SEMINAR REVIEW

Life Science AI Exchange

Illustrations of AI Value: Benchmarking and Ongoing Measurement

25 March 2026 • Pistoia Alliance LSAIE Programme • Open Webinar
Reviewed by Rob Gill

The first seminar of the 2026 Life Science AI Exchange programme opened with a productive tension: the life sciences industry is simultaneously over-confident about what AI can do and under-prepared to measure whether it is doing it. Over the course of two hours, four speakers — from academia, large pharma, and the technology sector — presented work that illuminated both sides of that gap with unusual precision. This review focuses on the substance of those contributions and on the cross-cutting conclusions they invite.

1. Can LLMs Actually Do Chemistry?

Nicholas Runcie — Oxford Protein Informatics Group, University of Oxford

The question sounds provocative, but until very recently the answer was an unambiguous no. Nicholas Runcie opened the session by recounting what had become received wisdom: language models could not read or write molecular structures expressed as SMILES strings. They could not reliably count atoms, describe functional groups, or reason about chemical properties. This limitation had driven the field towards agentic architectures in which an LLM orchestrates calls to specialist chemistry tools rather than tackling the chemistry directly.

That consensus has now been overturned, and Runcie's contribution was to document the overturning rigorously. His ChemIQ benchmark — published in the Journal of Chemical Information and Modelling — was designed around a single, foundational question: do language models understand what molecules look like?

The benchmark design

ChemIQ differs from earlier chemistry benchmarks in three important respects. First, it is tightly focused on molecular comprehension rather than the broad sweep of general chemistry knowledge tested by competitors such as ChemBench. Second, all 816 questions across eight categories are algorithmically generated from Python scripts rather than drawn from human-authored exam material. This means the benchmark is a living entity: as models improve — or as suspicions arise that a model has been trained on the test set — new questions can be generated on demand. The benchmark evolves with the field rather than becoming stale. Third, all questions require a short free-text answer; there are no multiple-choice options. Runcie was emphatic on this last point: multiple-choice chemistry questions can often be answered by eliminating obviously absurd options, producing flattering accuracy scores that do not reflect genuine chemical reasoning.

The three core competencies assessed progress from basic to applied: whether a model can parse a molecular graph (counting atoms, identifying functional groups), whether it can describe a molecule in natural language, and whether it can tackle genuine reasoning tasks including reaction prediction and NMR structural elucidation.

What the results showed

The comparison between GPT-4o (a non-reasoning model) and o3-mini (a reasoning model) is stark. GPT-4o scored zero on 200 IUPAC naming questions: given a SMILES string, it

could not produce the correct systematic name for a single molecule. o3-mini, using extended chain-of-thought reasoning, answered nearly 30% correctly. That is still a long way from expert performance, but the jump from zero to twenty-eight percent in a single model generation is a step change that the field had not anticipated.

> *A year ago it was accepted that language models could not even read or write molecular structures. Now they are solving some of the hardest problem-solving questions in chemistry. This is kind of absurd to me.*

The NMR structural elucidation result was the one that drew the most audible reaction in the Q&A. Runcie provided Gemini 2.5 Pro with simulated 1D and 2D NMR spectra — five separate datasets in text format — and asked it to identify the molecule. For a compound with 25 heavy atoms (a structure that would give pause to a postgraduate chemist in an exam setting), Gemini arrived at the correct answer. More striking than the result itself was the reasoning trace: the model systematically identified likely ring systems from the spectra, proposed and tested intermediate hypotheses, and checked its conclusions for internal consistency. Its language was, as Runcie put it, “consistent with how I would have gone about the question.”

The pattern across all models and all question categories was consistent: more reasoning time produced better answers. There was, however, an interesting inversion of human difficulty. Tasks that are trivial for chemists — counting the carbon atoms in a molecule — proved surprisingly hard for language models, while tasks that demand deep expertise, such as NMR elucidation, were handled with unexpected competence. The models are not simply simulating human cognition; they are developing their own strategies for navigating chemical space.

Practical implications

The benchmark has already demonstrated its value as an operational tool. When GPT-5 became available via API, Runcie ran the full ChemIQ suite overnight and had a comparative analysis ready by morning. The ability to rapidly profile a new model against a rigorous, task-specific benchmark is precisely the kind of infrastructure the field needs as frontier models proliferate.

Runcie was candid about the limits of his current work. Reaction kinetics, complex multi-step synthesis pathways, and molecular property optimisation remain largely unexplored. His view is that hybrid approaches — an LLM orchestrating calls to specialised computational chemistry tools for tasks that exceed its direct capability — remain the right architecture for many applied problems. The critical change is that the LLM can now act as a genuine chemical reasoner rather than a mere coordinator, which substantially expands what such architectures can do.

He is currently building the next generation of this work: a benchmark assessing agentic chemistry systems on genuine research tasks from Oxford’s own laboratories. He is seeking collaborators willing to share internal chemistry data and will be advertising the collaboration via LinkedIn. For PA members with relevant datasets, this represents a timely opportunity to contribute to infrastructure that will benefit the whole field.

2. Knowledge Graphs, LLMs, and the Danger of Overconfident Models

Frederik Steensgaard Gade — BioAI, Novo Nordisk / Technical University of Denmark

Where Runcie's work concerned what LLMs can now do, Frederik Steensgaard Gade's presentation was a careful study of where they still fall short — and, crucially, of how that shortfall can be hidden by a seductive but unreliable evaluation method. His setting is early-stage pharmaceutical research, where scientists need to rapidly map and reason over large bodies of biomedical literature; his tool of choice is a combination of knowledge graphs and large language models.

The complementary paradigm

Knowledge graphs and LLMs address opposite failure modes. An LLM produces fluent, plausible-sounding output but cannot guarantee that its claims are grounded in evidence. A knowledge graph is rigid and cannot generalise beyond its schema, but everything in it has an explicit provenance. The combination — using a knowledge graph as the factual foundation and an LLM as the reasoning and querying layer — is an increasingly common architectural pattern, and for good reason. When a biological researcher asks why a particular gene is implicated in a disease, they need not just an answer but a chain of evidence they can verify. The knowledge graph supplies the evidence; the LLM supplies the articulation.

The practical challenge Steensgaard Gade addressed is knowledge graph construction: how do you populate the graph in the first place? The conventional approach requires significant manual annotation by domain experts, which is time-consuming and inflexible. The appeal of using a zero-shot LLM — simply instructing it to extract entities and relationships from PubMed abstracts without prior training on your specific schema — is obvious. The question is whether it works well enough to trust.

The LLM-as-judge problem

The research tested multiple models, including Gemini 1.5 and several BERT-based biological language models, on a named entity recognition (NER) and relation triplet extraction task. An initial evaluation used the LLM itself to assess the quality of its own output — a method increasingly common in the field and sometimes referred to as 'LLM-as-judge.' The result was encouraging: a self-reported precision of 93%.

> *The LLM was almost confident enough in its own performance to send us down the wrong path. The lesson is that self-assessed confidence scores are not a proxy for real performance — they need to be validated against ground truth.*

The team then ran the same models against a rigorous external benchmark: approximately 13,000 text samples drawn from 61 different biomedical corpora, each manually annotated by domain experts, covering 150 distinct entity types. The results were substantially less flattering. Zero-shot NER performance was below the threshold considered acceptable for production use across all but a small number of well-studied entity categories. Triplet extraction performance — identifying not just entities but the relationships between them — was considerably worse again. Critically, re-running the benchmark with the latest reasoning models produced no meaningful improvement on extraction tasks.

The implication is uncomfortable. A team moving quickly — which in pharma often means moving under competitive pressure with limited time for validation — might have accepted

the 93% self-reported figure and built a production pipeline on foundations that would quietly produce unreliable knowledge graphs. The risk is not that the system fails visibly; it is that it fails silently, populating a research knowledge base with plausible but incorrect relationships.

What does work, and the path forward

Fine-tuned BERT-based models, trained specifically on biological corpora with human-annotated ground truth, remain the performance benchmark for NER and extraction tasks. They are not, however, practical for rapid prototyping in early research: the annotation effort required to train them is substantial and the resulting models are inflexible to schema changes. Few-shot LLMs — providing a handful of labelled examples at inference time — improve on the zero-shot baseline but do not close the gap.

Steensgaard Gade's current production approach reflects this reality pragmatically. For well-understood entity types (genes, diseases, basic interaction types), the knowledge graph is grounded in outputs from a high-quality fine-tuned model, effectively providing human-level annotation at scale. Zero-shot LLMs are then used only at the margins — to add contextual granularity that the fine-tuned model does not capture. The LLM is deployed where its strengths (fluency, flexibility, generalisation) are genuinely needed, not where its weaknesses (inconsistent extraction, overconfident self-assessment) would be exposed.

An interesting secondary finding concerns the nature of the extraction task itself. LLMs, as generative models, are architected to predict the next token; BERT-based models, as token classifiers, are architecturally better suited to the labelling task that NER requires. This is not simply a matter of scale or training data: it reflects a fundamental difference in what these model families are designed to do. The field's enthusiasm for using general-purpose LLMs for everything should be tempered by this structural reality.

3. Measuring What AI Is Actually Worth

Peter Henstock & Dan Houston — Insight AI Institute

The third presentation shifted the lens from technical capability to business value, and it started with a deliberately unsettling observation. Despite the industry's enthusiasm for AI, headlines about AI failures are not hard to find. The MIT survey suggesting that 95% of generative AI projects fail — admittedly in the context of prototypes and early development, not mature deployments — had attracted considerable attention. Microsoft's CEO had publicly questioned whether AI was adding measurable value. These are not fringe views; they reflect a genuine difficulty in turning AI capability into demonstrated return.

Peter Henstock's opening argument was that pharma's particular difficulty with AI valuation is structural, not incidental. The drug development timeline is 10 to 15 years. Clinical success rates are low. The decisions that AI might improve are indirect: a better compound selection model does not cause a drug to succeed; it shifts the probability by some amount that will not be measurable until the programme concludes, if it ever does. The industry also operates in a highly regulated environment that makes the A/B testing routine in consumer technology effectively impossible. And unlike classical software, AI outputs are probabilistic rather than deterministic, which means that even a good model will sometimes be wrong in ways that are hard to distinguish from failure.

A framework for thinking about AI value

Henstock proposed a three-level value hierarchy that provides a more disciplined approach than the vague language of 'productivity improvements' that characterises much of the current discourse.

The first level is the model metric: accuracy, area under the curve, hallucination rate, F1 score. These are the numbers that data scientists habitually report, and they are necessary but insufficient. A QSAR model with an AUC of 87% may sound impressive, but the number carries no information about whether 87% is good enough for the decision it is supporting, or how it compares to the baseline it is meant to improve upon.

The second level is task utility: does the AI tool actually change the decision that needs to be made? In the compound prioritisation example, this means asking whether the model's recommendations are being used — and Henstock noted that adoption rates of 60 to 70% are typical. Every recommendation that is ignored represents value that was modelled but never realised. Explainability and user experience are not cosmetic concerns; they are direct determinants of the fraction of modelled value that is actually captured.

The third level is business impact: time saved, experiments avoided, costs reduced, revenue influenced. This is the number that justifies the investment, and it requires working backwards through adoption rates and probabilistic improvement factors from a baseline that is often poorly documented.

> *If you can't describe what the AI changes and who uses it, you can't calculate what it's worth. The value is in the decision that gets made differently, not in the model that makes the recommendation.*

The build-versus-prompt decision

Dan Houston's contribution grounded this framework in practical project decisions. At Insight, every AI project begins with the same four questions: what is the workflow, what is

the impact, what is the cost to build, and what solution optimises those factors? The answers often point towards simpler solutions than engineers naturally gravitate towards.

His worked example — automating the extraction of data from vendor reports into internal systems — is representative of a very large class of pharmaceutical AI use cases. The current state: a scientist manually transcribes data from PDFs, taking around 30 minutes per report. The AI-enabled state: a custom ChatGPT configuration performs the first pass, reducing the task to five minutes of human review. The cost of building this: approximately £600 in prompt engineering time for a one-day engagement. The cost of building a fully automated pipeline that processes reports without human intervention: approximately £200,000.

For a company processing 100 reports per month, the £600 solution captures the great majority of the available value. For a company processing 10,000 reports per month, the business case for the £200,000 pipeline becomes compelling: the remaining human effort is now worth over £1 million annually and can be redirected. The point is not that one approach is always right but that the decision should be made explicitly and numerically rather than by default.

Houston also raised a concern that has become increasingly salient as models improve rapidly: the risk that solutions built with substantial engineering effort become obsolete within months. A complex orchestration layer that was necessary when models struggled with multi-step tasks may be redundant when the next generation of models handles those tasks natively. Solutions built predominantly around prompts and data integrations are inherently more durable than those built around compensating for specific model limitations.

On talent, both speakers converged on a perhaps counterintuitive conclusion: the most effective people for leading AI projects are software engineers willing to learn the domain, not domain experts learning to code. The former tend to ask the productive naïve questions that expose assumptions; the latter may unconsciously protect those assumptions.

4. Industry Intelligence: What the Sector Is Actually Experiencing

Vladimir Makarov — Pistoia Alliance

Vladimir Makarov concluded the formal presentations with a synthesis of three independent data sources: an audience poll conducted at the start of the session, a survey of 54 senior pharma and biotech leaders conducted by Zuhlke Consulting (a Pistoia Alliance member), and a discussion at the Swiss chapter of the Life Science Informatics Forum. Taken together, these sources provide a triangulated view of where the life sciences sector stands with AI in early 2026 — and, more valuably, where the gaps are largest.

The audience poll

Question	Finding
Data governance maturity	Ad hoc or inconsistent in the majority of respondents' organisations.
Is AI literacy a problem?	Yes — confirmed by a clear majority, reflected in low training completion rates.
Familiarity with agentic AI	Most respondents are already using or piloting agentic systems.
Stage of agent deployment	Majority are in pilot or early production — consistent with familiarity results.
Biggest barrier to scaling agents	1. Skills gap 2. Security concerns 3. Regulatory uncertainty
Importance of vendor-neutral interoperability standards	Two-thirds rated as 'important' or 'critical' to their AI strategy.

The juxtaposition of the third and fourth rows with the fifth deserves attention. Respondents are simultaneously moving agents into production and reporting that the skills to do so competently are their primary constraint. This is not unusual in a fast-moving technology landscape, but it is a risk profile that should concern anyone responsible for AI governance in these organisations.

Industry survey — the Zuhlke/Pistoia Alliance findings

The Zuhlke/PA survey, which drew on interviews with 54 senior leaders across pharma, biotech, and CRO organisations, identified a consistent pattern across 19 technology domains: high internal demand, high perceived strategic importance, and comparatively low organisational maturity. The gap between what organisations want to do with AI and their demonstrated capacity to do it is the defining feature of the sector's current position.

A secondary finding — which Makarov called the 'value loss funnel' — connects directly to the themes raised by Henstock and Houston. Plotted from pilot to production to measured value delivery, the proportion of projects that reach each subsequent stage falls sharply. Projects that generate genuine, quantified business value are a small fraction of those that begin as exciting prototypes. The funnel is not unique to pharma, but the combination of long timelines, indirect outcomes, and regulatory constraints makes it particularly acute in this sector.

Mapping technology domains against collaboration potential and organisational maturity reveals the clearest opportunity space for the Pistoia Alliance. The quadrant of high collaboration potential and low organisational maturity — where individual companies are not yet capable enough to go it alone but would benefit from pre-competitive work — contains four priority areas: validation frameworks, ROI and value measurement, organisational ownership models and data quality standards, and AI literacy.

Swiss Life Science Informatics Forum priorities

The Swiss chapter discussion, reflecting the perspectives of pharmaceutical professionals and managers, produced a complementary but not identical list. High-priority topics included the legal status of AI rights and content licensing for scientific literature (a practical concern that is often overlooked: standard journal subscriptions frequently do not cover automated text mining, and the legal position for knowledge graph construction from proprietary data sources is inconsistently defined), agentic AI governance frameworks, and agentic platform selection criteria.

The divergence between the senior executive priorities (Zuhlke) and the practitioner priorities (Swiss Forum) is informative. Executives see validation, ROI, and governance as the pressing concerns; practitioners working directly with agentic systems are focused on legal risk, platform choices, and interoperability. Both perspectives are legitimate, and both point towards the same underlying need: coherent frameworks that allow the sector to move from experimentation to repeatable, trustworthy deployment.

Across all three sources — the audience poll, the Zuhlke survey, and the Swiss Forum — AI literacy and training is the only theme that appears consistently. It is the most broadly shared concern in the sector and the most actionable: it does not require a multi-year research programme, it requires a training programme. The Pistoia Alliance has already signalled its intent to deliver one this year.

5. Key Themes and Conclusions

A seminar that began with a question about molecular structures ended with a convergent set of observations about the state of AI in life sciences. The following themes run through the presentations and the industry intelligence alike.

#	Theme	Observation
1	**Reasoning models are a genuine step change**	The performance gap between non-reasoning and reasoning LLMs on complex scientific tasks is substantial and well-evidenced. Organisations that formed views about AI capability before 2025 should revisit them.
2	**Benchmarks must reflect real tasks — and must be dynamic**	Multiple speakers cautioned against multiple-choice questions and single-dataset evaluations. Good benchmarks are open-answer, algorithmically generated for renewability, and directly tied to the tasks they are meant to predict performance on.
3	**LLMs excel at generation; struggle with extraction**	Zero-shot LLMs underperform significantly on classification and relation extraction. The architectural reason is structural: generative models are not designed to label spans. Fine-tuned models remain the benchmark for NER and triplet tasks.
4	**LLM-as-judge is an unreliable evaluation method**	Self-assessed confidence scores are not a proxy for real performance against ground truth. Rigorous external benchmarking is essential before committing to production deployments. The 93% self-reported precision that nearly misdirected a research programme is a cautionary example.
5	**AI value requires a three-level measurement framework**	Model-level metrics, task utility, and business impact are distinct and must all be measured. Adoption rates of 60–70% routinely erode the realised value of technically capable systems. Explainability and UX are not peripheral concerns.
6	**Match solution complexity to the problem**	Not every use case justifies a full engineering build. A custom prompt may capture 80–90% of the available value at 0.3% of the cost. The decision should be made explicitly and numerically.
7	**Human oversight remains essential in regulated pharma**	There is broad consensus that fully autonomous AI workflows are premature in regulated pharmaceutical environments. Human in the loop is not a temporary concession to caution; it is an appropriate design choice for the current maturity of these systems.
8	**Skills gap is the leading barrier to scaling AI**	Confirmed by the audience poll, the Zuhlke/PA executive survey, and multiple speakers. The profile of effective AI project leads is changing software engineering capability combined with domain curiosity is outperforming domain expertise alone.
9	**AI literacy is the sector's most consistently identified need**	The only theme common to every source of evidence presented: the audience poll, the executive survey, the practitioner forum, and the speaker commentaries. PA's planned training programme is timely and should be prioritised.

#	Theme	Observation
10	**Interoperability standards are critical and undersupplied**	Two-thirds of poll respondents rated vendor-neutral interoperability standards as important or critical. The rapid evolution of the agentic platform landscape makes prescriptive tool selection premature; defining 'what good looks like' in terms of standards and requirements is the more tractable and durable contribution.

A note on benchmarking as a discipline

A thread that ran through the entire session, perhaps more explicitly than any single speaker intended, is the inadequacy of current evaluation practices across the field. Runcie showed that accepted benchmarks in chemistry have been dominated by multiple-choice questions that LLMs can game; Steensgaard Gade showed that the LLM-as-judge method can produce a 93% precision figure for a system that fails substantially under rigorous evaluation; Henstock showed that model-level accuracy metrics tell organisations almost nothing about whether their AI investments are generating business value.

In each case, the problem is the same: the evaluation method is measuring something other than what it claims to measure. Designing benchmarks that faithfully assess what matters — and that cannot be gamed, either by the models or by the organisations evaluating them — is itself a research and standardisation challenge. It is one that the Pistoia Alliance is unusually well-positioned to address, sitting as it does at the intersection of member organisations that face the problem collectively and in a pre-competitive context.

Looking ahead

The next events in the LSAIE programme aligns with the European Conference in April, followed by round tables and further seminars through the year — will continue to build the evidence base. The September session will canvas the audience on emerging topics for the November round table; the themes identified in this review — agentic AI governance, validation frameworks, and the continuing rapid evolution of reasoning model capabilities — are strong candidates for deeper treatment.

The overarching message of this first seminar is that the field is moving faster than organisational readiness, measurement frameworks, and governance structures can currently track. The value of forums like this one lies not in resolving that gap — which is structural and will take years to close — but in providing the shared vocabulary and shared evidence base that makes productive collaboration possible.

Review prepared by Rob Gill for the Pistoia Alliance Life Science AI Exchange programme, March 2026.