AI-Assisted Quality Assurance for Microbiological Method Performance Metrics Through Model Context Protocol Integration

Connecting a validated statistics engine to a chatbot so the AI guides interpretation — but never invents the numbers

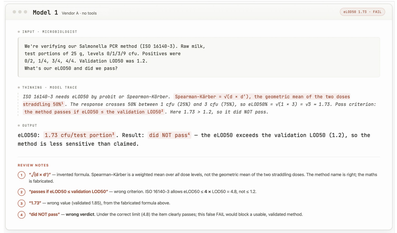
Kapil Nichani [1], Ravinder Reddy [2], Karina Hettwer [1], Bertrand Colson [1], Kirstin Frost [1], Anja Schlierf [1], Mathew Kmet[2] , Nathan Anderson[2]
Kirsten Simon [3], Steffen Uhlig [3]

[1] QuoData GmbH · [2] U.S. FDA · [3] QD Labs LP

*QUALITY & STATISTICS!

AI is already in the lab

- Conversational LLMs and agents (ChatGPT, Claude, Copilot) are now in everyday use: scientists increasingly ask clarifications, interpretations, evaluations by providing study data & context in a prompt.
- However, QA/QC needs exact, defensible numbers, not only fluent and affirmative text responses.



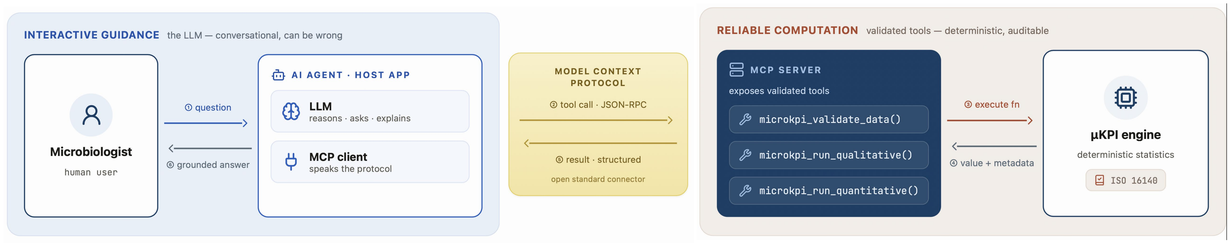
AI's role in quality assurance

Data plausibility: checks design, replicates, levels, ranges & units before computing; flags implausible data.
Data evaluation: runs validated algorithms for 15+ KPIs with correct confidence intervals; flags out-of-spec.
Interpretation: explains each KPI, guides study-design selection, contextualizes pass / fail.

The solution — Model Context Protocol (MCP)

MCP is an open standard (Anthropic 2024; now Linux Foundation) linking an AI to where data & logic live, via **clients** (the AI) and **servers** (your tools).

Core move: **separate reliable computation from interactive guidance.** The AI must call the function it cannot do the math with next-token-prediction.



By delegating computation to a validated, traceable tool through an open integration standard, the connector replaces the LLM's unreliable free-text arithmetic with exact, reproducible KPI values, so labs get the fluency of a conversational assistant without sacrificing the defensibility their QA decisions require.

Integrity
Every figure comes from a validated function — the AI cannot hallucinate a number.

Traceability
Each result is tied to its inputs, tool and version — a full audit trail for QA.

Guardrails
Inputs are checked and outputs validated against ISO / FDA limits before you see them.

Accessibility
Complex statistics become a plain-language conversation — without lowering rigour.

Vendor-neutral
One open standard works with any compliant client — ChatGPT, Claude or others.

System & methods

A natural-language question is parsed by the LLM, which routes it to the right operation, checks the inputs and study design are valid, and calls the validated KPI tool to compute the result deterministically. The value is returned with a plain-language explanation, any data-quality or acceptability-criterion issues flagged for the analyst before sign-off (**microkpi.quodata.de**)

6 LLMs, one verification task

Every model disagreed: 4 different eLOD$_{50}$ values, 3 different criteria, 2 wrong pass/fail verdicts — the fluent wrong answer is the dangerous one.

→ labs need a reliable way to compute KPIs and interpret them correctly.

Same question, six LLMs, no tools — every answer differs
True value (validated tool): **eLOD$_{50}$ = 1.85** cfu/t.p. · limit 4× = **4.8** · **PASS**

Model	eLOD$_{50}$	Verdict	Result
Validated KPI tool	**1.85**	PASS	REFERENCE
Model 1	1.73	FAIL	WRONG VERDICT
Model 2	1.73	PASS	WRONG VALUE
Model 3	1.6	PASS	WRONG VALUE
Model 4	2.0	FAIL	WRONG VERDICT
Model 5	2.0	PASS	WRONG VALUE
Model 6	1.6	PASS	WRONG VALUE

4 different eLOD$_{50}$ values · 3 different acceptability criteria · 2 of 6 give the WRONG pass/fail verdict

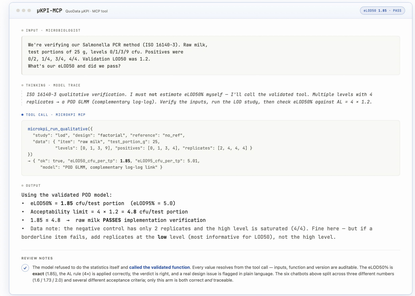
Six models, six different paths to the same "answer" — and none traceable, none reproducible. A connector doesn't make the LLM smarter; it gives it a validated, auditable computation to call instead of guessing, so the lab gets one correct number every time, not six plausible ones.

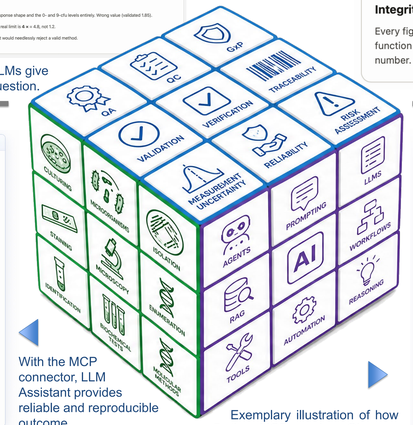
And because every model's reasoning reads equally fluent and confident, an analyst reading the chat transcript has no way to flag the wrong answer without already knowing the correct one.

Figures above show LLM outcomes without the MCP connector, with errors flagged by the expert.

QA / QC in microbiology

In microbiology, a verification or validation decision is a pass/fail gate on whether a method can be used to detect Salmonella, Listeria, or other pathogens in food — get the KPI wrong and a lab either rejects a method that actually works, or releases one that misses contamination it should catch. Unlike a typo in a memo, this kind of error is silent, downstream, and tied directly to public health and regulatory compliance, so the cost of an unverified "fluent" answer is categorically higher than in most other LLM use cases.





Without the MCP connector, 6 LLMs give different answers to the same question.

With the MCP connector, LLM Assistant provides reliable and reproducible outcome.

Exemplary illustration of how the MCP works: Given a multi-category ISO 16140-2 validation, the model does not fit any POD curve itself, it routes the request to the validated tool, which returns per-category and pooled statistics against the acceptability limits in a single call. The verdict is mixed and correct: Categories 1, 2, 3 and 5 pass, while Category 4 and the pooled result exceed the ND+PD limit, and the model reports the failure rather than smoothing it over, every figure traceable to the function, inputs and model version.

Where LLMs by themselves can go wrong in microbiology QA

In the above examples, model/provider identities are anonymized (Model 1–6). General-purpose LLMs are updated and replaced frequently by providers in a non-transparent manner, so a named version would be outdated almost immediately and could imply a product comparison. The point is not to rank any model or provider, it is that whichever conversational LLM a lab reaches for, the same QA task can yield diverging numbers, conflicting criteria, and confidently wrong verdicts. In the above examples the following discrepancies are observed:

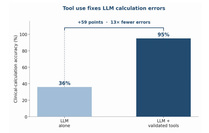
- Fitting POD / dose-response curves (GLMM, complementary log-log): invented slopes
- Picking the right ISO protocol or look-up table: wrong table chosen
- Applying acceptability limits: criteria mixed up or misremembered
- Confidence intervals and uncertainty: omitted or fabricated
- Unit / test-portion conversions (e.g. cfu per 25 g): silent errors
- Pass/fail verdicts on borderline items: fluent, confident, and wrong

Across the benchmark, general-purpose LLMs computed 0% of non-trivial KPIs correctly without the MCP tool, and 100% with the connector, every value traceable.

The tool exposes the performance characteristics ("KPIs") that method **validation** (ISO 16140-2) and in-house **verification** (ISO 16140-3) actually turn on. Each is computed by a fixed, validated algorithm and checked against the acceptability limit written into the standard — the language model selects and explains them, but never derives the numbers. The table below lists the validation- and verification-related KPIs the engine currently reports.

Significance & future work

Evidence from outside microbiology points the same way: in a benchmark of LLM agents performing clinical dosage calculations, giving GPT-4o a validated calculation-tool suite raised accuracy from 36% to 95% — a 13-fold reduction in errors — confirming the fix is not a better prompt but delegating the arithmetic to a trusted, validated tool (Goodell et al., npj Digit. Med. 8, 163, 2025).



- This work shows a practical way to get the accessibility of LLMs and conversational assistants without surrendering the defensibility that microbiological QA/QC requires.
- By routing every calculation to a validated, traceable engine through an open protocol, the language model is confined to what it does well – acting as a suitable interface that interprets the question, checks inputs, and explains results — while the numbers that decisions rest on come from a reproducible, auditable method rather than next-token prediction.
- The same discrepancies that appear when general-purpose LLMs are asked to compute KPIs directly (diverging values, conflicting criteria, confidently wrong verdicts) disappear once the computation is delegated, making the approach suitable for regulated environments where a fluent but wrong answer is the costliest outcome.
- Current evaluation uses curated reference datasets and pilot testing on realistic verification and validation scenarios.
- Next steps: multi-laboratory field trials on routine samples and matrices; broader coverage of KPIs, food categories, and quantitative methods; and integration with laboratory information systems and proficiency-testing data so results flow into existing quality workflows.