

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

Jente M. Houweling<sup>1,2</sup>, Matthias M.L. Arras<sup>3</sup>, Egon L. Willighagen<sup>2</sup>, Danyel Jennen<sup>4</sup>, Chris T. Evelo<sup>2</sup>, Anne Kienhuis<sup>1,5</sup>

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 deterministic, 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-4o-mini, o3-mini, gpt-4.1-nano). Performance was assessed via a confusion-matrix

framework using a fixed cosine-similarity threshold to derive completeness, precision, specificity, and accuracy.

## 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 effectively uses established LLM capabilities on extraction and summarisation to generate ToxTemp drafts. This shifts 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, toxicological test methods, large language models, documentation, LLM evaluation

## Author affiliation

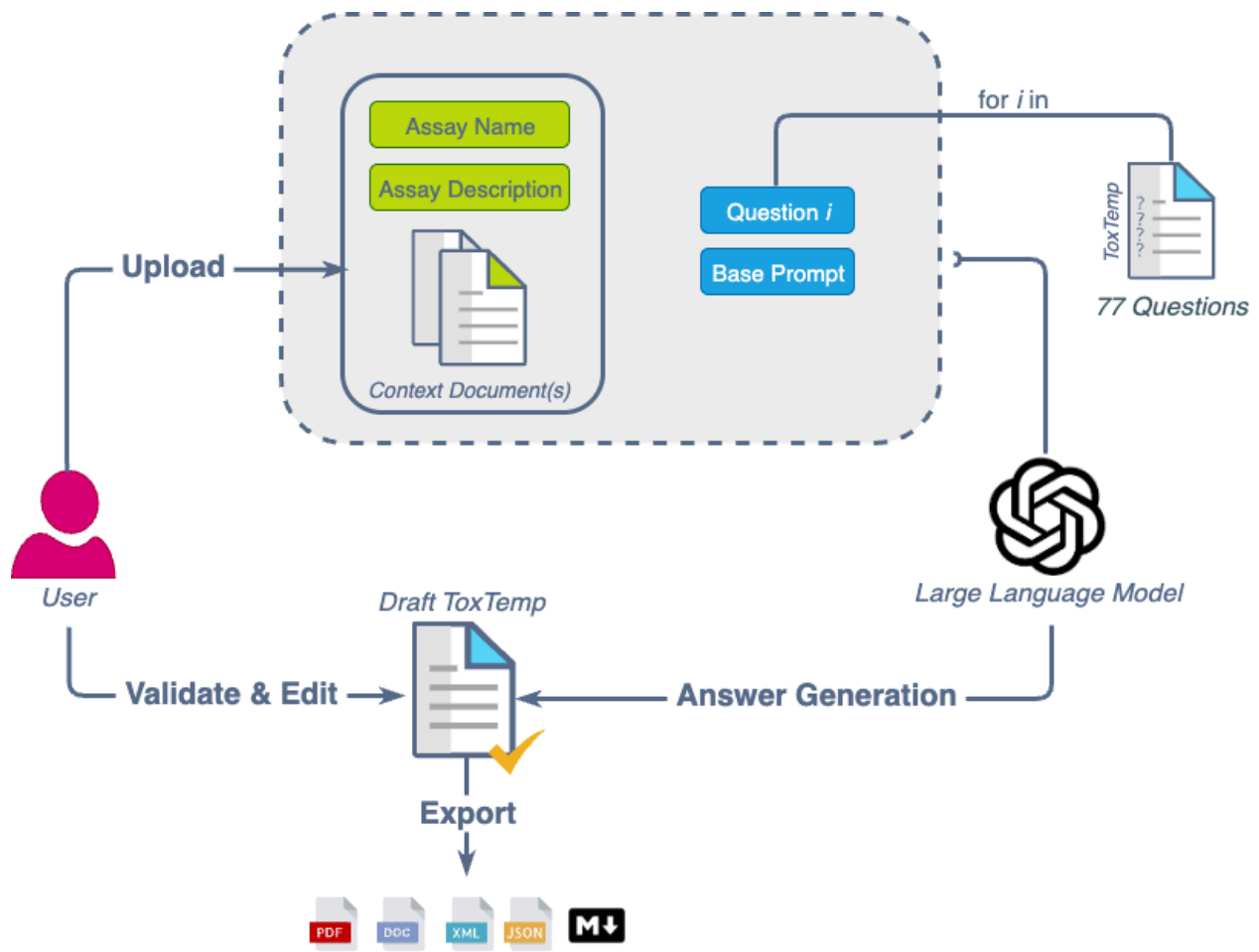
1. National Institute for Public Health and the Environment (RIVM), Bilthoven, Netherlands
2. Dept of Translational Genomics, NUTRIM Institute of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, Netherlands
3. Independent researcher

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.

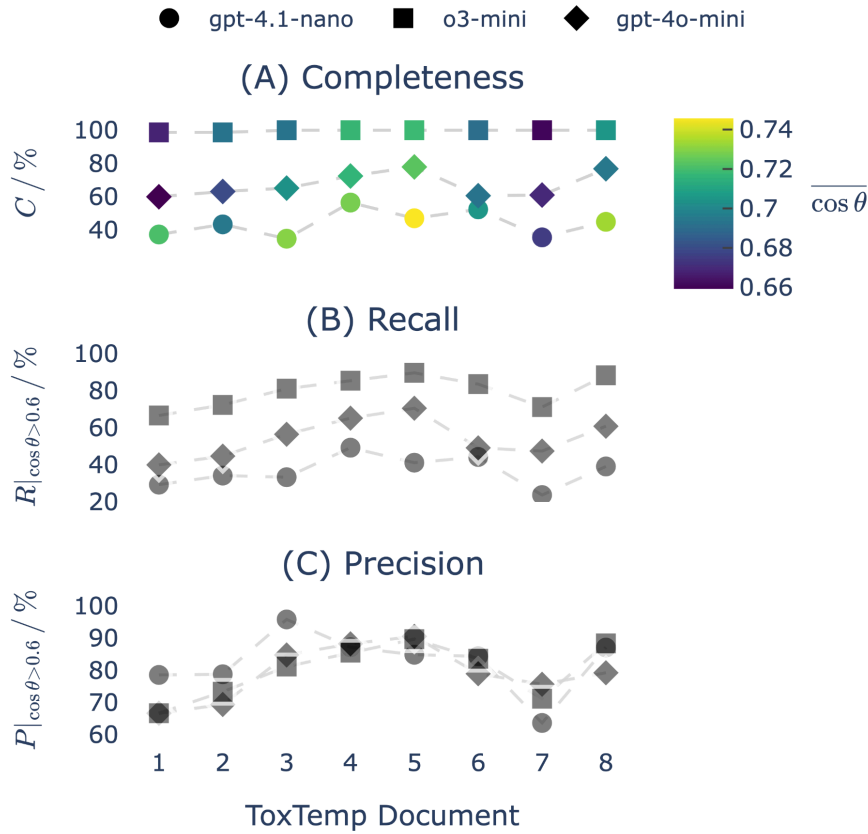
## Figures



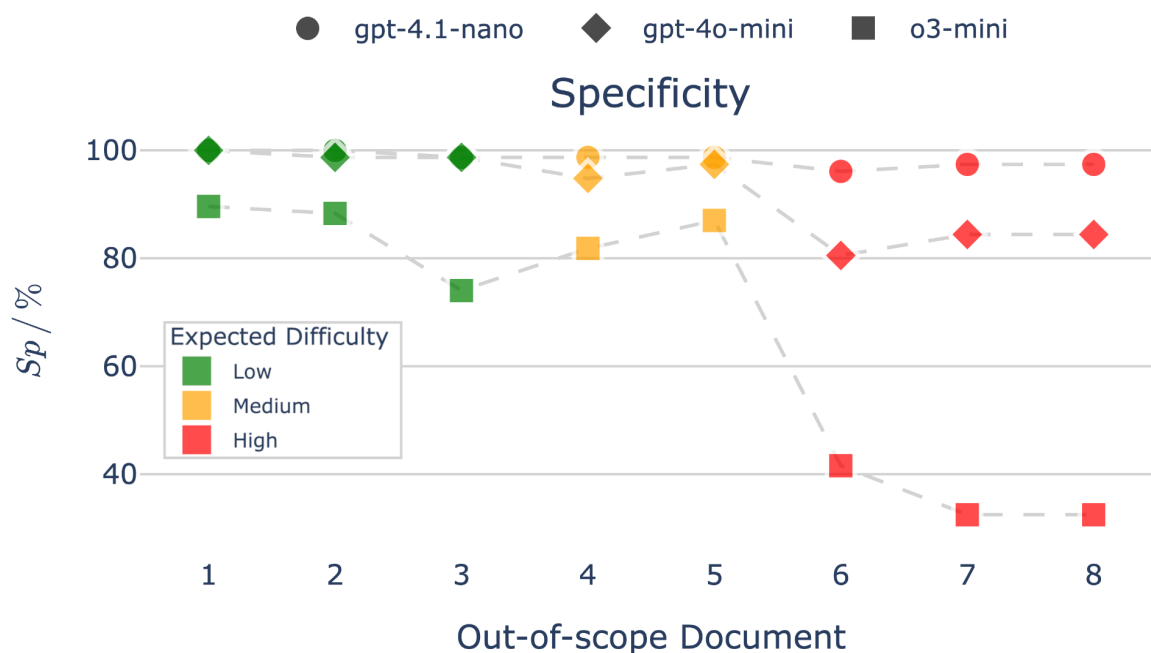
**Figure 1** | Schematic overview of the ToxTempAssistant workflow. The ToxTempAssistant enables automated drafting of ToxTemp templates using a large language model (LLM). Users upload contextual documents along with the assay name and short assay description. For each of the 77 ToxTemp questions, targeted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

		ToxTempAssistant Predicted		
		Answer Generated	False / No Answer Generated	
Actual	Answer in context	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 context	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 $A = \frac{TP + TN}{TP + FP + FN + TN}$

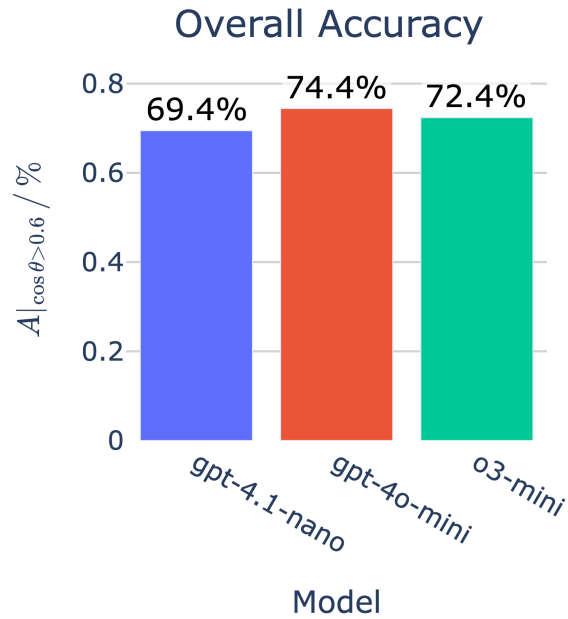
**Figure 2 |** Confusion matrix used to analyse the ToxTempAssistant evaluation data. It shows key metrics and corresponding equations. A trivial response is defined as “Answer not found in documents”. We use cosine similarity above threshold  $\tau$  ( $\cos \theta > \tau$ ) to distinguish between correct 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  $\tau$ , 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 3 |** Positive control test of ToxTempAssistant using eight publicly available, factually complete ToxTemp documents as inputs. (A) Shows the percentage of answered ToxTemp questions for each ToxTemp document and three LLMs as a measure for completeness. Colours represent mean cosine similarity  $\overline{\cos \theta}$  between model-generated responses and expert reference answers. Higher scores indicate stronger semantic similarity with the reference content. (B) Shows the proportion of well answered question of total answerable questions, or recall  $R|_{\cos \theta > 0.6}$  where the cosine similarity is above 0.6. (C) Shows precision  $Pr|_{\cos \theta > 0.6}$ , i.e. within the subset of questions the model chose to answer the percentage meeting the same similarity threshold as B. Shapes depict models: gpt-4.1-nano (circle), gpt-4o-mini (diamond), and o3-mini (square). Together, these panels convey that completeness and recall are sensitive to model selection, while cosine similarity and hence precision are only weakly affected.



**Figure 4 |** Negative control test of ToxTempAssistant using irrelevant 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. Marker shapes depict models: gpt-4.1-nano (circle), gpt-4o-mini (diamond), and o3-mini (square). Colours denote the manually assigned a priori difficulty level, ranging from Low (green) to High (red), based on how likely domain-adjacent terminology in the context document is to confuse ToxTempAssistant.



**Figure 5 | Overall accuracy of ToxTempAssistant pooled evaluation data.** Bar chart compares the prevalence-weighted accuracy  $A|_{\cos\theta>0.6}$  across three LLM models on the combined positive  $N_{pos} = 546$  and negative  $N_{neg} = 616$  control sets ( $\pi = 0.47$ ). Accuracy is the fraction of non-trivial answers whose cosine similarity to the expert reference is above 0.6 and (ii) correct abstentions on out-of-scope inputs.. Differences are modest, indicating that overall accuracy is relatively robust to model choice despite differing recall and specificity profiles.



## Tables

Document	Assay Name	$N_{missing}$	Reference
ToxTemp1	NPC1	2	1
ToxTemp2	NPC2-5	1	1
ToxTemp3	cMINC (UKN2)	8	1
ToxTemp4	NeuriTox assay (UKN4)	8	2
ToxTemp5	MitoMet assay (UKN4b)	9	2
ToxTemp6	MitoStressLUHMES (MSL)	16	2
ToxTemp7	MitoComplexesLUHMES(MCL)	18	2
ToxTemp8	PeriTox test (UKN5)	8	2

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

Model	Positive control						Negative control	Combined	
	$C$	$N_{non-trivial}$	$R _{\cos\theta>0.6}$	$Pr _{\cos\theta>0.6}$	$F_1 _{\cos\theta>0.6}$	$\overline{\cos\theta}$	$Sp$	$\pi$	$A _{\cos\theta>0.6}$
<i>gpt-4.1-nano</i>	0.443	242	0.368	0.831	0.510	0.716	0.984	0.470	0.694
<i>gpt-4o-mini</i>	0.681	372	0.542	0.796	0.645	0.694	0.924	0.470	0.744
<i>o3-mini</i>	0.998	545	0.797	0.798	0.797	0.693	0.659	0.470	0.724

**Table 2 |** Baseline performance of ToxTempAssistant across three Large Language Models under positive and negative control conditions. The table summarises how each model reconstructs expert-completed ToxTemp answers ( $N = 546$  question-answer pairs) using cosine similarity based metrics ( $\tau = 0.60$ ) and, in parallel, to remain silent when presented with eight off-target documents ( $N = 616$  questions with no correct answer). Metrics are calculated on a pooled confusion matrix (figure 2).

Completeness ( $C$ ) is the fraction of expert-answered questions for which the model returns a non-trivial response; Non-trivial count ( $N_{non-trivial}$ ) is the numerator for completeness; Recall and precision at  $\tau$  ( $R|_{\cos\theta>0.6}$  and  $Pr|_{\cos\theta>0.6}$ ) are computed by thresholding cosine similarity between model outputs and reference answers, with their harmonic mean reported as  $F1$  ( $F_1|_{\cos\theta>0.6}$ ); Mean cosine similarity ( $\overline{\cos\theta}$ ) summarises non-trivial responses only; Specificity ( $Sp$ ) indicates how often the model correctly abstains from answering; Accuracy ( $A|_{\cos\theta>0.6}$ ) is prevalence weighted ( $\pi$ ) fraction of correct responses and correct abstentions across the pooled dataset.

## 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*. (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*. (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. 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).
6. OECD. *Guidance Document for Describing Non-Guideline In Vitro Test Methods*. (2017). doi:10.1787/9789264274730-en.
7. 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).
8. Vinken, M. *et al.* Safer chemicals using less animals: kick-off of the European ONTOX project. *Toxicology* **458**, 152846 (2021).
9. 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).
10. Zhang, T. *et al.* Benchmarking Large Language Models for News Summarization. *Trans. Assoc. Comput. Linguist.* **12**, 39–57 (2024).
  11. Dagdelen, J. *et al.* Structured information extraction from scientific text with large language models. *Nat. Commun.* **15**, 1418 (2024).
  12. Peters, U. & Chin-Yee, B. Generalization bias in large language model summarization of scientific research. *R. Soc. Open Sci.* **12**, 241776 (2025).
  13. 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).
  14. Houweling, J. & Willighagen, E. Research Output Management. *Qeios* (2023)  
doi:10.32388/ZNWI7T.
  15. 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).
  16. 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.
  17. AI Act | Shaping Europe's digital future.  
<https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai> (2025).
  18. Ji, Z. *et al.* Survey of Hallucination in Natural Language Generation. *ACM Comput Surv* **55**, 248:1-248:38 (2023).
  19. 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).
  20. GPT-4o mini: advancing cost-efficient intelligence.  
<https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>.

21. OpenAI o3-mini. <https://openai.com/index/openai-o3-mini/>.
22. Introducing GPT-4.1 in the API. <https://openai.com/index/gpt-4-1/>.
23. 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).
24. LLM Evaluation Metrics: The Ultimate LLM Evaluation Guide - Confident AI.  
<https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation>.
25. Radford, A. *et al.* Language Models are Unsupervised Multitask Learners.
26. 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).
27. 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).
28. Models - OpenAI API. <https://platform.openai.com>.
29. 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).
30. Chang, Y. *et al.* A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **15**, 1–45 (2024).
31. 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.* (2024) doi:10.1136/bmjebm-2024-113199.
32. 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).
33. 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).
34. Reimers, N. & Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Preprint at <https://doi.org/10.48550/arXiv.1908.10084> (2019).
35. 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.
36. 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.
37. 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.
38. 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).
39. 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.

## 1. Introduction

New Approach Methodologies (NAMs) are transforming chemical hazard and risk assessment by providing human-relevant alternatives to animal testing<sup>3</sup>. Transparent and comprehensive documentation is a prerequisite for scientific confidence in their application<sup>4</sup>. To support researchers in preparing regulatory-compliant descriptions of non-guideline, cell-based toxicological test methods, one type of NAM, Krebs et al. (2019)<sup>5</sup> introduced the ToxTemp template. ToxTemp converts the broad requirements of Organisation for Economic Co-operation and Development (OECD) Guidance Document 211<sup>6</sup> 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. The template has since been adopted by research projects, such as the EU flagship projects RISK-HUNT3R<sup>7</sup> and ONTOX<sup>8</sup>, and a ToxTemp methods database is hosted within the EU-ToxRisk knowledge infrastructure<sup>9</sup>.

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 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<sup>10–12</sup>. 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 tasks<sup>13</sup>, such as research output management<sup>14</sup>. While proof-of-concept studies demonstrate the feasibility of using finetuned LLMs to fill regulatory templates<sup>15</sup> and combination of LLMs with knowledge graphs show improved chemical information retrieval from datasets<sup>16</sup>, widespread adoptions of validated, user-ready implementations in the field of toxicology are still limited.

Here we introduce *ToxTempAssistant*, an LLM-based tool that generates draft ToxTemp documents, enabling 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 measures<sup>17</sup>. 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 information<sup>17</sup>, potentially compromising scientific integrity.

Recognising that validation is central to trust in the use of AI-based tools<sup>19</sup>, 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 models (gpt-4o-mini<sup>20</sup>, o3-mini<sup>21</sup>, and gpt-4.1-nano<sup>22</sup>). 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*.

## 2. Methods

### 2.1 Development of ToxTempAssistant

To build *ToxTempAssistant*, the 77 questions of the ToxTemp template were extracted from the original publication by Krebs et al.<sup>5</sup> and converted into a structured, machine-readable JavaScript Object Notation (JSON) file. The resulting file preserves the ToxTemp’s hierarchy 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.



### 2.1.1 ToxTempAssistant implemented as a web-application

The beta version of the tool is hosted at [<sup>23</sup>https://toxtempassistant.vhp4safety.nl](https://toxtempassistant.vhp4safety.nl) 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. Currently accepted formats for LLM input include PDF, TXT, HTML, DOCX, Markdown (MD). Support for additional formats (e.g., ODF, PNG, JPG) may be added in the future;
2. Receive LLM-generated draft responses for each ToxTemp question
3. Inspect, edit the draft ToxTemp, and if necessary (re-)generate answers at the single or multiple questions;
4. Export the ToxTemp template in various formats (PDF, DOCX, XML, HTML, JSON, MD, Extensible Markup Language (XML)).

During the VHP4Safety project<sup>23</sup>, a group of test users were engaged to guide defining and prioritising the features. Additionally, beta-user feedback on the usefulness, clarity, and scientific accuracy of *ToxTempAssistant* is systematically collected to inform ongoing refinement.

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 environment<sup>20</sup>. 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 structured base prompt designed to answer individual questions from the ToxTemp with strict adherence to source material provided by the user. 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 full prompt text is provided in Supplementary Materials **S1** and archived on Zenodo (DOI 10.5281/zenodo.1234567).

## 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. The design follows current recommendations for LLM evaluation while keeping the method simple and reproducible<sup>24</sup>.

The current system uses gpt-4o-mini<sup>20</sup>, and we additionally test o3-mini<sup>21</sup> and gpt-4.1-nano<sup>22</sup>. 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.

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) the model's output matches the reference answer closely enough. Here we employ a threshold  $\tau$  of the cosine similarity  $\cos \theta$  to distinguish relevant from non-relevant answers.

We then cast the evaluation as a confusion-matrix problem (**Figure 2**) 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.

### 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 regurgitate 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.

Using expert-completed ToxTemps both as c. It isolates the model’s intrinsic capabilities. Unlike classical rule-based systems, LLMs generate text based on probabilities and offer no guarantees of consistency, completeness, or contextual accuracy<sup>25</sup>. Consequently, even with a filled ToxTemp as input, an LLM may still hallucinate 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 ToxTemp documents 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<sup>1</sup> and the EFSA Pilot Project on New Approach Methodologies (NAMs) for Tebufenpyrad Risk Assessment<sup>2</sup>. One may note that these ToxTemp documents were authored under 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. A spot 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, to analyse the quality of the selected ToxTemp documents, each expert-generated ToxTemp question-answer pair is independently assessed using an LLM-as-a-judge approach<sup>26</sup> (gpt-4o), and assigned 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.

### *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 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*

In light of the confusion matrix (**Figure 2**) 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  $\tau$ , a commonly used metric for this task<sup>27</sup>. Text embeddings, i.e., semantic representations of the text as vector, are generated using OpenAI’s text-embedding-3-large<sup>28</sup> 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 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 2**). The threshold value  $\tau$  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*<sup>29</sup> or true positive rate at threshold  $\tau$  ( $R|_{\cos \theta > \tau}$ ) 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*<sup>29</sup> or conditional correctness at threshold  $\tau$  ( $Pr|_{\cos \theta > \tau}$ ) 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  $Pr|_{\cos \theta > \tau} = \frac{R|_{\cos \theta}}{C}$ . 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.

#### *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 in correctly refusing to answer, 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 not 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

$$A|_{\cos\theta>\tau} = \pi * R|_{\cos\theta>\tau} + (1 - \pi) * Sp \text{ where } \pi = \frac{N_{pos}}{N_{neg} + N_{pos}}$$

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 ( $8 \times 77$ ), respectively.

### 3. Results

Here we aim to quantify how accurate *ToxTempAssistant*, a 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 abstain from answering when no relevant content is present (negative control). We cast the evaluation as a confusion-matrix problem (**Figure 2**). 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-missing 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% with as strong precision (79.6%) as the previous model. 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 3**. 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 Zenodo repository (DOI: 10.5281/zenodo.17047715) for further examples.

**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 a 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 ( $0.719 \pm 0.124$  for gpt-4.1-nano,  $0.714 \pm 0.126$  for gpt-4o-mini, and  $0.727 \pm 0.118$  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 factional disagreement: concise paraphrases aligned with the reference scored higher, whereas expansive answers introducing factually correct details were penalised by cosine similarity metric, as exemplified here:

**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 (N=616 questions with no correct answer), 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 4**; 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 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 5** and **Table 2**).

#### 4. Discussion

*ToxTempAssistant* represents a practical implementation of LLMs in toxicology that aligns with current technological capabilities<sup>11</sup>. 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 challenging<sup>30-31</sup>.

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, underscoring that aggregate robustness can mask differing failure modes<sup>11</sup>. Moreover, recent evidence that errors correlate across models cautions simple model substitutions may not eliminate shared failure modes<sup>32</sup>.

Cosine similarity between model drafts and expert references averaged  $\approx 0.70$  across models, indicating substantial semantic resemblance. Cosine 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 BERTScore<sup>33</sup> and Sentence-BERT<sup>34</sup> 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)<sup>35-36</sup>. Accordingly, we treat cosine similarity as a coarse alignment signal rather than a fine-grained quality metric and pair it with completeness/abstention metrics. 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 approach<sup>37</sup>.

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. 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. Methodological, our design parallels work by Madhusudhan et al.<sup>37</sup>: 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 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.

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. 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 data are 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 Act<sup>17</sup>. 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<sup>38 39</sup>. 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 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; 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

All figures (main and supplementary), input documents, model outputs, and analysis scripts used in this study are available on Zenodo (DOI: 10.5281/zenodo.17047716).

*ToxTempAssistant* is open-source under the GNU Affero General Public License v3. Source: <https://github.com/johannehouweling/ToxTempAssistant>. An archived, citable release is available on Zenodo (DOI:10.5281/zenodo.15607642). [10.5281/zenodo.16749296](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