestions that accompany the ToxTemp abstract instructions. This is corrected for in the analyses.

Additionally, the quality of the selected expert-completed ToxTemp documents is assessed by independently evaluating each expert-generated question-answer pair. Using an LLM-as-a-judge model (gpt-4o), a predefined rubric is applied to rate whether answers adequately address the corresponding ToxTemp question, assigning one of three quality ratings: 'High': Fully or nearly fully addresses the question; content is relevant and complete, 'Medium': Partially addresses the question or lacks clarity. 'Low': Does not address the question or is missing. This so-called LLM-as-a-judge approach provides a scalable and internally consistent proxy for manual review, though it may fail to capture domain-specific nuances and could introduce model-specific biases²⁸.

Input preparation

The following processing steps are applied to all context documents prior to LLM inference: 1) section titles and corresponding answer texts are retained to preserve context, 2) question texts are omitted to prevent semantic leakage (*i.e.*, inadvertent matching of model queries to co-located answers based on lexical overlap of the question text rather than information extraction), 3) assay name is retained to support contextual grounding during LLM prompting, and 4) figures are omitted since the current system accepts only text as input.

Evaluation metrics

Cosine similarity is a widely used metric for comparing the meaning of two pieces of text. After texts are converted into numerical representations of their meaning, cosine similarity measures how closely these representations align (evaluated by the angle between their respective vector representations, hence cosine in the name), producing a score between 0 (unrelated meaning) and 1 (very similar meaning)²⁹. In this study, higher cosine similarity indicates closer agreement between a model-generated answer and the expert reference. However, depending on the underlying text representation,

scores rarely reach exactly 1 and should be interpreted comparatively rather than absolutely³⁰. See second paragraph of the Results section for an example of an answer-answer pair with cosine similarity of 0.7 for an intuitive reference point.

In light of the confusion matrix (**Figure 4**) and the goal to calculate overall accuracy (see Section 2.2.3) we are required to calculate the recall R , but calculate other insightful metrics like precision Pr as well.

Completeness $C = \frac{N_{non-trivial}}{N_{questions}}$, quantifies the fraction of ToxTemp questions for which the model generates a non-trivial response. The quality or the degree to which model-generated answers preserve the meaning of the reference answers is assessed using cosine similarity ($\cos \theta$) above threshold τ , a commonly used metric for this task²⁹. Text embeddings, *i.e.*, semantic representations of the text as vector, are generated using OpenAI's text-embedding-3-large³¹ for both the model output and reference answers. It is a simple algebraic step using the dot-product to calculate the cosine of the angle between two normalised vectors. Here, we consider values of $\cos \theta \leq \tau = 0.6$ as unrelated, partitioning the non-trivial *ToxTempAssistant* generated answers into two groups: correct answers and false answers (vis. **Figure 4**). The threshold value τ of 0.6 is chosen empirically and will depend on the embedding model. Spot tested answers with $\cos \theta \approx 0.6$ still show utility despite being overly verbose.

The *recall*³² (also called sensitivity or true positive rate) at threshold τ ($R|_{\cos \theta > \tau}$) measures the fraction of all questions for which the model produces, both, a non-trivial response and exceeds the threshold for the cosine similarity. It is given by:

$$R|_{\cos \theta > \tau} = \frac{TP}{TP + FN} = \frac{N_{correct}(\tau)}{N_{questions}}, \text{ where } N_{correct}(\tau) = \sum_i^{N_{questions}} 1[non - trivial \wedge \cos \theta_i > \tau]$$

The *precision*³² or conditional correctness at threshold τ ($Pr|_{\cos \theta > \tau}$) reflects the quality of the subset of questions the model answered and is defined as:

$$Pr|_{\cos\theta>\tau} = \frac{TP}{TP + FP} = \frac{N_{correct}(\tau)}{N_{non-trivial}}$$

Since $N_{non-trivial} = C \times N_{questions}$, it follows $Pr|_{\cos\theta>\tau} = \frac{R|_{\cos\theta>\tau}}{C}$. We will therefore focus our discussion on completeness and precision, since those metrics seem to be the most intuitive and tabulate the remainder.

2.2.2 Negative control: performance under irrelevant input conditions

To evaluate the specificity of *ToxTempAssistant* we ask whether it can correctly recognise when a document contains no information relevant to cell-based toxicological test methods and therefore abstains from generating a non-trivial response. This negative control was conducted using out-of-scope inputs that resemble scientific texts or contain overlapping terminology (e.g., “cells,” “exposure,” “assays”), yet no relevant information on cell-based toxicological test methods. See also **Figure 3** right hand side for a summary of the negative control.

Document selection

Again, eight documents were selected across diverse domains (e.g., policy documents, environmental reports, and biomedical reviews). Each document was manually rated for expected difficulty for *ToxTempAssistant*, i.e., how likely it is to be confused, based on its terminological and structural similarity to cell-based toxicological test method descriptions. The full list of the documents, rated difficulty levels, and the rationale for each document are provided in the supplementary materials (**S2**).

Input preparation

No text cleaning or reformatting was applied; the PDFs were ingested *as-is*. No assay name or description was supplied.

Evaluation metrics

Specificity was calculated per document (PDF) and per model.

$Sp = \frac{TN}{TN + FP} = \frac{N_{trivial}}{N_{unanswerable\ questions}}$ where $N_{trivial}$ is the number of “Answer not found in document.” responses.

2.2.3 Overall accuracy

Given the metrics of Recall and Specificity have been established in Sections 2.2.1 and 2.2.3, it is now possible to calculate the pooled accuracy. It reflects the fraction of questions answered sufficiently or correctly left unanswered ($A|_{\cos\theta>\tau}$) and is given by

$A|_{\cos\theta>\tau} = \pi * R|_{\cos\theta>\tau} + (1 - \pi) * Sp$ where $\pi = \frac{N_{pos}}{N_{neg} + N_{pos}} = \frac{N_{pos}}{N_{questions}} = 0.47$. For the present study $N_{pos} = 546$ and $N_{neg} = 616$ are the number of questions

ToxTempAssistant is tasked to answer for the positive ($8 \times 77 - \sum N_{missing}$) and negative control (8×77), respectively.

3. Results

Here we aim to quantify how accurate *ToxTempAssistant*, an LLM-assisted tool for generating ToxTemp drafts, retrieves and structures information under controlled conditions. We evaluate whether *ToxTempAssistant* can do so when relevant content is present in the provided context (positive control) and abstains from answering when no relevant content is present (negative control). We cast the evaluation as a confusion-matrix problem (**Figure 4**). Three alternative LLMs back-ends (o3-mini, gpt-4o-mini, gpt-4.1-nano) contrast the effect of model type. We show that accuracy is high and robust despite model choice, due to compensating shortcomings in completeness, recall and specificity.

Starting with reporting on the results of the positive control section of **Table 2**, of $N = 546$ answerable items (eight ToxTemps * 77 questions minus expert-unanswered fields), o3-mini produced near-ideal completeness of 99.8% with a precision of 79.8%. This means about 20% of the questions were deemed not sufficiently answered (*i.e.*, they did not exceed the set cosine similarity threshold of $\tau = 0.6$). GPT-4o-mini showed a lower, intermediate completeness of 68.1%, while maintaining precision comparable to the previous model (79.6%). Gpt-4.1-nano was most conservative in providing answers with a completeness of only 44.3%, while achieving the highest precision of all models

(83.1%). The model-averaged mean cosine similarity $\overline{\cos \theta}$ among non-trivial responses was similar across all tested models (0.693–0.716), which directly follows from the limited spread in precision. This consistency, reflected in the similarity of the curves across samples, presents itself already on document-level, as shown in **Figure 5**. The distributions of $\cos \theta$ on answer-level also follows a similar trend for all documents and models (Supplementary Material **S3**), yet with a much wider spread. In the following we provide the reader with an example of an answer-answer pair with a cosine similarity of 0.7. The reader is encouraged to look into our GitHub repository (https://github.com/johannehouweling/ToxTempAssistant/tree/main/myocyte/toxtempass/evaluation/positive_control/output) for further examples.

ToxTemp Question: *“Indicate whether the test method is linked to an OECD Test Guideline (how, and which) or other regulatory guidance (e.g. EMA).”*

Expert-answer: *“Test is not linked to regulatory guidelines.”*

ToxTempAssistant (o3-nano, $\cos \theta = 0.7$): *“The test method is not linked to any OECD Test Guideline or other regulatory guidance such as EMA guidelines. This is stated explicitly in the document `_`(Source: NPC2-5.pdf, page 156)`_`.”*

Our subjective evaluation is that this is a helpful answer. One may note, that the *ToxTempAssistant* paraphrases the question much more in the answer, the actual brief answer of the expert makes up only 1/3rd of the *ToxTempAssistant*’s answer. A source-attribution string is included at the end. The source string enables direct attribution to the correct document especially in multi-document contexts. Despite the high utility of this answer the example also highlights the importance of the expert-answer quality. We therefore analysed the quality of expert answers with an LLM-as-a-judge approach to classify the expert-answers into three quality categories: high, medium and low (see Section 2.2.1). 310 (56.8%) were scored as high-quality, 117 (21.4%) as medium-quality, and 119 (21.8%) as low-quality. Inspection of answers categorised as low-quality, despite being expert-generated, demonstrates a considerable proportion of answers contains omissions, incomplete phrasing, or lacks sufficient clarity to be self-contained. For all models, statistical analysis shows that

cosine similarity is positively correlated (Spearman $\rho = 0.324$, $p < 0.001$ overall), with ordered quality levels (low to high), but the effect stagnates between medium and high (Supplementary Material **S4**).

To assess whether performance patterns persist on directly comparable outputs, we evaluated cosine similarity for the subset of questions answered by all three models ($N = 206$), potentially representing an easier subset of the full question set favouring high coverage models. Within this subset, mean cosine similarity $\overline{\cos \theta}$ exceeded 0.7 for all models (0.719 ± 0.124 for gpt-4.1-nano, 0.714 ± 0.126 for gpt-4o-mini, and 0.727 ± 0.118 for o3-mini), with 83.0% of gpt-4.1-nano and gpt-4o-mini outputs and 89.8% of o3-mini outputs passing the set cosine similarity threshold ($\tau = 0.6$). Distributions were otherwise comparable, although o3-mini showed slightly higher pass rates and reduced variability, indicating more consistent alignment with reference answers (Supplementary Material **S5**). When restricting the comparison to the paired, question-level differences, only marginal but interpretable contrasts emerged. o3-mini achieved a statistically higher cosine similarity than gpt-4o-mini, while differences relative to gpt-4.1-nano were not statistically significant (Supplementary Material **S5**). Inspection of individual questions revealed that the largest differences between models stemmed from verbosity rather than factual disagreement: concise paraphrases aligned with the reference scored higher, whereas expansive answers introducing factually correct details were penalised by cosine similarity metric. Notably, some of this added detail reflected imprecise NAM validation language phrasing carried over from the source document (e.g., repeating the phrasing “formal OECD GD34 validation study” even though GD34 is guidance rather than a test guideline), underscoring the value of expert review, as exemplified here:

ToxTemp Question: *“If the method has undergone some form of validation/evaluation, give its status. (9.4)”*

Expert-answer: *“readiness analysis: readiness score A, according to Bal-Price et al. (2018)”*

Expert-answer-quality (LLM-as-a-judge): Medium, justification: *“The answer provides some information about the validation status by mentioning a readiness score and a reference, but it lacks clarity and completeness. It does not explicitly state the validation status or provide details on the evaluation process.”*

ToxTempAssistant (gpt-4.1-nano, $\cos \theta = 0.68$): *“The assay has undergone a readiness analysis and has obtained a readiness score of A, according to Bal-Price et al. (2018) (Source: NPC1.pdf).”*

ToxTempAssistant (gpt-4o-mini, $\cos \theta = 0.44$): *“The neural progenitor cell proliferation assay (NPC1) has undergone a readiness analysis and obtained a readiness score of A, according to the criteria published by Bal-Price et al. (2018). However, no formal OECD 34 validation study, such as ring trials with a standard set of known positive and negative controls, has been conducted to date. The assay has successfully tested 123 unique compounds (as defined by unique DTXSIDs) _(Source: NPC1.pdf)_.”*

ToxTempAssistant (o3-mini, $\cos \theta = 0.61$): *“The method was evaluated by a readiness analysis and obtained a readiness score A (Bal-Price et al., 2018). Although it has been applied to 123 unique compounds successfully, no formal OECD validation study—such as ring trials with standard controls—has been performed. (Source: NPC1.pdf)”*

In the following, the results of the negative control, see **Table 2**, are presented. On eight out-of-scope input documents, models differed in refusal behaviour. gpt-4.1-nano achieved the highest specificity of 98.4%, gpt-4o-mini slightly lower (92.4%), and o3-mini the lowest specificity of 65.9%. Stratified by *a priori* difficulty of the out-of-scope input documents (**Figure 6**; Supplementary Material **S2**), all models correctly abstain in 74% to 100% of the cases from answering for low/medium difficulty inputs. However, strong divergence emerged for high difficulty input, where gpt-4.1-nano still refrained from answering in 95% of cases, gpt-4o-mini in 80-85%, and o3-mini leaves only 45% of questions unanswered, meaning 55% of questions are answered in error. Analysis of o3-mini’s false positives reveals recurring causes: (1) classifying domain-adjacent

information about cells, measurements, or exposure thresholds as in-scope content; (2) an answer-by-default tendency, *i.e.*, less likely to follow the base prompt to provide no answer (“know-it-all”); and (3) returning real document metadata (titles, authors, affiliation) that are grounded in the source documents but nonetheless out-of-scope for ToxTemp questions. The first two factors explain o3-mini’s low specificity, while metadata misplacements constitute a less harmful error that is also observed in the other two models.

Combining positive- and negative control data at prevalence of $\pi = 0.470$, gpt-4o-mini attained the highest overall accuracy ($A_{\cos\theta>0.6}$) of 0.744, followed by o3-mini (0.724) and gpt-4.1-nano (0.694) (**Figure 7** and **Table 2**).

4. Discussion

ToxTempAssistant represents a practical implementation of LLMs in toxicology that aligns with current technological capabilities¹². By using established LLM strengths in information extraction and summarisation, the tool addresses a fundamental bottleneck affecting the toxicology community: the labour-intensive process of converting experimental documentation into standardised, regulatory-compliant formats. Rather than attempting to automate complex scientific reasoning or decision-making, *ToxTempAssistant* is confined to the well-defined task of structuring information, where LLMs have demonstrated reliable performance and where evaluation protocols can be systematically applied. While building LLM applications is increasingly accessible, their automated evaluation remains challenging³³⁻³⁴.

Under controlled conditions, *ToxTempAssistant* achieved high accuracy across alternative LLMs, but via different balances of completeness (recall), precision and specificity. In other words, performance is robust to model substitution in aggregate, yet the underlying error profiles are model-dependent. Prior work in scientific information-extraction tasks likewise reports task-specific shifts in precision/recall between models, underscoring that aggregate robustness can mask differing failure

modes¹². Moreover, recent evidence that errors correlate across models cautions simple model substitutions may not eliminate shared failure modes³⁵.

Cosine similarity between model drafts and expert references averaged ≈ 0.70 across models, indicating substantial semantic resemblance. Cosine similarity increased with reference-answer quality but showed attenuated gains from “medium” to “high”, implying that once the essential content is present, this metric is less sensitive to incremental refinements. This supports curating/adjudicating references, or restricting analyses to high-quality items, when benchmarking systems intended to produce complete, stand-alone prose. We note that widely used BERTScore³⁶ and Sentence-BERT³⁷ compare texts via cosine similarity and typically correlate with human judgments, yet have known caveats. They can saturate at higher quality bands and even mislead, *i.e.* producing false negatives (lexically different but semantically equivalent phrasing) and false positives (fluent and on-topic text that *e.g.* uses the incorrect reagent or omits critical parameters)^{30–38}. Accordingly, we treat cosine similarity as a coarse alignment signal rather than a fine-grained quality metric and pair it with completeness/abstention metrics. Notably, cosine similarity is used in this manuscript only for evaluation purposes. In the deployed application there is no hidden reference answer, so cosine similarity is neither calculated nor used for accept/reject decisions. Instead, all LLM-generated responses are treated as drafts that require expert review. We conducted limited, non-systematic manual inspection to check borderline cases, but did not perform structured random-sample audit, which remains a limitation of this evaluation. As possible future extensions we may add a metric for faithfulness (ensuring answers are grounded in the provided context) via deterministic extract-compare and/or a rubric-guided LLM-as-a-judge approach³⁹.

Negative control results expose the principal risk: answering out-of-scope. All models abstained reliably on low/medium-difficulty out-of-scope inputs, but divergence emerged on high difficulty inputs, *i.e.*, domain-adjacent terminology that is easily confused with in-scope content. Here, o3-mini most often produced answers likely driven by the reasoning model design which may dilute the intended strictness of the base prompt. By

contrast, gpt-4.1-nano maintained high refusal rates under the same conditions, consistent with a higher internal threshold for responding. On the one hand, for applications where erroneous inclusion of out-of-scope content is costly, this conservative behavior may be preferable even at the expense of missed valid extractions. On the other hand, users of ToxTempAssistant are experts on their method and thus the likelihood of them providing off-topic grounding documents is low. Methodological, our design parallels work by Madhusudhan et al.⁴⁰: both treat abstention ability as a black-box property, quantify it via an Answerable-Unanswerable confusion matrix built from positive/negative controls, and deliberately probe confusable, domain-adjacent inputs. Although framed as open-domain multiple-choice question-answering, they likewise find that state-of-the-art models often fail to abstain, especially on reasoning, conceptual, and problem-solving questions and under-represented domains. They further show that techniques like stricter prompting and Chain-of-Thought enhance abstention ability and that these gains frequently coincide with better overall QA performance (e.g., higher precision and competitive answerable accuracy), underscoring the value of abstention-focused evaluation and prompting.

Design implications follow directly. First, model selection should be task- and risk-sensitive: use a high-specificity model (e.g., gpt-4.1-nano) for triage/abstention and a higher-recall model (e.g., o3-mini) for candidate drafting, with cross-checking before acceptance. However, when combined with metrics of cost and time, given that overall precision is somewhat the same, we for now decided to keep gpt-4o-mini in the production version of *ToxTempAssistant* and keep monitoring both capability and price development as new models become available. New GitHub issues for improvement ideas based on the insights of the evaluation like using a conservative model first to select relevant documents, then use a less costly model to answer based on the filtered documents and potentially allowing the user to select a specific model were created.

Our evaluation shows that *ToxTempAssistant* satisfies the *prima facie* requirements for a human-in-the-loop documentation assistant, and our metrics (completeness,

precision, specificity, accuracy) delineate an empirical envelope of behaviour under controlled conditions, revealing both strengths and failure modes. However, these scenarios remain abstractions of routine practice.

A limitation of our positive control is that it functioned as an “open-book” regurgitation exercise, since the model can extract information from already-completed ToxTemps. More broadly, performance is sensitive to input preparation: changes in input type and structure can substantially affect LLM accuracy, and this effect is not uniform across ToxTemp question types. Low difficulty ToxTemp question items are expected to be explicit and local (e.g., readily stated method identifiers, materials/cell line names, exposure times, endpoints reported in a single paragraph/table) and therefore benefit from straightforward retrieval even when the input is heterogeneous. By contrast, higher difficulty ToxTemp questions may be distributed, conditional, or interpretive (e.g., validation status and evidence, applicability domain/limitations, data-quality considerations, rationale/justifications, handling of deviations, or questions requiring synthesis across sections and careful distinction between “not reported” vs “not applicable”). The latter are more sensitive to how the source is segmented, whether section context is preserved, and whether the prompt scaffolds selective extraction and abstention. A next step is an end-to-end validation using the original base publications (and other primary sources) to test selective extraction and correct refusal under realistic, messy inputs.

A further source of variability not captured in this study is the user-provided assay description. For this evaluation we left it blank, but expect that this free-text input may differ markedly in detail and terminology per user and will therefore influence output quality by priming the model toward relevant parts of the supplied documents or by introducing ambiguity. To avoid conflating this effect with document-bounded performance, we intentionally left the assay description field empty for both positive and negative controls. We therefore flag assay description as another practical factor that should be addressed in future work.

Guided by Donald E. Knuth’s maxim—“premature optimization is the root of all evil”—we deliberately defer optimisation of *ToxTempAssistant* on the present dataset. Given that the *ToxTempAssistant* is now available online, our next step is to gather real-world evidence before tuning, spanning the full range of heterogeneous assay documentation (e.g. standard operating procedures, lab protocols, scientific articles, structured metadata, and figures). With explicit consent, we will invite users to share the input documents they supply to the assistant and their completed, expert-verified ToxTemps. This will create a paired, practice-derived corpus linking context, model draft, human edits, and final outputs, enabling user-based evaluation of effectiveness (e.g., edit burden, time-to-completion, and abstention correctness) and a clearer view of failure modes. Only once such a dataset becomes available will we consider targeted optimisation, ensuring that improvements are driven by observed practice rather than benchmark artefacts. Nonetheless, having an established evaluation pipeline will not only be useful for evaluating real-world data but is also required because both LLM behaviour and regulatory expectations evolve.

From a governance perspective, the tool does not attempt end-to-end automation and requires expert verification of every generated statement. This human oversight, combined with transparent provenance (all model parameters embedded in export metadata), supports reproducibility audits and is consistent with a “limited-risk” posture under the EU AI Act¹⁸. The system’s model-agnostic architecture and open-source prompts/evaluation scripts facilitate adaptation as newer, cheaper, or more capable models appear, enabling straightforward re-implementation and re-evaluation.

Beyond the technical aspects, *ToxTempAssistant* exposes a broader community challenge: the absence of consensus on documentation requirements for cell-based toxicological test methods and other NAMs^{41 42}. This tool helps bridge scattered research outputs and regulatory templates, but sustained harmonisation will require clearer, widely accepted standards and more precise question formulations that reduce ambiguity. By reducing the documentation burden, *ToxTempAssistant* creates room to

increase the specificity of ToxTemp, without pushing complexity beyond a tractable level, an important step toward reducing ambiguity.

While *ToxTempAssistant* offers technical infrastructure to potentially streamline ToxTemp completion, its true value depends on demonstrable end-user demand. The central question is no longer whether LLMs *can* structure documentation, but whether stakeholders (*i.e.* regulators, researchers, and industry) actively consult ToxTemps, and whether this documentation format will meaningfully facilitate the acceptance of cell-based NAMs. Early indications are encouraging: ToxTemps have already been reviewed in OECD Integrated Approaches to Testing and Assessment case studies and may be included in formal validation exercises led by the European Centre for the Validation of Alternative Methods. We therefore invite the community to test the publicly available instance and contribute to its development via the GitHub repository; such engagement will not only justify the continued refinement of *ToxTempAssistant* but also shape its role in advancing the regulatory acceptance of NAMs.

5. Data and software availability

The data that support the findings of this study are openly available on Zenodo at <https://doi.org/10.5281/zenodo.17047715>.

All experiments can be re-run end-to-end using the provided scripts and input documents to reproduce the generated outputs within the limits of non-determinism of LLM implementations. The code is available on GitHub including the results of the positive and negative control discussed in this manuscript as well as further documentation at

<https://github.com/johannehouweling/ToxTempAssistant/tree/main/myocyte/toxtempass/evaluation>

A *ToxTempAssistant* instance is hosted at <https://toxtempassistant.vhp4safety.nl/> and is open-source under the GNU Affero General Public License v3. An archived, citable release is available on Zenodo at <https://doi.org/10.5281/zenodo.15607642>

(latest archived version). The specific version used for the evaluation in this paper is *ToxTempAssistant* v1.3.1, archived at <https://doi.org/10.5281/zenodo.16749296>.

6. Disclosure statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7. Acknowledgments

The authors thank Ozan Cinar (ORCID: 0000-0003-0329-1977) for his technical support in deploying the application on the VHP4Safety Platform. We gratefully acknowledge Julia Meerman (ORCID:0000-0001-6487-1655), Jelle Vriend (ORCID:0000-0001-5758-0614), Mirthe Klaassen, Marie Corradi (ORCID:0000-0001-8185-5913), and Rob Beffers for their contributions to the initial conceptualisation and development of *ToxTempAssistant* during the 4th VHP4Safety Hackathon. We also thank Mirjam Luijten (ORCID:0000-0002-5277-1443) for her thoughtful review and valuable feedback.

LLMs were used as part of the methodological approach, as described in the Methods section. In addition, ChatGPT (OpenAI) and Claude (Anthropic) were used occasionally to improve the readability of the manuscript. All LLM-assisted content was reviewed and verified by the authors, who take full responsibility for the final version.

ORCID IDs

Jente Houweling, jente.houweling@rivm.nl, ORCID: 0009-0005-3680-0645

Matthias Arras, matthias.arras@gmail.com, ORCID: 0000-0002-4714-9086

Egon Willighagen, egon.willighagen@maastrichtuniversity.nl, ORCID: 0000-0001-7542-0286

Danyel Jennen, danyel.jennen@maastrichtuniversity.nl, ORCID: 0000-0002-8618-2487

Chris Evelo, chris.evelo@maastrichtuniversity.nl, ORCID: 0000-0002-5301-3142

Anne Kienhuis, anne.kienhuis@rivm.nl, ORCID: 0000-0002-6465-4498