

SUPPLEMENTAL MATERIALS

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}

¹National Institute for Public Health and the Environment (RIVM), Bilthoven, Netherlands

²Dept of Translational Genomics, NUTRIM Institute of Nutrition and Translational Research in Metabolism, Maastricht University, Maastricht, Netherlands

³Independent researcher

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

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

ARTICLE HISTORY

Compiled December 30, 2025

Note. This PDF accompanies the main article and contains five supplementary items (S1–S5) with methods, results, and additional figures/tables. File names and persistent links are listed below.

Contents

S1 ToxTemp question set	2
S2 Overview of negative control documents used to test abstention behaviour	6
S3 Distributions of cosine similarity between model outputs and expert references across positive control ToxTemp documents	8
S4 Relationship between cosine similarity and LLM-assigned reference-answer quality	10
S5 Pairwise comparison of cosine similarity across the questions answered by all models	11

S1. ToxTemp question set

The ToxTemp is a standardized documentation framework for cell-based toxicological test methods, consisting of 77 questions across 11 sections. Reference: <https://doi.org/10.14573/altex.1909271>

Item	Question text
1. Overview	
1.1	Provide a descriptive title using normal language without technical terms or acronyms.
1.2	Please write an abstract on the cell-based toxicological test method in no more than 200 words:
1.2.a	Which toxicological target (organ, tissue, physiological/biochemical function, etc.) is modelled? (8.1)
1.2.b	Which test system and readout(s) are used? (4.1; 5.2)
1.2.c	Which biological process(es) (e.g. neurite outgrowth, differentiation) and/or toxicological events (e.g. oxidative stress, cell death) are modelled/reflected by your test method? (8.1)
1.2.d	To which (human) adverse outcome(s) is your test method related or could be related? (8.1; 9.2; 9.3)
1.2.e	Which hazard(s) do(es) your test method (potentially) predict? (8.1; 8.6)
1.2.f	Does the test method capture an endpoint of current regulatory studies? (9.5)
1.2.g	If the method has undergone some form of validation/evaluation, give its status. (9.4)
2. General information	
2.1	Provide the original/published name, as well as the potential tradename.
2.2	Provide the original deposition date of first version and date of current version.
2.3	This only applies to updated versions. If this is the original version, state 'original version'.
2.4	Normal text names often do not uniquely define the method. Therefore, each method should be assigned a clearly and uniquely defined database name.
2.5	Include affiliation.
2.6	Provide the details of the principal contact person.
2.7	For example, the principal investigator (PI) of the lab, the person who conducted the experiments, etc.
2.8	Supply number of supporting files. Describe supporting files (e.g. metadata files, instrument settings, calculation template, raw data file, etc.).
3. Description of general features of the test system source	
3.1	Describe briefly whether the cells are from a commercial supplier, continuously generated by cell culture, or obtained by isolation from human/animal tissue (or other).
3.2	Give a brief overview of your biological source system, i.e. the source or starting cells that you use. Which cell type(s) are used or obtained (e.g. monoculture/co-culture, differentiation state, 2D/3D, etc.)? If relevant, give human donor specifications (e.g. sex, age, pool of 10 donors, from healthy tissue, etc.).
3.3	List quantitative and semi-quantitative features that define your cell source/starting cell population. For test methods that are based on differentiation, describe your initial cells, e.g. iPSC, proliferating SH-SY5Y; the differentiated cells are described in section 4. Define cell identity, e.g. by STR signature (where available), karyotype information, sex (where available and relevant), ATCC number, passage number, source (supplier), sub-line (where relevant), source of primary material, purity of the cells, etc. Describe defining biological features you have measured or that are firmly established (use simple listing, limit to max. 0.5 pages), e.g. marker genes, surface antigens, doubling time, metabolic/transport capacity. Transgenic cell lines have additional requirements (transgene/vector, integration/deletion site(s), stability, etc.). Organoids and microphysiological systems (MPS) may require special considerations (Pamies et al. 2018; Marx et al. 2016).
3.4	Describe acceptance criteria (AC) for your initial cells. Which QC criteria must be fulfilled (e.g. pathogen-free)? Which functional parameters (e.g. reference-substance responses) are important? For iPSC maintenance: how is pluripotency controlled and stability ensured over passages? For primary cells: stability/identity of supply and stability of function (e.g. xenobiotic metabolism). Provide quantitative AC definitions where possible; include exclusion criteria.
3.5	Name known causes of variability of the initial cells/source cells (critical consumables; batch effects; handling steps; influencing factors). Include special considerations for genetically-modified cells and MPS (matrix chemistry/geometry; microfluidics; surface vs core cells). Provide recommendations to increase/ensure reproducibility and performance.

Continued on next page

Item	Question text
3.6	Describe the principles of the selected differentiation protocol, including a scheme/graphical overview (phases, media, substrates, manipulations). Include organoid/MPS-specific considerations where relevant.
3.7	Provide the SOP of the general maintenance procedure as a database link, including propagation outside experiments, purity, QC/AC per batch, valid passage numbers, GCCP/GIVIMP status, duration of batch use, frozen stock/cell bank preparation, and primary-cell sourcing/characterisation (inclusion/exclusion criteria).
4. Definition of the test system as used in the method	
4.1	Describe the test system as used in the test. If generation involves differentiation or complex manipulation, refer to 3.6. Provide culture protocol principles (e.g. collagen embedding, 3D structuring, mitotic inhibitors, hormones/growth factors). Report contaminating-cell percentages and subpopulation proportions in co-cultures; note differential cytotoxicity sensitivities if known.
4.2	What endpoints control that cultures are as expected at the start of toxicity testing (e.g. gene expression, staining, morphology, reference chemicals)? Describe AC and the analytical methods used (PCR, ATP, etc.). Specify required/forbidden values and actions if AC are not met; include historical-control performance where possible.
4.3	Describe AC for the test system (endpoints, parameters, required values, actions if unmet) and provide examples where applicable.
4.4	Give known causes of variability for the final test system state (consumables/batch effects; handling steps; influencing factors). Indicate positive/negative controls, expected values, accepted deviations, and recommendations to improve reproducibility/performance.
4.5	What is known about endogenous metabolic capacity (phase I/II) and transporter activity?
4.6	Are transcriptomics or other omics data available describing the test system (without compounds)? Briefly list/describe (type; where deposited/published).
4.7	Where does the test system differ from the mimicked human tissue, and what gaps of analogy should be considered?
4.8	Are elements of the test system protected by patents or other means?
4.9	If section 3 has not been answered, provide the SOP link for general maintenance (see 3.7).
5. Test method exposure scheme and endpoints	
5.1	Provide an exposure scheme (timelines, supplements/compounds, sampling) in the context of the overall culture scheme. Include medium changes, re-plating, compound re-addition, and critical supplements.
5.2	Define toxicity-testing endpoint(s) (e.g. cytotoxicity, migration). Indicate whether cytotoxicity is primary; list secondary endpoints. Describe reference/normalisation endpoints used for normalising the primary endpoint.
5.3	Describe analytical method principles and key steps sufficient for understanding (not full replication). For multi-endpoint designs, state whether measured in the same well (parallel) or independently. For imaging, summarise quantification and approximate numbers of cells imaged.
5.4	Provide machine settings, standards, processing and normalisation procedures; for imaging endpoints, provide the detailed algorithm (also covered in SOP; see 6.6).
5.5	MCC: list up to 10 manipulations/chemicals that plausibly change the endpoint; justify; describe expected data; highlight those used for routine performance monitoring and AC setting; indicate pathway up/down controls if available.
5.6	Positive controls: which are used and what are the expected signals/uncertainties? Are in vivo reference data and relevant thresholds known?
5.7	Negative controls: which are used and what is the expected background/noise? Provide rationale for concentrations; list any unspecific controls and why they are used.
5.8	Discuss apoptosis sensitivity/resistance; subpopulation sensitivity; marker availability; distinguishing slowed proliferation vs death; considerations for repeated dosing and short-term endpoints (and delayed cytotoxicity assessment, if applicable).
5.9	Which rule defines whether a run is within the normal performance frame? How is the decision documented? What actions are taken if AC are not met?
5.10	Indicate real data points per month (count three working weeks per month). Each concentration is a data point; calibration/AC controls are excluded; technical replicates count as one data point. Indicate repeated-measures extent and justify.
6. Handling details of the test method	
6.1	Overview of volumes, labware/instruments, temperature/lighting, media/buffers for dilution, solvent decision rules, solubility testing, stock preparation/verification/storage, dilution preparation, sterility filtering, and compound addition details (where/volume/timing relative to medium change).

Continued on next page

Item	Question text
6.2	How are daily procedures documented (lab book/templates)? How are concentration calculations documented? How are plate maps defined/reported? (Also in SOP; see 6.6.)
6.3	How is pipetting timing established and documented? How is adherence to plate maps ensured? How are intermediate steps and errors documented? What well-use pattern is followed? (Also in SOP; see 6.6.)
6.4	How is the concentration range defined (single vs serial dilutions; dilution factors; number of concentrations)? Rules for starting dilutions? For functional endpoints, how is test concentration defined (e.g. EC10 viability for gene expression)? (Also in SOP; see 6.6.)
6.5	Problematic compounds (interference, solubility, precipitation, fluorescence/colour); hard-to-control variables; critical handling steps; robustness issues; known pitfalls/operator mistakes.
6.6	SOP format (ideally DB-ALM): https://ecvam-dbal.m.jrc.ec.europa.eu/home/contribute . Refer to additional files for sections 3–4 details. Indicate whether deposited in a database and whether externally reviewed.
6.7	Specialised instrumentation needs (non-standard labs), custom-made material, or non-commercially available equipment.
6.8	Variations/modifications/extensions: (a) other endpoints, (b) other methods for same endpoint, (c) other exposure schemes, (d) experimental variations (medium, inhibitors/substrates, etc.).
6.9	Related tests (names and database names) with short descriptions and differences; indicate use for high-throughput transcriptomics/deep sequencing if applicable.
7. Data management	
7.1	Data format: describe raw data and provide an example (e.g. plate-reader export). Provide an example of processed data suitable for display and comparison. If non-proprietary, include a template; provide an EU-ToxRisk-style example schema if relevant.
7.2	Outlier definition/handling and documentation; general frequency.
7.3	Raw-to-summary processing steps (background correction through normalisation).
7.4	Overall test result derivation (curve fitting/model/software; uncertainty estimation; handling non-monotonic curves or other hard-to-fit features).
7.5	Storage duration, backup procedures/frequency, version identification.
7.6	Metadata documentation/storage and linkage to raw data; what metadata are (or should be) stored.
7.7	Example metadata file (if available) or state that formats are predefined by a named project.
8. Prediction model and toxicological application	
8.1	Scientific rationale linking test-method data to in vivo adverse outcomes; target modelled; processes/events modelled; related adverse outcomes; predicted hazards.
8.2	Benchmark response statistics (threshold/variance) for dichotomised, pseudo-dichotomised, and continuous outcomes; rationale for thresholds; model limitations; hit definition for screening.
8.3	Prediction-model setup (training/test sets; classifiers/statistics; documentation); one- vs two-sided behaviour and handling of opposite-direction effects.
8.4	Performance parameters (noise within/between assays; SNR; z-factor; specificity; sensitivity; uncertainty; detection limit; LOD/LOQ; inter-operator variation; historical controls).
8.5	Parameters for free compound concentration (medium/cell composition; volumes; surface area); IVIVE strategies/data; IVIVE-relevant considerations.
8.6	Expected strengths/failure modes across chemistries, mixtures/UVCBs, excluded domains; known interferences (fluorescent/coloured, etc.).
8.7	Fit into test batteries; restrictions; strengths/weaknesses; comparison to similar tests; gaps filled; tier placement; complementary assays.
9. Publication/validation status	
9.1	Published literature on the test and deviations from published descriptions; key publications and what they cover; prioritised further publications with brief notes on obtainable information.
9.2	AOP linkage (form; MIE/KE coverage); references and AOP-Wiki status if applicable.
9.3	Mechanistic validation (omics; MCC controls); evidence of reflecting relevant human biology/signalling/tissue organisation.
9.4	Qualification/pre-validation/validation activities (ring trials, etc.) and compounds/libraries tested.
9.5	Linkage to OECD Test Guidelines or other regulatory guidance (how/which).
10. Test method transferability	
10.1	Required experience; operator training; training/experience required for smooth performance.
10.2	Transfer to other labs; multi-operator use; inter-laboratory variability; transfer procedures and performance.

Continued on next page

Item	Question text
11. Safety, ethics and specific requirements	
11.1	Legal requirements and hazards affecting operators/bystanders/others (including waste).
11.2	SDS availability/storage for hazardous reagents and compounds; safe-handling procedures; exposure scenario availability.
11.3	Special permits/facilities/ethical approval requirements (and approval document).
11.4	Patents/other protections on method elements; type of protection and where licences/elements can be obtained.

S2. Overview of negative control documents used to test abstention behaviour

Table S2 lists the negative control documents used to evaluate ToxTempAssistant’s abstention behaviour, including their titles, summaries, manually-assigned difficulty levels, underlying rationale, and references.

Table S2. Negative control documents used to evaluate ToxTempAssistant’s abstention behaviour

ID	Title	Summary	Rationale / relevance	Diff.	Reference
pdf1	European State of the Climate 2023	Annual synthesis of European climate conditions and notable events in 2023.	Out of scope for toxicological test-method documentation; may prompt broad health-risk narratives rather than procedural details required by ToxTemp items.	Low	https://climate.copernicus.eu/esotc/2023
pdf2	White Paper on Artificial Intelligence: A European approach to excellence and trust	EU strategy paper on AI governance and use across sectors, with occasional health/biomedical references.	Mentions decision systems and data analysis but provides no toxicological test methods or cell-based assay protocols.	Low	https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en
pdf3	The Burden of FAIRness: A Longitudinal Study of Metadata-Induced Fatigue in Biomedical Researchers	Mock study in scientific format (abstract/methods/results) on fatigue from FAIR stewardship.	Superficially resembles a paper but contains no toxicology content and no method description aligned to ToxTemp.	Low	N/A
pdf4	Human exposure to bisphenol A	Review of human exposure pathways and reported BPA levels.	Contains overlapping vocabulary (exposure, thresholds) but lacks an <i>in vitro</i> toxicological test-method protocol.	Medium	DOI: 10.1016/j.envres.2007.08.008
pdf5	ECDC Annual Epidemiological Report — seasonal influenza	Surveillance summary and epidemiological trends for seasonal influenza.	Includes sampling/surveillance concepts but no chemical exposure design or cell-based toxicology method documentation.	Medium	https://www.ecdc.europa.eu/en/publications-data/seasonal-influenza-annual-epidemiological-report-20232024
pdf6	DMSO induces drastic changes in human cellular processes and epigenetic landscape <i>in vitro</i>	Multi-omics study of DMSO effects in human 3D microtissues.	Biological assay content and experimental details are present, but the paper is not a toxicological test-method description in the ToxTemp sense.	High	DOI: 10.1038/s41598-019-40660-0
pdf7	Application of the Virtual Cell-Based Assay	Computational simulation of <i>in vitro</i> chemical fate and exposure.	Uses “assay” framing but is modelling-focused and lacks a wet-lab protocol; likely to trigger partial but misleading matches to ToxTemp items.	High	https://publications.jrc.ec.europa.eu/repository/handle/JRC107407
pdf8	Optimization of cell viability assays to improve replicability in preclinical cancer research	Methodological work on viability assays and replicability in oncology research.	Contains assay and exposure terminology, but the domain (oncology screening) and reporting targets differ from toxicological test-method documentation.	High	DOI: 10.1038/s41598-020-62848-5

S3. Distributions of cosine similarity between model outputs and expert references across positive control ToxTemp documents

Figure S3 shows the distributions of cosine similarity for positive control documents, per LLM backend.

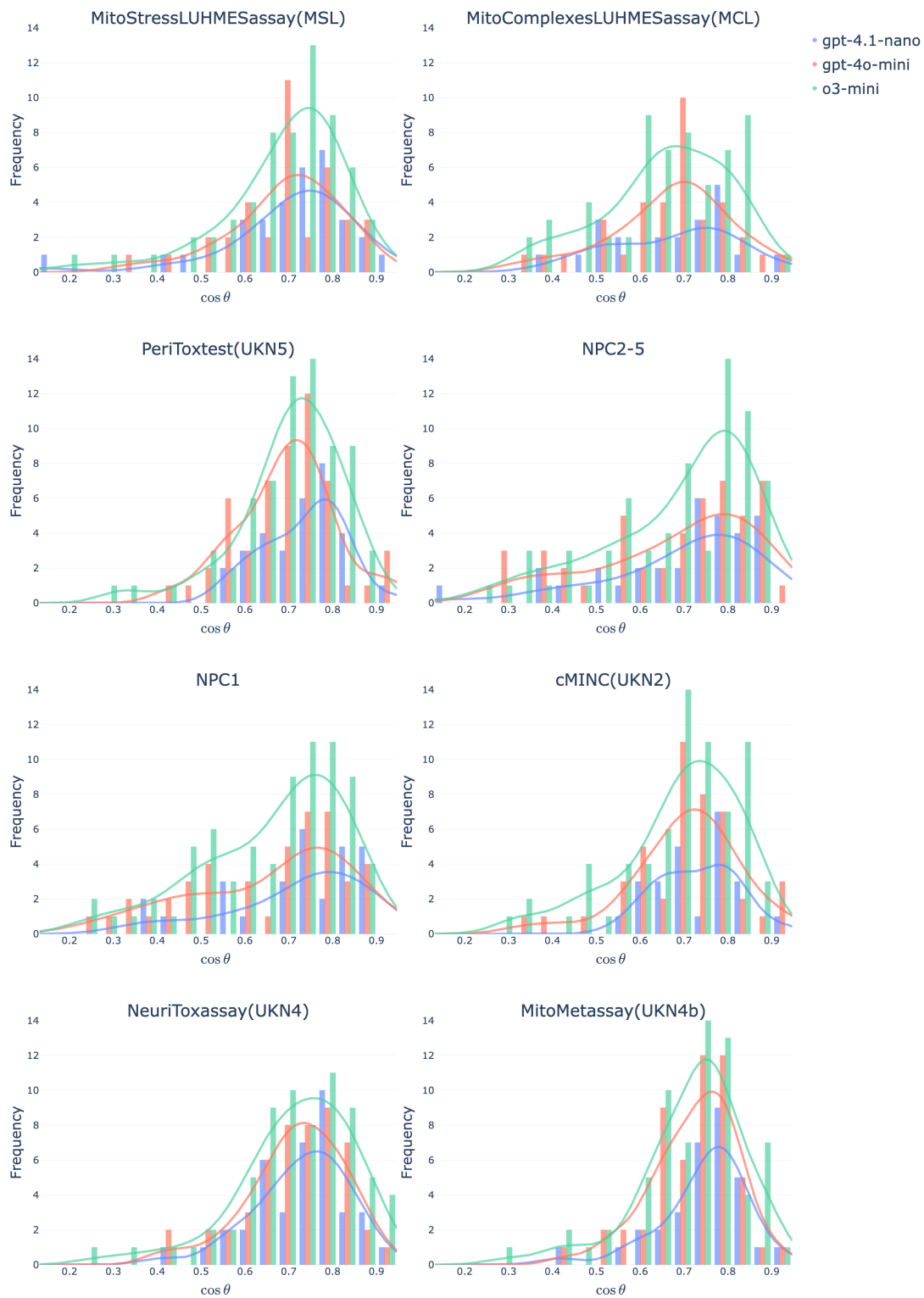


Figure S3. Positive control evaluation of ToxTempAssistant using eight publicly available, factually complete ToxTemp templates as input. Each panel corresponds to one ToxTemp document, with histograms showing the distribution of cosine similarity ($\cos \theta$) between model-generated responses and expert reference answers. Bar heights indicate question frequencies, coloured by model (GPT-4.1-nano, blue; GPT-4o-mini, orange; o3-mini, green). Lines represent kernel density estimates.

S4. Relationship between cosine similarity and LLM-assigned reference-answer quality

Figure S4 presents cosine similarity by reference-answer quality (Low, Medium, High) for each model (sample sizes shown above violins). Visual separation is clearest between Low and Medium/High, with comparatively modest shifts from Medium to High.

Monotonic trends were evaluated using Spearman correlation (ρ). Across all models, cosine similarity increased significantly with reference quality. This trend was strongest for `o3-mini`, followed by `gpt-4o-mini`, with `gpt-4.1-nano` showing only a weak correlation.

Distributional differences across quality levels were confirmed using the Kruskal–Wallis test. Significant effects were observed both overall and within each model. Post-hoc pairwise comparisons using Dunn’s test (Holm- and Bonferroni-corrected) showed that Low-quality references differed strongly from both Medium and High, whereas Medium vs. High did not reach significance. Cliff’s Δ (overall) supported this pattern: a pronounced jump from Low to Medium with a plateau thereafter.

Sample sizes differed across both models and reference-quality categories (e.g., “Medium” was more frequent than “Low” or “High”). The non-parametric tests used (Spearman, Kruskal–Wallis, Dunn’s, Cliff’s Δ) are robust to unequal group sizes, but pooled “overall” results are necessarily weighted toward larger groups.

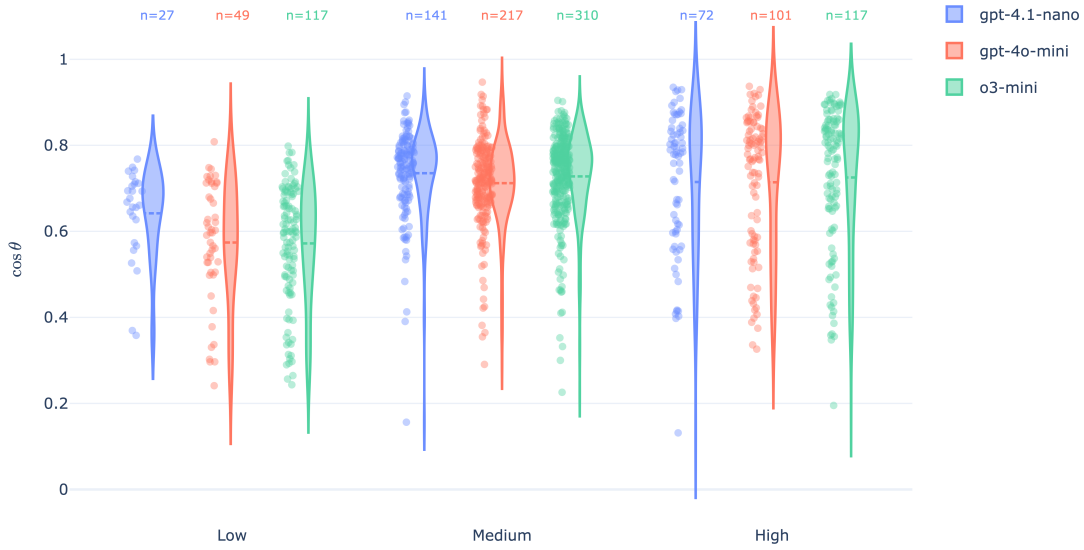


Figure S4. Cosine similarity ($\cos \theta$) distributions stratified by LLM-assigned reference-answer quality (Low, Medium, High). Violin plots show the score distributions for each model, with sample sizes indicated above each violin

S5. Pairwise comparison of cosine similarity across the questions answered by all models

We compared cosine similarity at the question–answer level across the $n = 206$ items answered by all three models. Paired differences were tested for normality (Shapiro–Wilk) before applying paired t -tests or Wilcoxon signed-rank tests.

o3-mini exceeded **gpt-4o-mini** (mean $\Delta = +0.014$, median $\Delta = +0.008$), with paired differences approximately normal ($p = 0.116$) and significant in a paired t -test ($t = 2.943$, $p = 0.004$). Comparisons involving **gpt-4.1-nano** yielded non-normal paired differences (Shapiro–Wilk $p < 0.05$) and were not significant in Wilcoxon tests (**o3-mini** vs. **gpt-4.1-nano**: $p = 0.160$; **gpt-4o-mini** vs. **gpt-4.1-nano**: $p = 0.073$). Taken together, **o3-mini** shows a small but statistically significant advantage over **gpt-4o-mini**, whereas differences involving **gpt-4.1-nano** are not statistically conclusive.

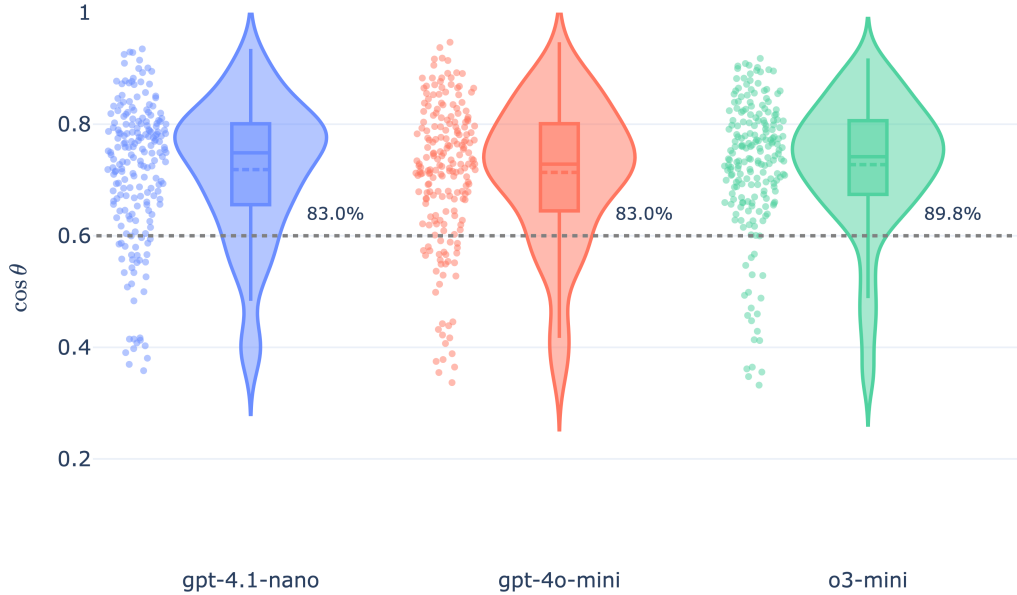


Figure S5. Cosine similarity ($\cos \theta$) distributions for the $n = 206$ questions answered by all three models (**o3-mini**, **gpt-4.1-nano** and **gpt-4o-mini**). Violin plots show the score distributions with embedded boxplots (median, quartiles), overlaid with individual question–answer points. The dashed line marks the evaluation threshold ($\cos \theta = 0.6$), with the percentage of responses exceeding this threshold shown above each distribution.