Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Large language models do not replace chemists in a closed-loop catalysis experiment

Andrew Ian Cooper

aicooper@liverpool.ac.uk

University of Liverpool https://orcid.org/0000-0003-0201-1021

Qi Jie Yeow

University of Liverpool

Ziheng Xiao

University of Liverpool

Mengjia Zhu

University of Liverpool https://orcid.org/0000-0002-7569-9886

Abdoulatif Cissé

University of Liverpool https://orcid.org/0009-0005-4767-5149

Xenophon Evangelopoulos

University of Liverpool

Charlotte Boott

University of Liverpool https://orcid.org/0000-0003-0902-030X

Article

Keywords:

Posted Date: July 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10032842/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Large language models do not replace chemists in a closed-loop catalysis experiment

Qi Jie Yeow,† Ziheng Xiao,† Mengjia Zhu,† Abdoulatif Cissé,† Xenophon Evangelopoulos, Charlotte E. Boott and Andrew I. Cooper*

Materials Innovation Factory and Department of Chemistry, University of Liverpool, Liverpool, L7 3NY, United Kingdom. † Equal contributions. * Email: aicooper@liverpool.ac.uk

Artificial intelligence (AI) is reshaping scientific research and laboratory automation[1]. Large language models (LLMs) can perform aspects of scientific reasoning, which could in principle reduce human decision-making as a rate-limiting step in closed-loop automated experiments[2-7]. Yet the reasoning performance of LLMs compared with human experts in complex, noisy, long-running laboratory experiments is largely unexplored. Here we benchmarked LLM reasoning head-to-head against a team of human domain experts for a noisy 25-dimensional closed-loop colloidal catalysis problem[8] explored using a mobile robot[9]. The LLM (GPT-5.1) navigated the available chemical space across a 528-experiment campaign, using background information and experimental data to propose an anionic surfactant that gave the largest single gain in catalyst activity, also adapting to a mid-campaign change in the measurement set-up. In a like-for-like final phase of 160 experiments, the human experts found a catalyst formulation that was, on average, more active than the best two formulations found in independent LLM runs. At the same time, the LLM reasoned 35 times faster and was estimated to be around 1,900 times less expensive than human reasoning. Inspection of the LLM reasoning traces found them mostly sound, but with some costly silent errors, logical inconsistencies, and apparent memory limitations, suggesting that LLMs are not out-of-the-box replacements for expert reasoning in problems of this complexity[10]. This points to a need to design closed-loop systems that combine the speed of machine reasoning with expert oversight.

Main

Laboratory automation can accelerate scientific discovery in areas such as materials research and biology[11]. Autonomous or self-driving[4,12,13] laboratories use feedback and adaptive decisions to afford a level of agency over their own actions[14]. In a sense, autonomy follows naturally from automation: if we automate experiments extensively, then human decision making can become the bottleneck. This has led to methods such as Bayesian optimisation [4,15-25] for automating experimental decision making. Recently, there has been much interest in large language models (LLMs) and agentic frameworks to guide closed-loop chemistry experiments[2,3,5-7,26-38] due to their inherent capacity to inject chemical context into high-dimensional, non-convex exploration spaces.

Previously, a benchmarking study across 2,700 question and answer pairs suggested that LLMs could outperform human chemists for such knowledge tasks[39], albeit with limitations. However, as also noted by others[40], the real-time decision-making performance of LLMs over long-horizon closed-loop chemistry laboratory experiments is sparsely explored[2,3,5,34]. To our knowledge, there are no direct side-by-side comparisons between AI reasoning and human domain expert reasoning performance for high-dimensional experimental chemistry problems[38]. One recent study reported LLM-based multi-agent systems outperforming human “best-guess” hypotheses on biomedical research goals as judged by expert preference and internal ratings,[6] but those comparisons relied on one-shot human inputs and hypothesis-quality ratings rather than experimental outcomes, leaving it unclear whether LLM reasoning is competitive with iterative expert reasoning in a closed-loop experiment setting. A multi-agent system was also used to drive an iterative lab-in-the-loop discovery cycle in biology, generating hypotheses and analysing experimental data[7], but its human comparison assessed concordance between human and AI analysis of the same data, rather than competitive reasoning performance toward an experimental outcome.

In an earlier study, we benchmarked frontier LLMs as optimisers for a 10-dimensional closed-loop chemistry experiment using a predictive reaction model, finding that LLM reasoning was effective for optimising catalyst activity while providing a native ontology.[41] The use of a reaction model precluded comparisons with human reasoning because the best reaction conditions predicted by the model were already known to us. Also, this model was deterministic whereas real-life closed-loop experiments can exhibit random noise, systematic errors, measurement drift, or periodic experimental failures, all of which present challenges for machine reasoning.

Our primary goal here was to test LLM reasoning for a noisy, high-dimensional closed-loop catalysis experiment where the optimum solution was unknown. A second goal was to compare the LLM reasoning performance with that of human domain experts. As a test problem, we chose nanoscale covalent organic frameworks (nanoCOFs) as photocatalysts for hydrogen production (Fig. 1). We showed previously that nanoCOFs can exhibit a counterintuitive reverse concentration dependence[8], where higher photocatalytic activities are observed at a lower catalyst concentration. This renders the problem more complex than many photocatalyst systems, where catalyst activity tends to increase monotonically up to a limiting catalyst concentration. We also knew that particle aggregation levels have a marked influence on catalytic performance for nanoCOFs[8], and that these colloids are prone to aggregate over time or in response to changes in their chemical environment, making them challenging systems to explore using automated experiments that run over weeks or months. To enrich the problem further, we added the hypothesis that organic-inorganic or all-organic heterojunctions[42,43] might lead to higher catalyst activities. To this end, we introduced 12 inorganic semiconductor samples as possible heterojunction partners for four different nanoscale COFs (Fig. 1a-c). We also included two pH modifiers (HCl and NaOH), three surfactants, three candidate sacrificial hole scavengers, and a platinum cocatalyst precursor (H_2PtCl_6). All told, this created a 25-dimensional problem corresponding to a combinatorial space of 5.92×10^{16} potential catalyst formulations with the constraints and discretisation levels used here.

LLM reasoning before experiments (Phase 0)

The experiments were carried out using our agentic optimisation software, Opsight (Extended Data Fig. 1). At the start of the experimental campaign, we provided the LLM with the aim of the experiment and some experimental constraints (Methods). The single objective was to maximise hydrogen production, irrespective of other factors, such as cost of the chemicals. We provided 10 resources (see data repository) that included three background research papers[8,44,45], a categorisation of pH modifiers and surfactants, a table of the bandgaps of the 16 semiconductors, the measured particle size distributions and zeta potentials of all semiconductors (as *.csv files), redox potentials for the water splitting half reactions and the sacrificial donors, and the measured UV-visible absorbance spectra of the semiconductors plus the output spectrum of the photolysis light source (*.csv files).

Before commencing any experiments, the LLM generated an initial search strategy using these 10 resources plus additional literature that it sourced using search tools as detailed in a “Research Summary” in the optimisation log (data repository). Initially, the LLM targeted two regimes: a “single-photocatalyst mode” and a “heterojunction mode”. Using the bandgap information, the model then reasoned about the reducing power of the various semiconductors and their redox compatibility with the candidate sacrificial donors, postulating, for example, that nanoCOF1 would be compatible with ascorbic acid (AA), but not with triethanolamine (TEOA), which was borne out later by experiments. The LLM then considered the overlap between the light source output spectrum and the semiconductors’ UV-visible spectra by comparing the absorbance/output spreadsheets, concluding that “nanoCOFs are the main visible-light harvesters in the system. They should be central to any high-activity formulation”. This, too, was confirmed by subsequent experiments.

The LLM then concluded that “Nanoscale COFs will interact sensitively with surfactants (CTAB, SDS, PVP), which control aggregation, interparticle distances, and heterojunction contact with oxides”, and went on, using the zeta potential data, to predict that SDS was “Likely to adsorb to the positively charged COF surfaces, possibly reducing or reversing their zeta potential and enabling heteroaggregation with other positively charged species through bridging.” The model then attempted to scale published reaction conditions[8] to determine a starting concentration for the Pt cocatalyst precursor in these experiments but it made a conversion error by not considering the molecular weight of the Pt precursor, H_2PtCl_6. As discussed below, this was a costly mistake. The model also considered heterojunction concepts informed by literature that we provided[44,45], plus the likely influence of sacrificial agents, pH, and other additives. In total, this initial reasoning phase took 4 minutes and 13 seconds, including all tool use.

LLM reasoning improves catalyst performance (Phase 1)

The experiments were performed using a mobile robot[9] (Extended Data Fig. 2) instructed by our Opsight software via a web bridge, conducted in batches of 16 experiments as a compromise between reasoning efficiency[41] and experimental throughput. We designed the web bridge as a thin client to connect the laboratory orchestration to the reasoning functions in Opsight while keeping those two systems separate (Supplementary Information Fig. S8).

The experimental campaign comprised three phases (Fig. 1d), explained in the following sections and summarised in Supplementary Figure S10. Phase 1 was an open-ended closed-loop optimisation of catalyst activity. In the first 32 experiments (two batches), the most active catalyst formulation found (experiment 18, A in Fig. 1d) was a single semiconductor system (nanoCOF2 + ascorbic acid + H_2PtCl_6), modelled on previous literature[8], but with a target Pt concentration that was around two times lower, based on the aforementioned conversion error. Attempts to form heterojunctions (*e.g.*, nanoCOF3 + WO_3) yielded lower catalytic activities. The model had already started to assess experimental noise, stating that: “The best condition (nanoCOF2 0.25 mL, AscorbicAcid_0.5M 2.0 mL, H2PtCl6 0.5 mL) has multiple replicates between ~3.0–4.3 µmol, implying σ≈0.5 µmol (CV ~15%).” The next 77 experiments failed to improve on experiment 18 despite testing a range of hypotheses relating to Pt loading, surfactants (CTAB or PVP), other candidate heterojunctions (*e.g.*, TiO_2), HCl addition, and alternative nanoCOFs, commenting that “nanoCOF4 has shown near-top performance in single runs”. The model also reassessed measurement noise, noting that “some replicates underperform strongly, suggesting occasional experimental failures or strong sensitivity to unrecorded factors.” We did not initially provide the model with any information about the available experimental budget. Coupled with an explicit system prompt to carry out duplicates because of noise (data repository), this explains the exploitative, repeat-heavy strategy in Phase 1.

After 96 experiments, the model hypothesised that: “SDS surfactant as an alternative dispersion aid. Rationale: CTAB was detrimental, but no data exist for an anionic surfactant. Low SDS doses might improve charge separation or dispersion without the deactivation seen for CTAB. We include replication at SDS 0.25 mL and 0.5 mL to disentangle noise from true effects.”

Execution of these experiments led to a large jump in catalyst activity (experiments 109–112, B in Fig. 1d) and the model concluded that:

“Across 7 batches the baseline nanoCOF2 + AA + Pt formulation without surfactant has clearly drifted downward (early mean ≈3.7 µmol vs late mean ≈1.1 µmol for identical commanded settings), confirming a high and non-stationary noise regime, likely linked to dispersion/aggregation or experimental drift. In contrast, introducing SDS at nanoCOF2 0.5 mL, AA 2.0 mL, Pt 0.5 mL produces a large and reproducible performance boost …”

Note that the hydrogen amounts in the reasoning logs are the raw measurements obtained directly by gas chromatography, neither normalised for headspace volume nor for photolysis time (*c.f.*, Fig. 1d).

After this discovery, SDS was included in 206 out of 208 experiments up to the end of Phase 1 (experiment 320; point C in Fig. 1d). Little improvement in catalyst activity was observed over these experiments, but human inspection of the gas chromatography (GC) peaks (Supplementary Information Fig. S6, S7) revealed that the detector was saturating above 60 μmol h^{-1} on the scale shown in Fig. 1d. It was therefore necessary to reduce the photolysis time to achieve a linear measurement response before proceeding further, as described in the next section (Phase 2).

The LLM was exploitative across the 320 Phase 1 experiments, sampling only 34 distinct categorical formulations, dominated by just five categorical formulations that account for 83% of experiments (Extended Data Fig. 3). Here we define a "categorical formulation" as having the same N non-zero components, irrespective of quantity, and zero values for the other (25-N) components.

To evaluate the stochasticity of the model, we regenerated 20 repeats of the first batch of experiments in Phase 1. As shown in Fig. 2a, the LLM strategy averaged across these 20 repeat runs targeted a narrow, focused area of the chemical space, even before any experiments. There was a strong focus on platinum (100% of formulations), nanoCOFs (77.5% of formulations contained one of the four nanoCOFs), and ascorbic acid (AA, 74.1%), building on the published study[8] that we provided as a resource. SDS was included in 21.9% of formulations across the 20 repeat first batches, but not in the experimental run itself, where SDS was not trialled until experiment 109 (B in Fig. 1d, also Extended Data Fig. 4b). The experimental batch in Phase 1 sits centrally in the distribution of the 20 repeats. Its centroid in the 25-dimensional space lies a mean Euclidean distance of 0.17 (range 0.08–0.24) from the centroids of the repeats, close to the mean repeat-to-repeat centroid distance of 0.18.

We also performed 20 repeat first-batch runs without the 10 appended resources to test their influence on the initial strategy (Supplementary Information section 6.3). Again, the strategy was targeted (Extended Data Fig. 5a), but with key differences from the runs including resources (Extended Data Fig. 5b). When not provided by us, the LLM found our earlier publication on nanoCOFs[8] through literature searches in only 25% of runs. This led to higher recommended nanoCOF volumes, on average, because the model did not make a connection to the reverse concentration dependence discussed in that earlier study[8]. We ascribe the omission of HCl with resources (1/320 versus 40/320 without resources) to the influence of the same study, which did not use HCl. The surfactant strategy also differed: without resources, SDS was suggested in just 8 formulations, versus 70 cases with resources. We traced this difference to a publication[46] located in 50% of the 20 runs without resources that showed that organic nanoparticles without SDS had much higher hydrogen evolution rates. Likewise, the bias toward using PVP without resources seemed to be connected to a study[47] on the positive effects of PVP, albeit for inorganic composite catalysts, that was found in 25% of the non-resource runs.

The best-performing categorical formulation at the end of Phase 1 [nanoCOF2 + AA + H_2PtCl_6 + SDS] was essentially absent from the prior distribution of the repeat first batches. Even with the attached resources, it appeared in only 2 of 320 starting points (0.6%) in a batch that articulated an explicit [nanoCOF2 + AA + H_2PtCl_6 ± SDS] hypothesis. As such, the positive effect of SDS addition (B in Fig. 1d) was found by experiment, prompted by the model's reasoning, rather than recovered from latent prior knowledge.

The evolution of the LLM hypotheses through Phase 1 and beyond is summarised in Fig. 2c, which ranks the hypotheses into 7 clustered themes. The dominant theme, labelled c5 in Fig. 2c, focused on varying the categorical [nanoCOF + Pt + AA + SDS] formulation. Inorganic-organic heterojunctions were explored briefly in Phase 1 but proved ineffective, although this theme was reinvestigated, again without success, in Phase 3 (c1 in Fig. 2c). Early attempts were made to rank nanoCOF2/3/4 in Phase 1 (c4 in Fig. 2d; batches 2–7), but unlike in human efforts, discussed below, this hypothesis was not returned to in Phase 3.

Renormalising to reduced photolysis time (Phase 2)

To bring the hydrogen peak within a linear measurement range, we reduced the illumination time from 60 to 15 minutes from experiment 321. We explained the rationale for this in an instruction prompt (Supplementary Information section 5.3.1). The LLM followed this instruction and spent the next 48 experiments (experiments 321–368) renormalising or "anchoring" the GC measurements, choosing to combine multiple repeats of the historical best formulation plus replicates of the same composition without SDS to "quantify the surfactant benefit at 15 min" also varying the Pt concentration, speculating that "under 15-min illumination and non-saturated GC conditions, the optimal Pt loading may shift slightly". The model then re-quantified noise for this shorter illumination time. This revealed that the H_2 production scaled down with shorter illumination time although not by a factor of precisely four, which caused the jump in time-normalised activity observed in Phase 2 (Fig. 1d). The LLM therefore concluded correctly that "historical 60 min data cannot simply be rescaled".

Reasoning within a fixed experimental budget (Phase 3)

Until now, we had set no explicit experimental budget for the LLM. For Phase 3, we set a termination condition of 160 further experiments (10 more batches). We also compared the performance of the LLM in Phase 3 with a human domain expert team (four of the coauthors here, see Methods). Both the human team and the LLM could draw on all experimental data from Phases 1 and 2. As discussed later, the human team also had access to other information, such as visual cues for phase behaviour, that were unavailable to the LLM. To make the comparison as fair as possible, we did not carry out any additional offline measurements during the human-led experiments.

Human domain expert reasoning (Phase 3a)

The full details of the human team hypotheses for these 160 experiments (Phase 3a) are provided in the Supplementary Information (section 9). In total, we explored 26 hypotheses over 160 experiments. The first notable advance in catalyst activity was found in experiment 375 (D in Fig. 1d) for a formulation focused on nanoCOF4, motivated by our hypothesis that the SDS benefit found by the LLM for nanoCOF2 might translate to the other three nanoCOFs, essentially revisiting an LLM hypothesis theme from Phase 1 (c4 in Fig. 2c). Following that, we also explored nanoCOF/nanoCOF blends (looking for synergic mixtures; experiments 399–416), pH modification, higher inorganic component loadings, and inorganic heterojunction partners not tested previously by the LLM, all without improvement. We also trialled the use of inorganic semiconductors without nanoCOFs (experiments 441–464), both with and without SDS addition; this area of the chemical space had been ignored by the LLM, but these experiments yielded no hydrogen. The next substantial increase in catalyst activity was found in experiments 471 and 472 (E in Fig. 1d), where we increased the Pt loading for the best [nanoCOF4 + SDS + AA + H_2PtCl_6] formulation, which was related to the Phase 1 LLM hypothesis theme c3 (Fig. 2c). Beyond that, we found only marginal improvements in activity within the experimental budget by increasing the Pt loading further (experiment 482).

LLM reasoning (Phase 3b,c)

We next prompted the LLM that the remaining experimental budget was 160 experiments (Supplementary Information section 5.3), and its task was to maximise the photocatalyst activity. We also carried out 20 repeat virtual batches at the beginning of Phase 3, again to

evaluate stochasticity (Methods), which showed that the LLM was focused consistently its initial Phase 3 search on a small subset of the 25 variables (Fig. 2b).

Initially, the LLM outperformed the human team, finding an active nanoCOF2 formulation, also with higher Pt loading (experiment 374, F in Fig. 1d; Phase 3b). The LLM improved further on this in experiment 459 (G in Fig. 1d) by using a mixture of nanoCOF2 and nanoCOF4 at high Pt loading. Up to experiment 464, the LLM had reasoned correctly about experiments conducted using 15 min irradiation time. For experiments 465–528 (Phase 3b1 in Fig. 1d), a reasoning error occurred, arising from context-window limitations and unsuccessful execution of the Python data analysis step. After three failed execution attempts, the workflow was scripted not to retry further, preventing condition-specific re-evaluation of the 60-minute and 15-minute photolysis results. Because the raw, unnormalised H_2 production was higher at 60 minutes, this caused the reasoning to revert incorrectly to the best formulation from Phase 1. We addressed this by rolling the reasoning chain back to experiment 464 and adding an instruction to focus on the data associated with 15 min photolysis time, before recommencing experiments from experiment 465 onwards (Phase 3b2, Fig. 1d). This led to a somewhat higher best-so-far catalyst activity (H in Fig. 1d), although in fact, the best formulation found was the same composition for both Phase 3b1 and the re-prompted Phase 3b2 (Table 1), the difference in activity stemming from batch-to-batch variation.

We also repeated these 160 experiments (Phase 3c) with an adapted prompt (Supplementary Information 5.3.3), again to reinforce the focus on the 15-minute photolysis time. This run performed less well than Phase 3b, but again the best formulation found by the model (Table 1) had increased Pt loadings (I and J in Fig. 1d), in this case remaining focused on nanoCOF2.

Comparison of performance and strategy for human and LLM reasoning

The three most active catalyst formulations found in Phase 3 by the human team and by the LLM are summarised in Table 1. They are all similar: the human team favoured nanoCOF4 and more elevated Pt concentrations, while the best formulations found by the LLM have lower Pt levels and favoured either nanoCOF2 (Phase 3c) or nanoCOF2/nanoCOF4 mixtures (Phase 3b). In fact, the LLM captured the key categorical formulation components in the first 120 experiments of Phase 1 (point B in Fig. 1), and improvements in Phase 3 were essentially small modifications around a platform categorical formulation, [nanoCOF + AA + H_2PtCl_6 + SDS] (Table 1). Neither the LLM nor the human team found catalytically productive organic-inorganic heterojunctions. Indeed, the gains in catalyst performance made by the human team in Phase 3a (D and E in Fig. 1d) were effectively a result of returning to earlier hypothesis themes explored by the LLM in Phase 1 (c3 and c4, Fig. 2c).

The human team and the LLM in Phase 3 followed somewhat different exploration strategies (Fig. 3a,b). The humans explored 31 distinct categorical formulations across their 160 experiments, compared with 18 each for the two independent LLM runs Phase 3b2 and Phase 3c (Fig. 3a). The human team was more explorative, also trialling more unproductive formulations (Extended Data Figs. 6 & 7), such as inorganic-only formulations. Based on the best formulation compositions found, each one averaged over 15 separate repeats (Table 1 and Supplementary Information, Fig. S15), the human team's best formulation (126 ± 8 μmol h^{-1}) had the same catalytic activity, within error, as the LLM's stronger single run, Phase 3b (116 ± 17 μmol h^{-1}, $p = 0.082$) and was 22% more active than its weaker run, Phase 3c (103 ± 14 μmol h^{-1}, $p = 3.5 \times 10^{-5}$).

We kept precise records of all team reasoning time, including both solo data analysis and team meetings to discuss strategies for the next experimental batch (Supplementary Information Table S17). The total 'clock time' for human reasoning in Phase 3a was 34.7 person hours over the 160 experiments. Analysis of the optimisation logs for Phases 3b

and 3c showed that corresponding average reasoning time for the LLM in Phase 3 was 59.2 minutes; that is, 35 times faster than the human person-hour reasoning time. This also translated to a two-fold improvement in experiment throughput because there was almost no pause between batches to consider the next experiments; the average LLM reasoning time between batches in Phase 3 was about 6 min. Based on an annual industry FTE rate of $350,000 (pharmaceutical sector) and average European Union working hours (Fig. 3c), the LLM reasoning cost in Phase 3 was around 1,900 times less expensive than the human reasoning cost. These costs simply refer to the expense of running the reasoning with LLMs and humans in Phase 3, not the overall effectiveness. For example, the LLM made reasoning errors in Phase 0 and Phase 1 (Extended Data Fig. 8). Such errors incur additional experimental costs, as do potentially unnecessary repeat experiments made by the LLM (Fig. 3d). Likewise, these experiments used a single agent, and multiagent strategies would increase costs.

Phase	H_2PtCl_6 (mL)	nanoCOF2 (mL)	nanoCOF4 (mL)	AA (mL)	SDS (mL)	Best H_2[1] (μmol h^{-1})	Repeats[2]	Mean H_2 in run[3] (μmol h^{-1})	Final mean H_2[4] (μmol h^{-1})
1	0.50	0.50	-	2.00	0.50	68	77	60 (8)	-
2	0.75	0.50	-	2.00	0.50	91	3	89 (4)	-
3a	2.00	-	0.50	2.00	0.50	145	8	134 (6)	126 (9)
3b1	1.25	0.25	0.25	2.00	0.50	128	1	128 (NA)	117 (17)
3b2	1.25	0.25	0.25	2.00	0.50	140	8	127 (11)	117 (17)
3c	1.50	0.50	-	2.00	0.50	111	4	106 (3)	103 (15)

Table 1. Composition of best catalyst formulations found by human reasoning (Phase 3a) and by LLM reasoning (Phase 3b,c). [1] Best H_2 evolution rate found in run. [2] Number of exact repeats targeted for best formulation in human and LLM runs in Phase 3. [3] Mean H_2 evolution rate for best formulation found, averaged over the number of exact repeats of that formulation in the run. [4] H_2 evolution rate averaged over 15 separate repeats for each of the three best catalyst formulations. Standard deviations in parentheses.

Rationalising the effect of SDS addition

Before any experiments, the LLM predicted that SDS was likely to adsorb to the positively charged COF surfaces, possibly reducing or reversing the zeta potential. While there was no evidence for SDS producing catalytically productive heteroaggregates with inorganic semiconductors, there was evidence that the surfactant was causing partial nanoCOF aggregation. This was evident to us, but not to the LLM, because we could observe a change in phase behaviour in the reaction vials (Fig. 4a). Offline measurements after Phase 3 showed that this was associated with a reduction in zeta potential in the presence of SDS (Fig. 4b), as predicted by the LLM in Phase 0. Additional tests showed that other anionic surfactants, such as ammonium lauryl sulphate and sodium dodecylbenzene sulfonate, gave similarly elevated H_2 production rates in the best Phase 1 formulation (Supplementary Information, Fig. S16). Photoluminescence (PL) measurements (Fig. 4c,d) indicated that SDS altered the optoelectronic properties of the nanoCOF dramatically, with markedly enhanced PL intensity in the presence of SDS. This may be an aggregate-induced emission (AIE) effect[48], where SDS-induced aggregation suppresses non-radiative decay pathways and enables radiative recombination to compete.

At the end of Phase 3, the LLM suggested that it was “likely” that SDS “improves nanoCOF2 dispersibility and prevents aggregation, increasing effective surface area and light absorption” or “modifies surface charge (consistent with zeta_potential data) to favor separation of photogenerated carriers or better interaction with ascorbate and Pt” (Phase 3b1). The first of these conclusions contradicts earlier (correct) arguments made by the LLM in Phase 0 that SDS could cause aggregation. It is also inconsistent with observed phase behaviour and our offline tests (Fig. 4), but the LLM did not have access to that information.

The model did, however, suggest in “Recommendations for Future Experiments” (Phase 3b1) that we should “correlate H_2 activity with zeta_potential and dls_size_data to quantitatively understand how aggregation state and surface charge relate to performance”, which we subsequently did (Fig. 4b-d and Supplementary Information Fig. S17, S18).

Discussion

We have shown that an LLM (GPT-5.1) can direct long-term closed-loop catalysis experiments in a complex 25-dimensional space and cope with high experimental noise, measurement drift (in Phase 1), and measurement renormalisation (in Phase 2), albeit with some reasoning failures. The re-prompting step in Phase 2 is representative of a range of process monitoring scenarios, and it illustrates the potential power of LLMs to cope dynamically with complex, real-life experimental set ups.

Benchmarked on a per experiment basis, the domain expert team outperformed the LLM in Phase 3 to give a catalyst formulation that was more active, on average, than those found by LLM-only reasoning (Fig. 1d, Table 1). However, the LLM reasoning was 35 times faster, around 1,900 times less expensive, and the overall experimental throughput using the LLM was twice as high (Fig. 3c). Hence, it is possible that the LLM would have outperformed the human team if a time or cost budget had been used, rather than a total number of experiments. These speed and cost advantages must be weighed against the LLM’s higher rate of redundant experiments and its silent reasoning errors (Fig. 3d, Extended Data Fig. 8), both of which incur experimental cost. The best division of labour between machine and human reasoning is likely to depend on the per-experiment cost regime: machine reasoning is favoured where experiments are fast, inexpensive and low-hazard, and can be run in large numbers, whereas expert reasoning, or close human oversight of machine reasoning, becomes more valuable as each experiment becomes slower, more costly, or more hazardous. Designing autonomous systems that allocate reasoning effort along both axes, rather than defaulting to full autonomy, may be more resource-efficient than either mode alone.

While our study demonstrates that LLMs can produce practical results in complex, real-life experiments, perhaps even outperforming human teams by some resource metrics, it is less clear that they always reason scientifically[10,40]. A recent study across simulations within eight domains found that LLM performance and behaviour can be explained overwhelmingly by the base model[10], and found that LLM-based agents execute scientific workflows but do not exhibit the epistemic patterns that characterise scientific reasoning. We also found that LLM and human reasoning operated quite differently in these real-life experiments (Fig. 3a,b,d) and, in some cases, the model ignored experimental evidence (Extended Data Fig. 8).

The LLM exhibited reasoning and analysis errors (Extended Data Fig. 8), plus a tendency to be highly exploitative (Fig. 3c, Extended Data Fig. 6) and to perform potentially unnecessary experiments (Fig. 3d), which impacted both its effectiveness and its efficiency. These errors were made silently and confidently, and they were caught only by human inspection. The impact of such errors may depend on the problem and the level of experimental automation. For example, exploring unnecessary points is harmful for slow, costly experiments that are performed manually, whereas fast, inexpensive, autonomous experiments could tolerate this better. The consequences also depend on the chemistry itself. Our experiments were carried out in water, on a small scale, and posed low physical hazards, so erroneous decisions cost only wasted experiments. In higher-hazard or larger-scale processes, the same confident, unflagged errors could be more serious, for example prompting an unsafe combination of reagents or conditions. Before autonomy is extended to such settings, LLM-driven systems would likely need hard-coded checks against process constraints and risk assessments, such that machine reasoning operates within bounds that are guaranteed independently of the model.

Our comparisons here rest on a single language model, a single team of human experts, and a single closed-loop catalysis optimisation problem with a fixed list of available chemicals. As such, the specific performance margins summarised in Fig. 3c should not be over-generalised. Nonetheless, we expect that the observed trade-off between the speed and cost of machine reasoning and the depth and reliability of expert reasoning will apply to other laboratory experiments. Some of the LLM's apparent weaknesses were, in part, artefacts of how we framed the task, rather than intrinsic limits. The exploitative, repeat-heavy behaviour in Phase 1 was encouraged by our instruction to replicate because of noise and by the absence of any information about the experimental budget. Some deterministic mistakes, such as the Pt stoichiometry miscalculation, could be guarded against by requiring the model to perform exact calculations through called tools, whereas the long-horizon memory failures point to a need for more robust persistent state, which remains an open problem.

We performed this study with an off-the-shelf LLM to test a tool that is available to all researchers. Stronger performance might be achieved by using other LLMs, multiagent platforms, or by using models that are fine tuned to the specific chemistry domain. It may also be possible to implement a stronger Retrieval-Augmented Generation (RAG) system.

In these experiments, the LLM had only one data source to reason about: the amount of hydrogen produced. It could not detect the GC detector was saturated, nor observe phase behaviour in the samples (Fig. 4a). In principle, the LLM could have detected the phase-behaviour changes itself—for example, by equipping the mobile robot with a camera—thus moving these platforms closer to physically embodied AI. However, the diversity of possible laboratory experiments and failure scenarios makes it challenging for AI to fully replace human sensory perception, much as the long tail of edge cases has slowed full autonomy in self-driving vehicles. In this work we did not include any computational descriptors, mostly because of the difficulty in computing useful descriptors for complex multi-length-scale colloidal assemblies (Fig. 1b), but for other problems, agentic computational modalities[32] could enrich the LLM's reasoning power further.

This study included a direct, like-for-like comparison of human and LLM reasoning for complex, noisy, real-world experiment. Such head-to-head tests are a necessary first step, since only by comparing both approaches quantitatively can we design systems that combine them effectively. Here, the two reasoning methods proved broadly complementary: the LLM reasoned quickly, cheaply, and without fatigue, recalled and applied chemical knowledge, and proposed productive hypotheses. By contrast, the human team explored the design space more widely and with less repetition and redundancy (Fig. 3d), used additional perceptual judgement, and caught silent reasoning errors that the LLM itself did not flag.

The natural conclusion is for the two reasoning modes to work together, with the humans positioned not merely to set goals and to review outputs but to contribute ideas alongside the LLM as the campaign progresses[49]. Our Opsight software is configured to allow this (Extended Data Fig. 1), but how to orchestrate such human-in-the-loop frameworks remains a question: for example, how to weight potentially conflicting hypotheses in a multiagent, hybrid AI/human team. We see this as one of the central challenges for the next generation of self-driving laboratories.

Methods

Robot specifications

The mobile robot was a LBR iiwa 14 R820 Kuka Mobile Robot mounted on a KMP 200 KUKA Mobile Platform base (Extended Data Fig. 2).[9] The movement velocity of the robot was restricted to 0.5 m s^{-1} for safety reasons. A multipurpose gripper allowed the robot to

grasp individual 10 mL gas chromatograph sample vials, solid dispensing cartridges, and the 16-position sample racks.

Experimental stations
The workflow comprised six steps, each with its own station (Extended Data Fig. 2). Solid dispensing was carried out with a Quantos QS30 instrument (Mettler Toledo). Liquid dispensing was carried out with a bespoke system that used a 200 series Mini Peristaltic Pump (Williamson) and a PCG 2500-1 scale (Kern), to dispense liquids gravimetrically using a feedback loop and a 7-needle dispensing head (Supplementary Information Fig. S4). A bespoke instrument[9] (Labman) was used to perform sample inertisation with nitrogen (to exclude oxygen) and cap crimping in one operation. Photolysis was carried out at a photolysis station that uses a vibratory motor to agitate liquids and a G2V Sunbrick^{TM} AAA LED solar simulator as the light source (Supplementary Information, Fig. S5). Gas chromatography measurements were performed with a 7890B GC and a 7697A Headspace Sampler from Agilent GC. Communication between the process management system and the stations was achieved using various communication protocols (TCP/IP over WIFI/LAN; RS-232).

Design space: The 25 chemical components were: four organic semiconductors (nanoCOFs1-4), 12 inorganic semiconductors (V_2O_5, TiO_2 rutile-1, TiO_2rutile-2, TiO_2 anatase-1, TiO_2 anatase-2, Fe_2O_3-1, Fe_2O_3-2, SnO_2-1, SnO_2-2, WO_3-1, WO_3-2 and ZnSe), three sacrificial agents (ascorbic acid, L-cysteine and triethanolamine), 3 surfactants (CTAB, PVP and SDS), a base (NaOH), and acid (HCl), and a co-catalyst precursor (H_2PtCl_6) (Fig. 1a). Where there was more than one sample of an inorganic (*e.g.*, WO_3-1 / WO_3-2) they differed in average particle size (Supplementary Information Table S1).

Materials and synthetic procedures
NanoCOFs1-4 were prepared according to a modified literature procedure (section 1.2 of Supplementary Information)[8]. Precursors for the nanoCOFs and other materials were purchased from commercial suppliers and used as received, with exception of tin(IV) oxide (≥99.99% trace metals basis), which was ground before use in a pestle and mortar to increase flowability.

Datasheets: All formulation components were characterised before the experiments, and this information was provided to the LLM as resource files in csv format (data repository). Semiconductors were characterised by dynamic light scattering (particle size), zeta potential (surface charge) and UV-vis spectroscopy (absorption profile and band gap). Surfactants, NaOH and HCl were classified based on their charge (anionic, cationic or neutral), sodium content, and acidic or basic nature. The emission spectrum of the light source (Supplementary Information Fig. S5) was also provided, again as a csv file.

Experiment goal: Maximise absolute amount of hydrogen produced (μmol) after a fixed irradiation time (60 min in Phase 1, 15 min in Phases 2 and 3).

Experimental constraints: (1) The sum of all liquid parameters should not exceed 5 mL. (2) The sum of all solid components should not exceed 10 mg. (3) There must be at least one light absorbing semiconductor in the formulation. (4) The target mass of individual solid components in the formulation should not be 0.2 mg because this is too close to the dispensing accuracy. (5) The number of non-zero components in the formulation should not exceed ten; this was introduced to prevent redundant complexity and to prevent the experimental throughput from becoming excessively slow. (6) Acid and base should not be added simultaneously.

LLM-driven closed-loop optimisation
The exploration and optimisation of the parameter space was performed using our in-house chemical discovery web tool, Opsight (https://opsightlab.com/, Extended Data Fig. 1)[41]. Opsight supports hypothesis-driven scientific optimisation by enabling an intuitive human-optimiser interaction that allows the user to set up optimisation parameters, and to define the problem context and any initial hypotheses. The software also provides comprehensive graphical and text explanations of the ongoing discovery progress. Opsight offers multiple optimiser choices ranging from Bayesian optimisation, human-in-the-loop BO,[50] LLM reasoning[41], and LLM/BO hybrids[51]. Parametrically, the catalysis problem studied here was characterised by high, uncorrelated dimensionality and, in Phase 1, experimental measurement drift, which is challenging for BO methods. Significant noise was observed in these experiments because of the intricate and sensitive nature of the formulation chemistry, and Gaussian observation noise modelling, as often used in BO tends not to capture this well[52]. We used GPT-5.1 (medium reasoning level) for these experiments as a representative frontier model at the time that the experiments commenced (February 2026). The LLM-driven optimisation was handled through a sequence of prompts via API calls. The LLM-driven optimisation was orchestrated through a staged prompt stack delivered via API calls. We refer to the campaign-level context prompt as an “experiment card” (Supplementary Information section 5.1), which includes a detailed description of the chemistry task at hand, the optimisation objective, any parameter space constraints and, optionally, additional resources such as materials characterisation data and related literature (data repository). The LLM was free to choose its own optimisation strategy in each run, which encourages replicate-based robustness when needed while also guiding broader optimisation decisions, but the Opsight software was scripted to produce certain standard sub-sections, including literature search, initial scientific hypotheses, step-by-step optimisation overviews, and final key observations, summaries, and recommendations for future experiments. Further architectural details for Opsight are included in the Supplementary Information (Fig. S8, S9).

LLM tools
Opsight was equipped with an agentic framework capable of autonomous tool calling, ranging from web searches for literature scanning, Python code interpreters for data analysis and resource retrieval, and autonomous memory lookup operations. The agentic framework was also equipped with RAG-based memory integration (https://github.com/langchain-ai/langgraph) that enhances the LLM's ability to query past hypotheses and experimental outcomes.

Lab-optimiser communication
To allow Opsight to guide the experiments in the laboratory, we built a bridging software layer (Extended Data Fig. 1) that translates Opsight's formulation decisions into actionable laboratory protocols enabling a fully autonomous self-driving discovery setup, outside of loading up the initial chemicals and consumables. This software allows the automation workflow management system to retrieve pending experimental batches from Opsight and return the experimental results. This software bridge can handle sample format translation, authentication, and workflow synchronisation, while Opsight oversees the optimisation state, optimiser model updates, and batch recommendation logic (Supplementary Information Fig. S8).

Comparing human and LLM performance in Phase 3
The human team in Phase 3 comprised a subset of the coauthors here (QJY, ZX, CB, AIC). Collectively, this team had broad experience in sacrificial photochemical hydrogen production[9], COFs (including nanoCOFs)[8], colloids, and materials chemistry. Between batches, team members analysed the data either individually or together, but the final strategy for each next batch was always agreed by discussion as a team. We used a variety of tools to analyse the data including the plotting functions within Opsight and other offline data visualisation packages, such as TIBCO Spotfire, but we did not use LLMs for additional analyses beyond the summaries that were provided natively within Opsight. We kept strict records of all time spent on data analysis and team discussions and in total the team spent 34.7 person hours on data analysis over the course of Phase 3 (Supplementary Information section 9.1). LLM reasoning times were extracted from the Opsight logs. The stochasticity tests shown in Fig. 2a,b and Extended Data Fig. 5 were done at the end of Phase 3: that is, after all human and LLM experiments had completed. The LLM reasoning cost for Phase 3 was determined empirically by tracking individual expenditures for each of the three independent runs (Phase 3b1, 3b2, and 3c). The total costs for these three runs were then summed and averaged to provide the final reported cost per Phase 3 campaign. We note that these costs depend strongly on the tools used: for example, image analyses this would increase the average cost, as would the use of multiple agents (a single agent was used here).

Data availability
All data are included in the following repository: https://anonymous.4open.science/r/llm-vs-chemist-closed-loop-catalysis-experiments/ The repository includes resource files, raw optimisation logs from Opsight, experimental data (Phases 1–3), batch 1 and batch 24 stochasticity tests (20 repeats each), ablation studies (with and without attached resources), system prompts, instruction prompts, and a video of the robotic workflow in operation.

Acknowledgements
The authors acknowledge funding from the Engineering and Physical Sciences Research Council (EPSRC) (EP/V026887/1), the Advanced Research and Invention Agency (ARIA) (AI scientists programme), the China Scholarship Council (CSC), the European Research Council under the European Union's Horizon 2020 research and innovation program (grant agreement no. 856405), the EPSRC AIchemy hub (EP/Y028775/1 and EP/Y028759/1), and the Leverhulme Trust *via* the Leverhulme Research Centre for Functional Materials Design. A.I.C. thanks the Royal Society for a Research Professorship (RSRP\S2\232003).

Author contributions
Q.J.Y. and Z.X. performed the catalysis experiments. Q.J.Y. maintained the autonomous workflow and upgraded the liquid dispensing and photolysis stations. Z.X. synthesised and characterised the nanoCOFs and performed the offline analysis. Q.J.Y., Z.X., C.E.B and A.I.C. provided the hypotheses for the human-led experiments in Phase 3a. A.C. and M.Z. jointly implemented the Opsight software and performed all data analyses, with A.C. leading the Opsight implementation and M.Z. leading the data analysis. A.I.C. conceived and led the project with support from X.E. and C.E.B. A.I.C. wrote the paper with contributions from all authors.

Competing interests
The authors declare no competing interests.

References

1 Lu, C. S. *et al.* Towards end-to-end automation of AI research. *Nature* **651** (2026). https://doi.org/10.1038/s41586-026-10265-5
2 Boiko, D., Macknight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. *Nature* **624**, 570–578 (2023). https://doi.org/10.1038/s41586-023-06792-0
3 Bran, A. *et al.* Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6**, 13 (2024). https://doi.org/10.1038/s42256-024-00832-8
4 Tom, G. *et al.* Self-driving laboratories for chemistry and materials science. *Chem. Rev.* **124**, 9633–9732 (2024). https://doi.org/10.1021/acs.chemrev.4c00055
5 Song, T. *et al.* A multiagent-driven robotic AI chemist enabling autonomous chemical research on demand. *J. Am. Chem. Soc.* **147**, 12534–12545 (2025). https://doi.org/10.1021/jacs.4c17738
6 Gottweis, J. *et al.* Accelerating scientific discovery with Co-Scientist. *Nature* (2026). https://doi.org/10.1038/s41586-026-10644-y
7 Ghareeb, A. E. *et al.* A multi-agent system for automating scientific discovery. *Nature* (2026). https://doi.org/10.1038/s41586-026-10652-y
8 Zhao, W. *et al.* Nanoscale covalent organic frameworks for enhanced photocatalytic hydrogen production. *Nat. Commun.* **15**, 6482 (2024). https://doi.org/10.1038/s41467-024-50839-3
9 Burger, B. *et al.* A mobile robotic chemist. *Nature* **583**, 237–241 (2020). https://doi.org/10.1038/s41586-020-2442-2
10 Ríos-García, M. *et al.* AI scientists produce results without reasoning scientifically. *arXiv e-prints*, arXiv:2604.18805 (2026). https://doi.org/10.48550/arXiv.2604.18805
11 King, R. *et al.* Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* **427**, 247–252 (2004). https://doi.org/10.1038/nature02236
12 Nikolaev, P. *et al.* Autonomy in materials research: a case study in carbon nanotube growth. *npj Comput. Mater.* **2**, 6 (2016). https://doi.org/10.1038/npjcompumats.2016.31
13 Abolhasani, M. & Kumacheva, E. The rise of self-driving labs in chemical and materials sciences. *Nat. Synth.* **2**, 483–492 (2023). https://doi.org/10.1038/s44160-022-00231-0
14 Canty, R., Koscher, B., Mcdonald, M. & Jensen, K. Integrating autonomy into automated research platforms. *Digit. Discov.* **2**, 1259–1268 (2023). https://doi.org/10.1039/d3dd00135k
15 Häse, F., Roch, L., Kreisbeck, C. & Aspuru-Guzik, A. Phoenics: A Bayesian optimizer for chemistry. *ACS Central Sci.* **4**, 1134–1145 (2018). https://doi.org/10.1021/acscentsci.8b00307
16 Shields, B. J. *et al.* Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590**, 89–96 (2021). https://doi.org/10.1038/s41586-021-03213-y

17 Zhu, Q. *et al.* An all-round AI-Chemist with a scientific mind. *National Science Review* **9** (2022). https://doi.org/10.1093/nsr/nwac190
18 Angello, N. *et al.* Closed-loop optimization of general reaction conditions for heteroaryl Suzuki-Miyaura coupling. *Science* **378**, 399–405 (2022). https://doi.org/10.1126/science.adc8743
19 Siemenn, A., Ren, Z., Li, Q. & Buonassisi, T. Fast Bayesian optimization of needle-in-a-haystack problems using zooming memory-based initialization (ZoMBI). *npj Comput. Mater.* **9**, 13 (2023). https://doi.org/10.1038/s41524-023-01048-x
20 Yoo, H. J. *et al.* Bespoke metal nanoparticle synthesis at room temperature and discovery of chemical knowledge on nanoparticle growth via autonomous experimentations. *Adv. Funct. Mater.* **34**, 2312561 (2024). https://doi.org/10.1002/adfm.202312561
21 Rankovic, B., Griffiths, R., Moss, H. & Schwaller, P. Bayesian optimisation for additive screening and yield improvements – beyond one-hot encoding. *Digit. Discov.* **3** (2024). https://doi.org/10.1039/d3dd00096f
22 Slattery, A. *et al.* Automated self-optimization, intensification, and scale-up of photocatalysis in flow. *Science* **383** (2024). https://doi.org/10.1126/science.adj1817
23 Wang, C. S. *et al.* Autonomous platform for solution processing of electronic polymers. *Nat. Commun.* **16** (2025). https://doi.org/10.1038/s41467-024-55655-3
24 Xu, J. G. *et al.* Autonomous multi-robot synthesis and optimization of metal halide perovskite nanocrystals. *Nat. Commun.* **16** (2025). https://doi.org/10.1038/s41467-025-63209-4
25 Gao, D. P. *et al.* Autonomous closed-loop framework for reproducible perovskite solar cells. *Nature* **653**, 707–714 (2026). https://doi.org/10.1038/s41586-026-10482-y
26 Zheng, Z., Zhang, O., Borgs, C., Chayes, J. & Yaghi, O. ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. *J. Am. Chem. Soc.* **145**, 18048–18062 (2023). https://doi.org/10.1021/jacs.3c05819
27 Zheng, Z. L. *et al.* ChatGPT Research Group for optimizing the crystallinity of MOFs and COFs. *ACS Central Sci.* **9**, 2161–2170 (2023). https://doi.org/10.1021/acscentsci.3c01087
28 Ruan, Y. X. *et al.* An automatic end-to-end chemical synthesis development platform powered by large language models. *Nat. Commun.* **15** (2024). https://doi.org/10.1038/s41467-024-54457-x
29 Macknight, R. *et al.* Rethinking chemical research in the age of large language models. *Nat. Comput. Sci.*, **5**, 715–726 (2025). https://doi.org/10.1038/s43588-025-00811-y
30 Zhang, Y. *et al.* Large language models to accelerate organic chemistry synthesis. *Nat. Mach. Intell.* **7**, 1010–1022 (2025). https://doi.org/10.1038/s42256-025-01066-y
31 Huang, L. *et al.* Natural-language-interfaced robotic synthesis for AI-copilot-assisted exploration of inorganic materials. *J. Am. Chem. Soc.* **147**, 23014–23025 (2025). https://doi.org/10.1021/jacs.5c05916
32 Zou, Y. H. *et al.* El Agente: An autonomous agent for quantum chemistry. *Matter* **8** (2025). https://doi.org/10.1016/j.matt.2025.102263
33 Mitchener, L. *et al.* Kosmos: An AI scientist for autonomous discovery. arXiv:2511.02824 (2025). https://doi.org/10.48550/arXiv.2511.02824
34 Fei, Y. *et al.* Agentic LLM reasoning in a self-driving laboratory for air-sensitive lithium halide spinel conductors. arXiv:2604.11957 (2026). https://doi.org/10.48550/arXiv.2604.11957
35 Ramos, M. C., Michtavy, S. S., White, A. D. & Porosoff, M. D. Bayesian optimization of catalysis with in-context learning. *ACS Central Sci.* **12**, 599–615 (2026). https://doi.org/10.1021/acscentsci.5c02418
36 Tamura, R. *et al.* LLM-guided phase diagram construction through high-throughput experimentation. arXiv:2604.20304 (2026). https://doi.org/10.48550/arXiv.2604.20304

37 Wang, K. Y. *et al.* Synthesis of highly crystalline covalent organic frameworks using large language models. *J. Am. Chem. Soc.* **148**, 10094–10104 (2026). https://doi.org/10.1021/jacs.5c23233

38 Han, D. *et al.* ChemBOMAS: Accelerated BO in chemistry with LLM-enhanced multi-agent system. arXiv:2509.08736 (2025). https://doi.org/10.48550/arXiv.2509.08736

39 Mirza, A. *et al.* A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nat. Chem.* **17** (2025). https://doi.org/10.1038/s41557-025-01815-x

40 Bisht, H., Kumar, V., Maik Jablonka, K., Mausam & Anoop Krishnan, N. M. Agentic AI scientists are not built for autonomous scientific discovery. arXiv:2605.08956 (2026). https://doi.org/10.48550/arXiv.2605.08956

41 Cissé, A., Cooper, M. E., Zhu, M., Evangelopoulos, X. & Cooper, A. I. Can we automate scientific reasoning in closed-loop experiments using large language models? *Digit. Discov.* **5**, 1132–1160 (2026). https://doi.org/10.1039/D5DD00520E

42 Moon, H. S. *et al.* Spatial separation of cocatalysts on Z-scheme organic/inorganic heterostructure hollow spheres for enhanced photocatalytic H_2 evolution and in-depth analysis of the charge-transfer mechanism. *Adv. Mater.* **35** (2023). https://doi.org/10.1002/adma.202200172

43 He, Y. *et al.* Boosting artificial photosynthesis: CO_2 chemisorption and S-scheme charge separation via anchoring inorganic QDs on COFs. *ACS Catal.* **14**, 1951–1961 (2024). https://doi.org/10.1021/acscatal.4c00026

44 Xu, Q. L., Zhang, L. Y., Cheng, B., Fan, J. J. & Yu, J. G. S-scheme heterojunction photocatalyst. *CHEM* **6**, 1543–1559 (2020). https://doi.org/10.1016/j.chempr.2020.06.010

45 Sun, S. C. *et al.* Engineering interfacial band bending over bismuth vanadate/carbon nitride by work function regulation for efficient solar-driven water splitting. *Sci. Bull.* **67**, 389–397 (2022). https://doi.org/10.1016/j.scib.2021.10.009

46 Bruder, J. *et al.* Photocatalytic hydrogen generation in surfactant-free, aqueous organic nanoparticle dispersions. *Small* **20** (2024). https://doi.org/10.1002/smll.202406236

47 Rao, V. N. *et al.* Surfactant controlled metal oxide shell layer deposition for enhanced photocatalytic solar hydrogen generation: CdSe/TiO_2 nanocomposite a case study. *Mater. Lett.* **298** (2021). https://doi.org/10.1016/j.matlet.2021.130025

48 Barbieri, M. *et al.* Supramolecular dye polymers for aggregation-induced photocatalysis. *Nat. Chem.* (2026). https://doi.org/10.1038/s41557-026-02151-4

49 Soni, A. *et al.* Closed-loop, human-AI–driven optimization of geological hydrogen production. *ChemRxiv* **2026** https://doi.org/10.26434/chemrxiv.15002058/v1

50 Cissé, A., Evangelopoulos, X., Carruthers, S., Gusev, V. V. & Cooper, A. I. HypBO: Accelerating black-box scientific experiments using experts' hypotheses. *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*, 3881–3889 (2024). https://www.ijcai.org/proceedings/2024/0429.pdf

51 Cissé, A., Evangelopoulos, X., Gusev, V. V. & Cooper, A. I. Language-based Bayesian optimization research assistant (BORA). *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)*, 4967–4975 (2025). https://www.ijcai.org/proceedings/2025/0553.pdf

52 Tautvaisas, S. & Zilinskas, J. Heteroscedastic Bayesian optimization using generalized product of experts. *J. Glob. Optim.* **91**, 393–413 (2025). https://doi.org/10.1007/s10898-023-01333-5

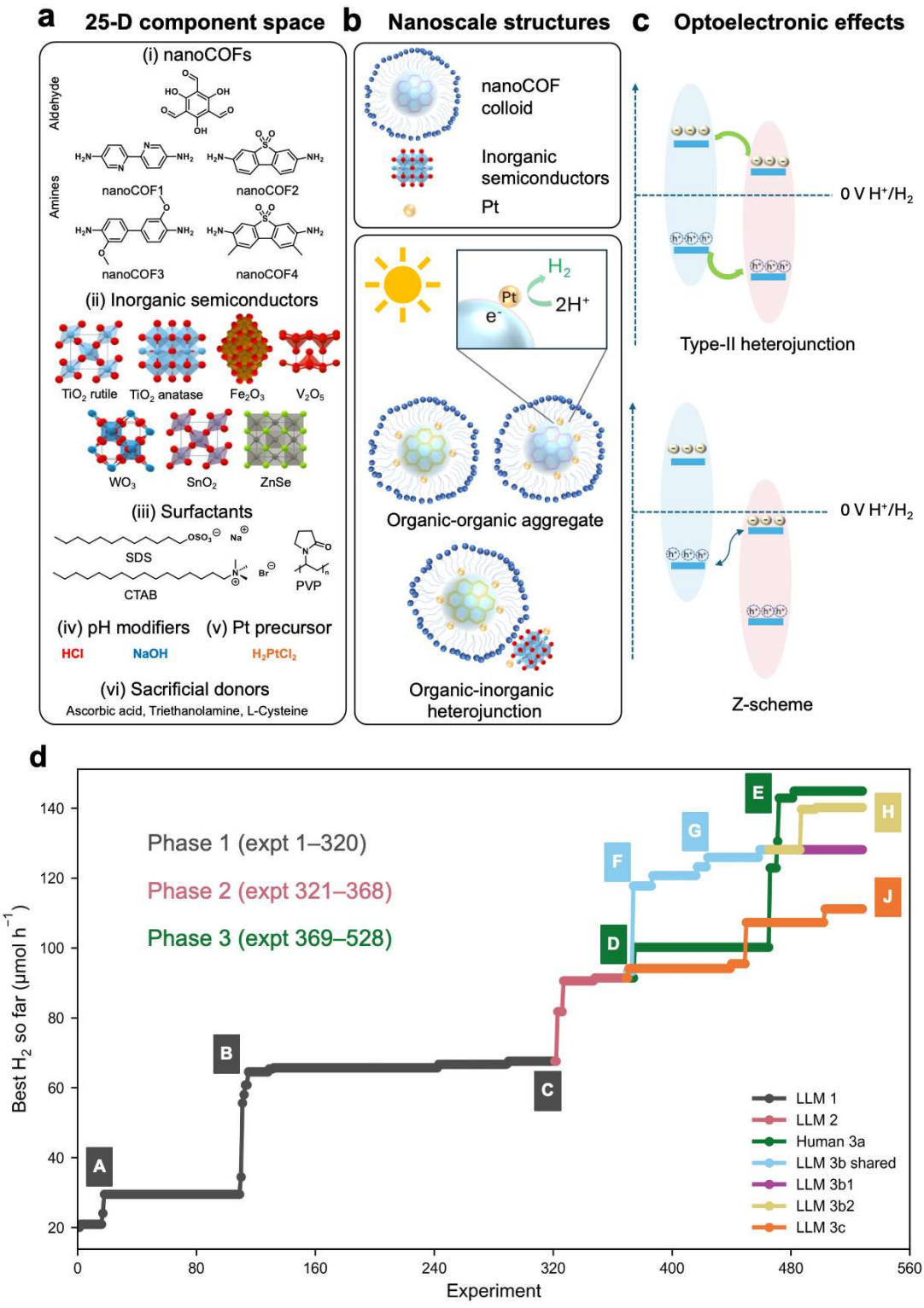


Fig. 1. | Optimisation of catalyst activity in a 25-dimensional photocatalyst formulation space using LLM reasoning and human reasoning across 528 experiments. a, The 25 catalyst formulation components across six categories, (i)–(vi). Some inorganic semiconductors were included twice with different particle sizes. **b**, These 25 components could form nanoscale structures including aggregates and **c**, heterojunctions. **d**, Plot of best H_2 value found so far against experiment number across three experimental phases: Phase 1 (LLM reasoning, 320 experiments); Phase 2 (LLM reasoning); Phase 3 (LLM or human reasoning, 160 experiments, branching the campaign after experiment 320). Phase 3a = 4-person human team; Phase 3b1, 3b2 and 3c = LLM.

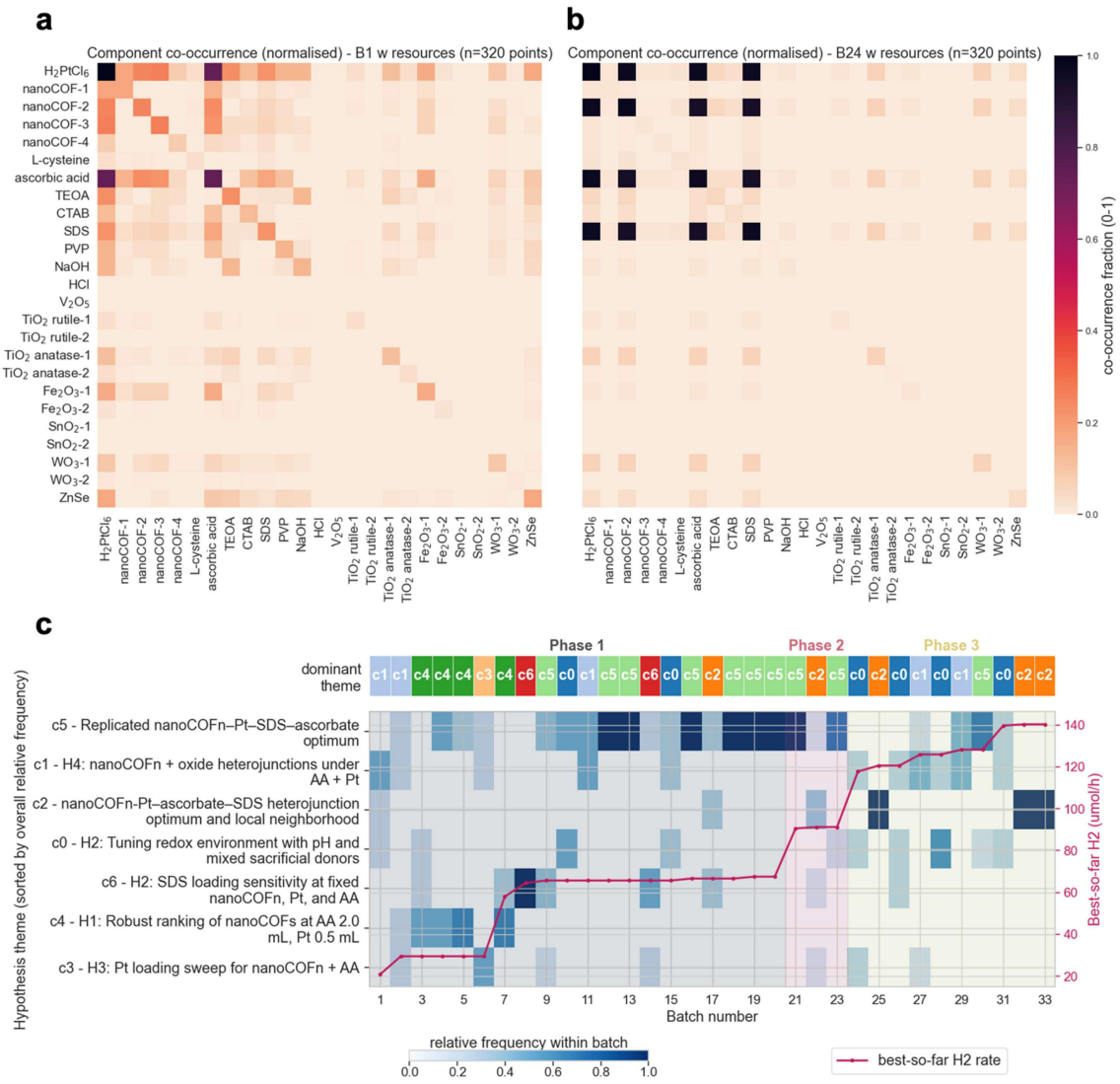


Fig. 2. | Analysis of LLM strategies at the start of Phase 1 and Phase 3 (20 repeats of each) and thematic evolution of LLM hypotheses over Phases 1, 2 and 3 for the full experimental campaign. **a**, Co-occurrence of the 25 formulation components over 20 repeat initialisations for batch 1 (experiments 1–16) in Phase 1 with inclusion of the 10 data resources (see also Extended Data Fig. 5). **b**, Same plot for first batch in Phase 3b2, again based on 20 repeats. Co-occurrence is defined as both components having non-zero values in the formulation (off-diagonal points); the on-diagonal points show the relative occurrence of the 25 components across the 20 repeats. **c**, Evolution of hypotheses over the experimental campaign (Phase 3b2 variant) with hypotheses grouped into seven themes of related hypotheses. The dominant hypothesis theme by batch throughout the three phases of the campaign is shown at the top of the figure. Some LLM hypothesis themes, such as c4, were trialled in Phase 1 and not returned to by the model, but they were revisited by the human team (Phase 3a), which led to improved catalyst activities.

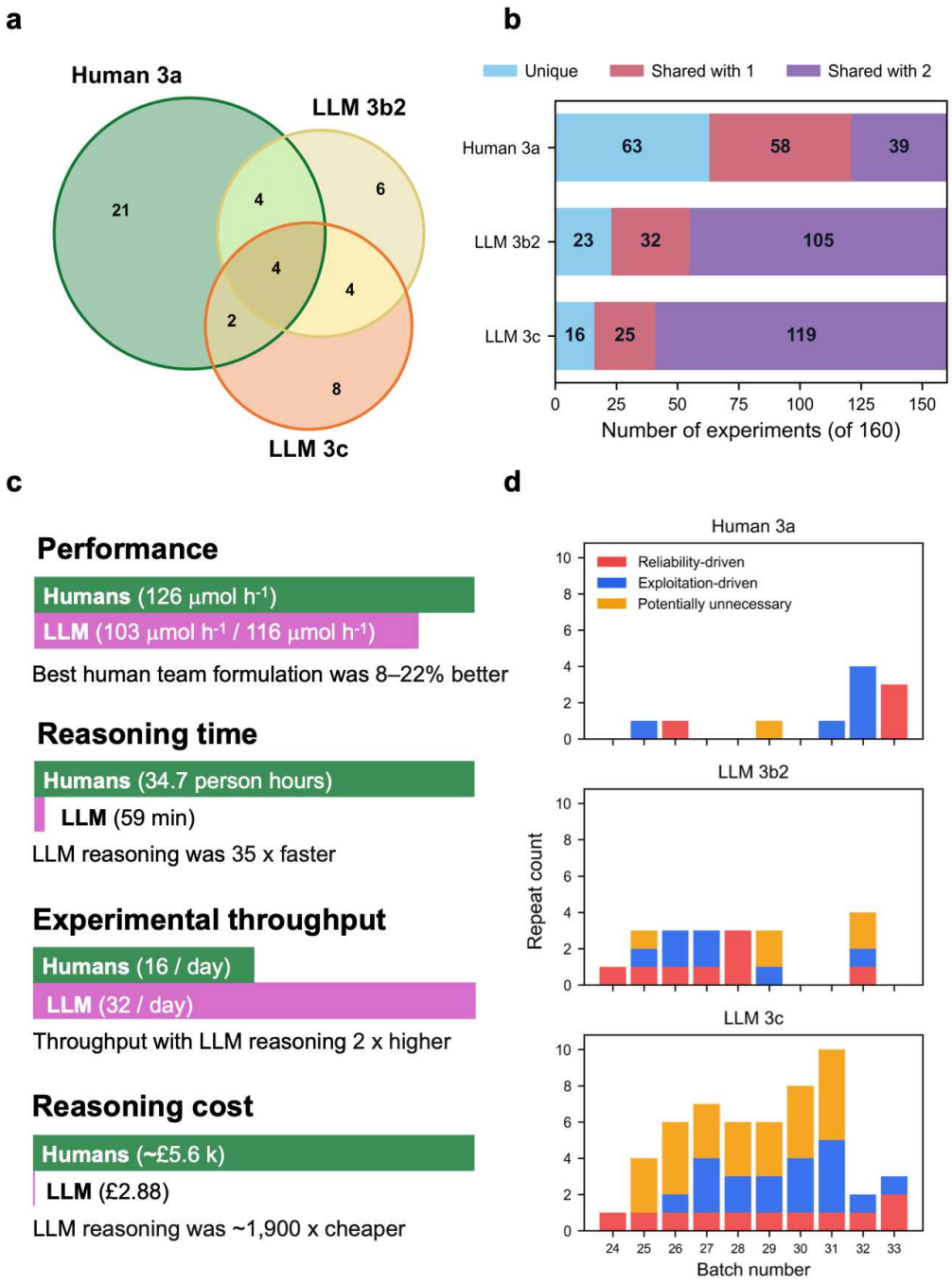


Fig. 3. | The human team explored more unique formulations in Phase 3 with less repetition and redundancy to find a superior catalyst formulation, but LLM reasoning was much faster and cheaper. a, Overlap of unique categorical formulations suggested by the human team and the LLM, ignoring repeats. Circles proportional to number. In total, the human team explored 31 unique categorical formulations while the LLM explored 18 (in both Phase 3b2 and 3c). **b**, Repeat aware comparison between human and LLM search strategies in Phase 3. **c**, Performance comparisons for best formulation from human Phase 3a and the average of the best two formulations found by the LLM in Phase 3b1/2 and 3c (Table 1). Reasoning cost assumes an annual loaded FTE rate for a pharmaceutical researcher of $350,000 USD, and the 2025 European Union average of 45 working weeks per year (36 h week). **d**, Compared to the human team, the LLM performed more identical repeats in successive batches and more potentially unnecessary repeats. Plots show exact repeat formulations per batch in Phase 3, relative to the immediate previous batch. Potentially unnecessary repeats are repeats of reproducible/stable but low- to mid-performing formulations; full definitions of these classifications are given in the Supplementary Information (Fig. S12).

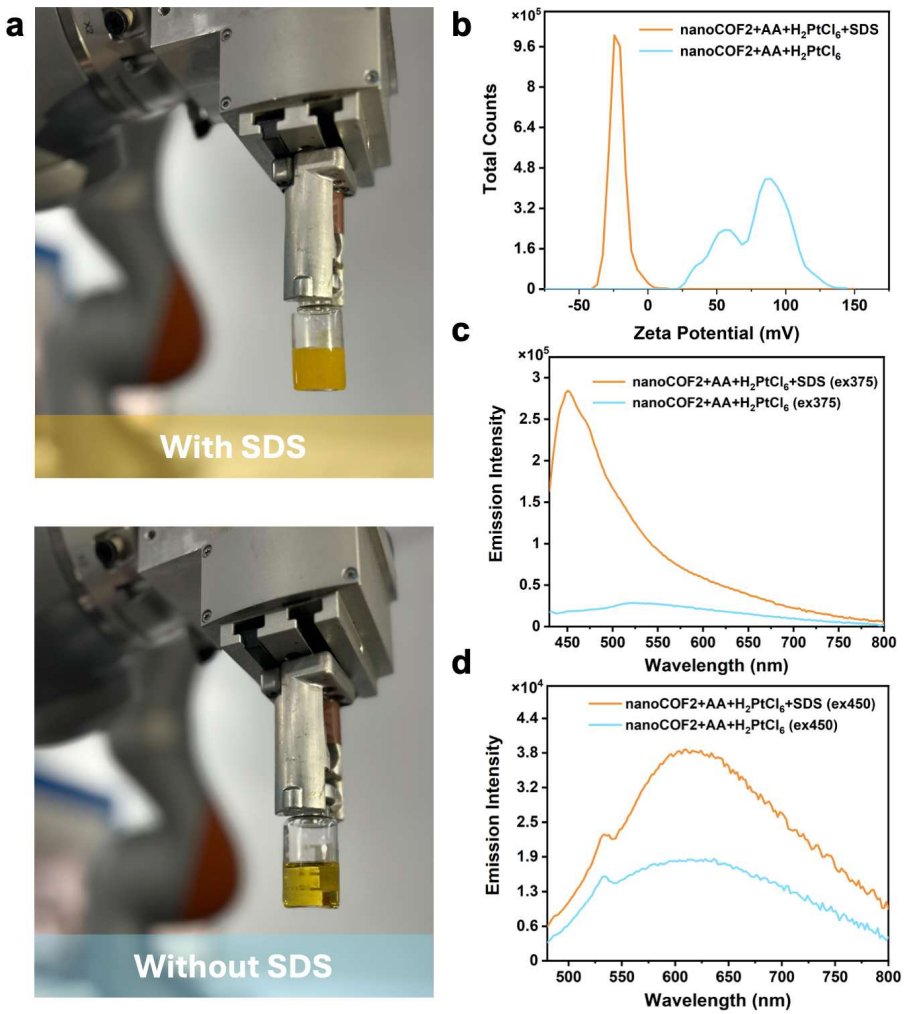
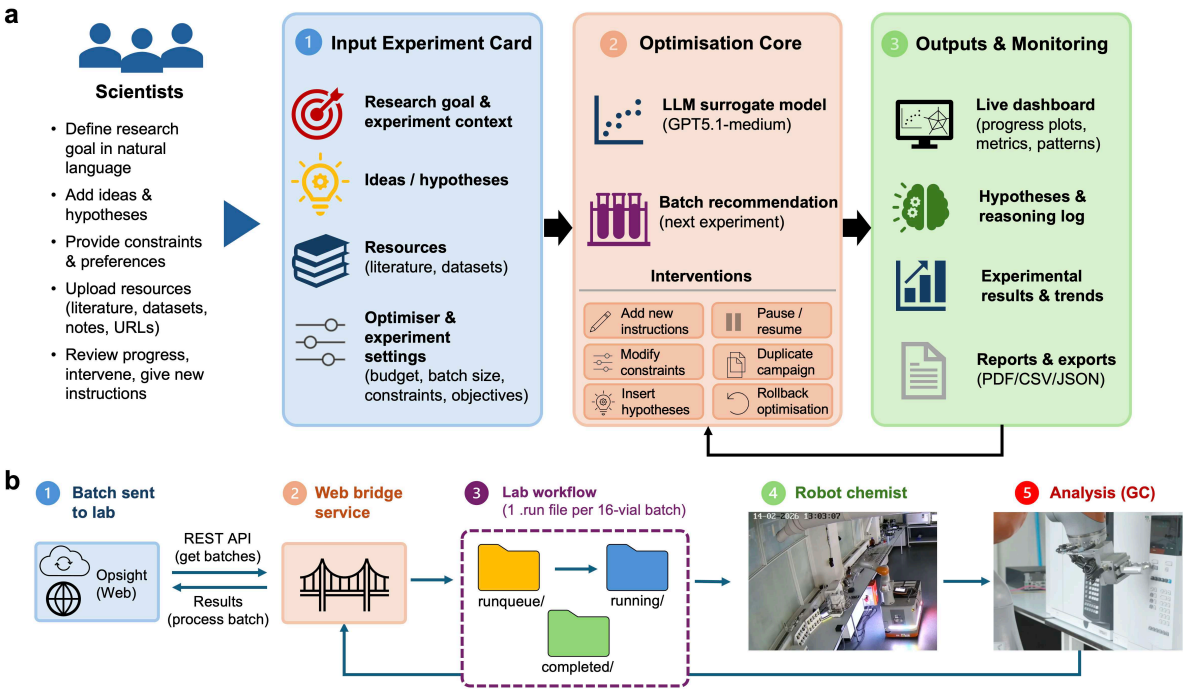
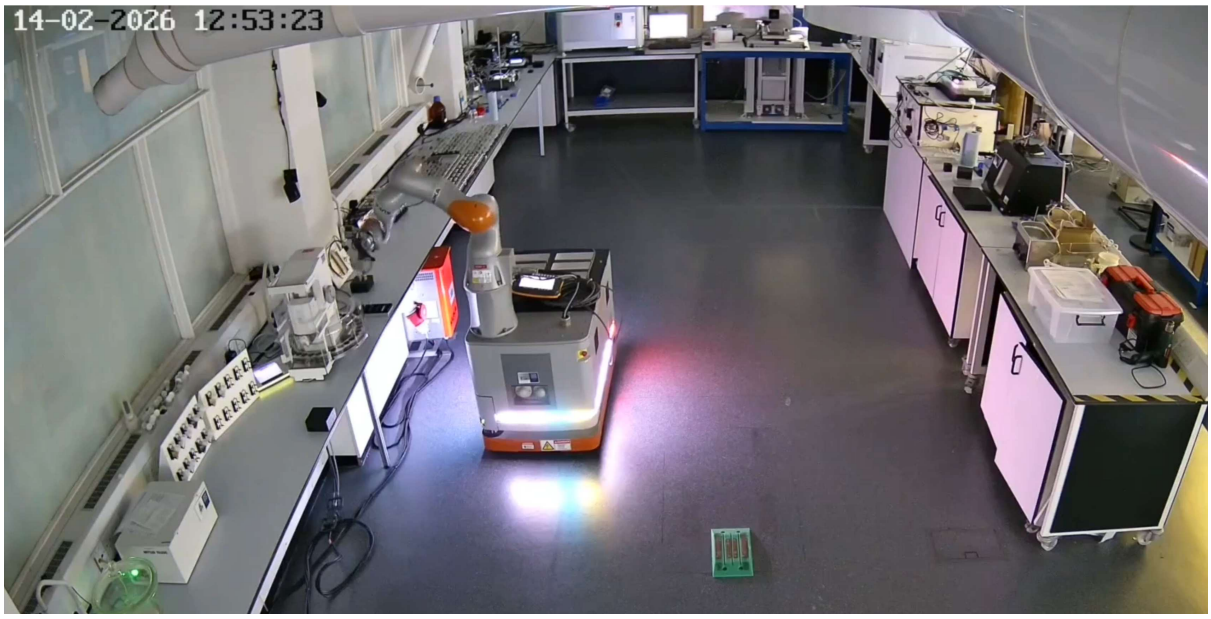


Fig. 4. | SDS addition partially aggregated nanoCOF2, reducing zeta potential and increasing photoluminescence, leading to higher photocatalytic activity. a, Photographs of samples nanoCOF2 / ascorbic acid (AA) / H_2PtCl_6 formulations with and without SDS addition; the sample with SDS is optically opaque, indicating partial aggregation of the colloids. **b**, Zeta potential plots with and without SDS addition. **c**, **d**, Photoluminescence spectra with and without SDS addition at an excitation wavelength of 375 nm (**c**) and 450 nm (**d**). For particle size analyses with and without SDS, see Supplementary Information Fig. S17, S18.



Extended Data Fig 1. | Opsight workflow for LLM-guided closed-loop experimental optimisation. **a**, Online platform workflow. Scientists can start an optimisation campaign through an input experiment card that captures the research objective, experimental context, available resources, and initial hypotheses. This information is passed to the Opsight LLM-scientist platform, where the LLM analyses the evolving optimisation data and resources and recommends the next batch of experiments based on its reasoning capacity. During optimisation, scientists can remain in the loop by adding new instructions or new hypotheses by pausing and resuming the campaign (*e.g.*, Phase 1 to Phase 2 / Phase 3 transitions here). Live features of duplicating a campaign state or rolling back to an earlier checkpoint allow scientists to explore alternative discovery strategies from a shared experimental history. Outputs from the optimisation, including live dashboards, reasoning logs, experimental trends, and exportable reports, are returned to the platform and fed back into the optimisation core to support iterative decision-making. **b**, Autonomous closed-loop experiment execution. Opsight communicates with the self-driving laboratory through a lightweight Web Bridge (2) that translates platform-level batch recommendations into laboratory-specific run files (3) and returns measured results to Opsight. New batches are retrieved from Opsight through the REST API, written to the laboratory workflow queue, executed by the automated laboratory system (4), and analysed using gas chromatography (5). Completed results are then parsed and submitted back to Opsight, where they update the optimisation state and trigger recommendation of the next experimental batch. This architecture decouples optimisation (1), laboratory execution (4 & 5, and data synchronization (2), enabling autonomous experimentation while preserving human oversight, traceability, and reproducibility.

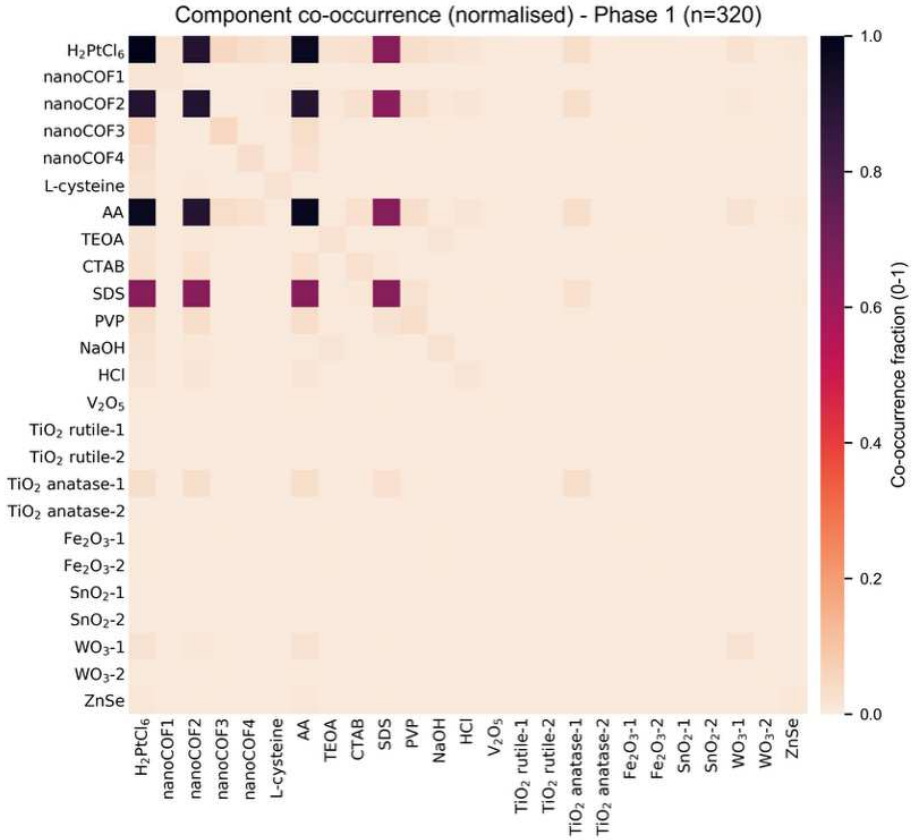


Extended Data Fig. 2 | Closed-loop ‘make-measure-reason’ cycles using a mobile robot coupled with an LLM reasoning tool. In this CCTV image, the robot was commencing a solid dispensing task (Phase 1, batch 10). Across the whole study (Phase 1–3, all LLM and human reasoning runs), the robot carried out 912 experiments and performed 4,526 liquid dispensing tasks and 143 solid dispensing tasks with low error rates in most cases (Supplementary Information Table S24). The laboratory was operated in consistently low-light conditions during the experiments, as shown, to avoid variations in the photocatalytic activity that result from changes in ambient light.

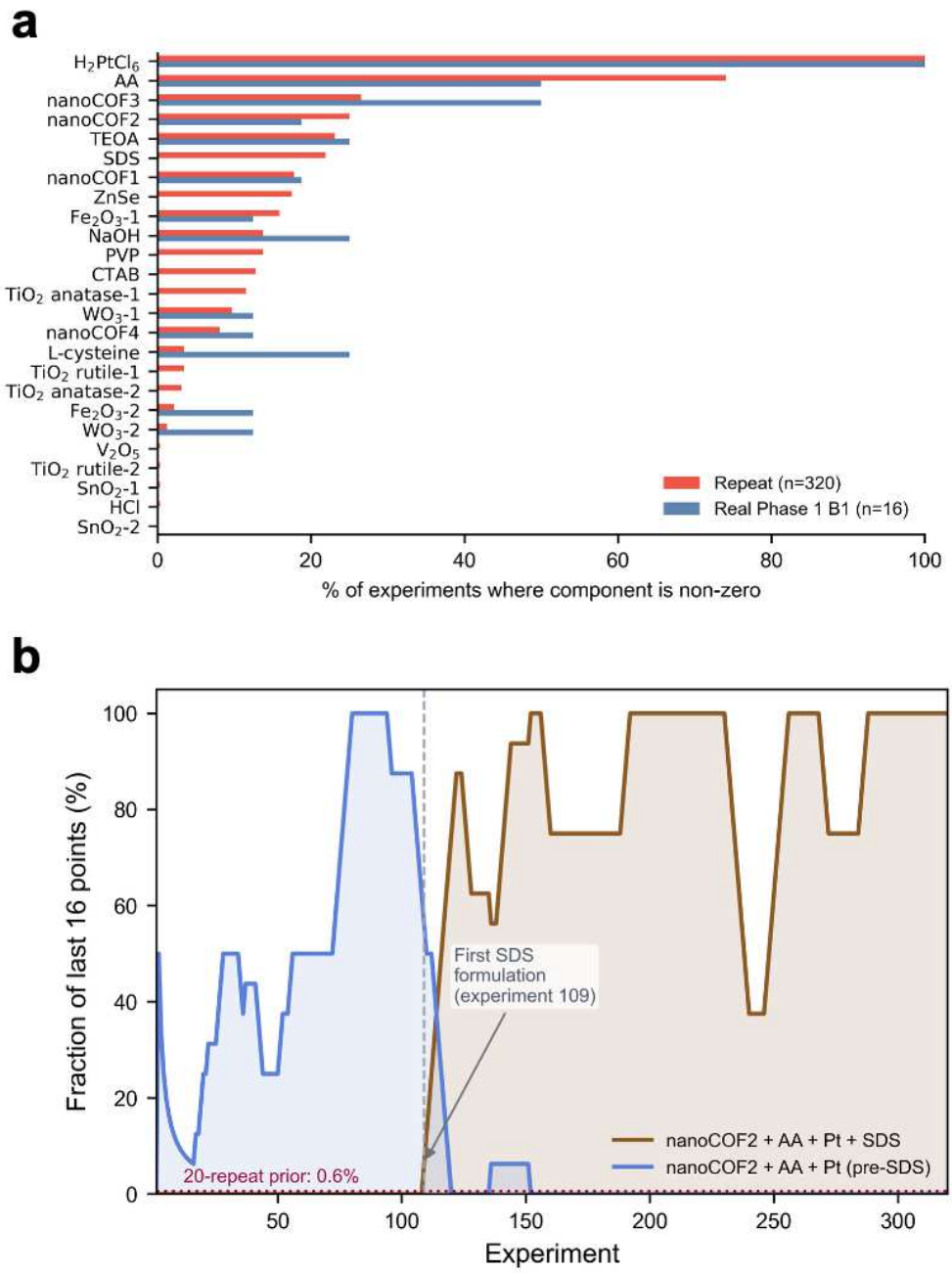
a

Rank	Count (pct)	Categorical Formulation
1	183 (57%)	nanoCOF2 + AA + H_2PtCl_6 + SDS (*the categorical optimum*)
2	60 (19%)	nanoCOF2 + AA + H_2PtCl_6 (*pre-SDS baseline*)
3	8 (2.5%)	nanoCOF3 + AA + H_2PtCl_6
4	8 (2.5%)	nanoCOF4 + AA + H_2PtCl_6
5	6 (1.9%)	nanoCOF2 + AA + H_2PtCl_6 + SDS + TiO_2 anatase-1 (*brief heterojunction probe*)

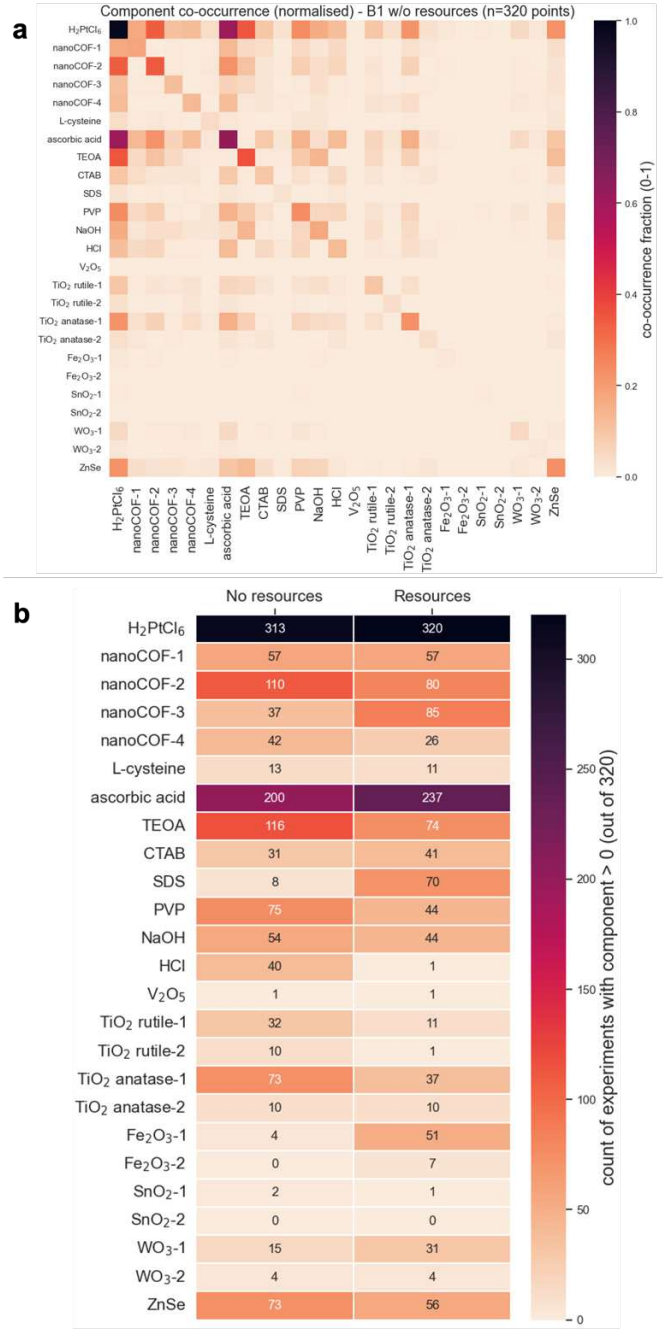
b



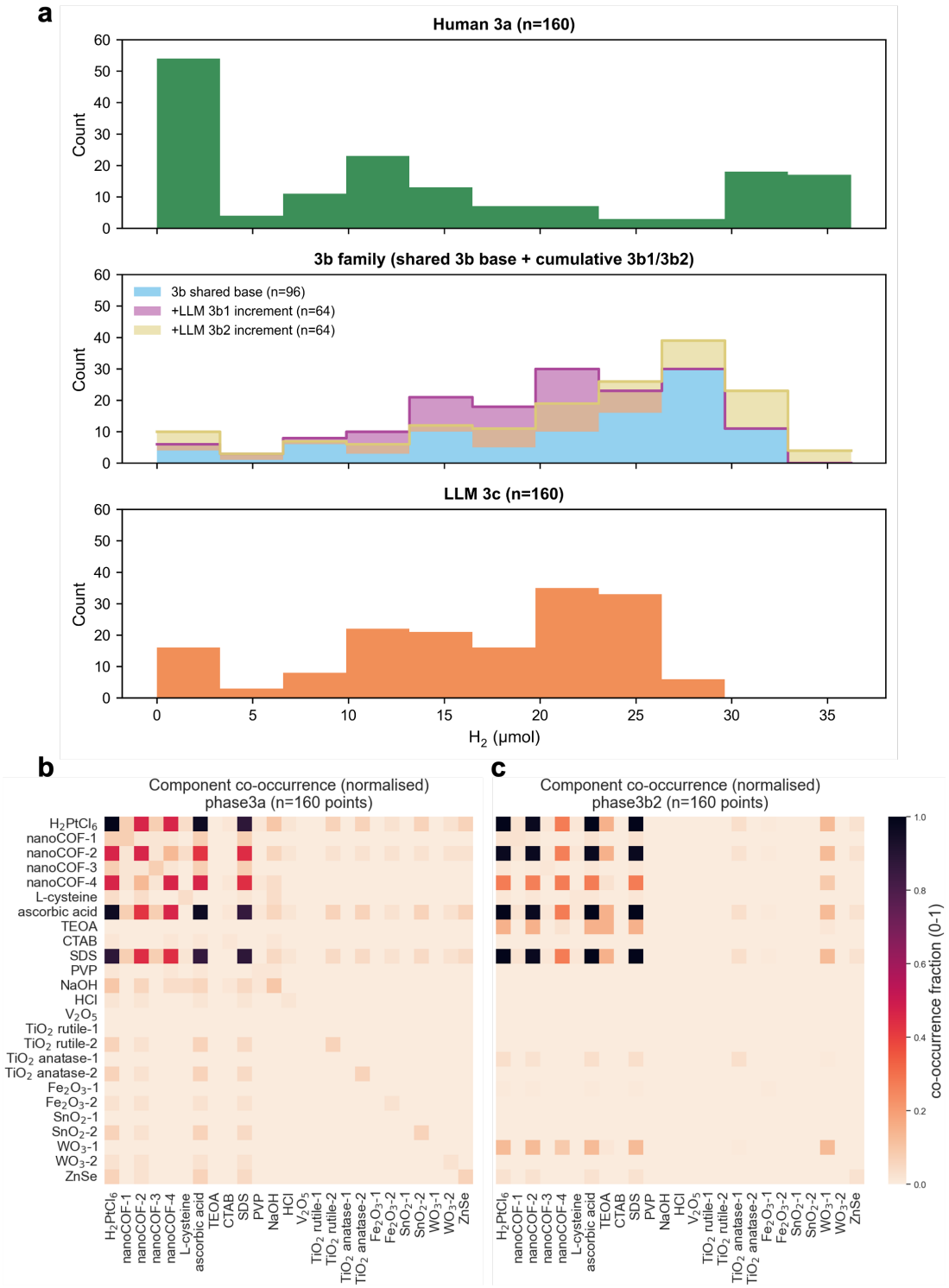
Extended Data Fig. 3 | Top five most sampled categorical formulations in Phase 1 show that the LLM was exploitative and repeat-heavy in Phase 1. a, **b**, Only 34 unique categorical formulations were explored across 320 experiments, and these top 5 categorical formulations account for 83% of experiments. Co-occurrence in **b** is defined as both components having non-zero values in the formulation (off-diagonal points); the on-diagonal points show the relative occurrence of the 25 components across the 20 repeats.



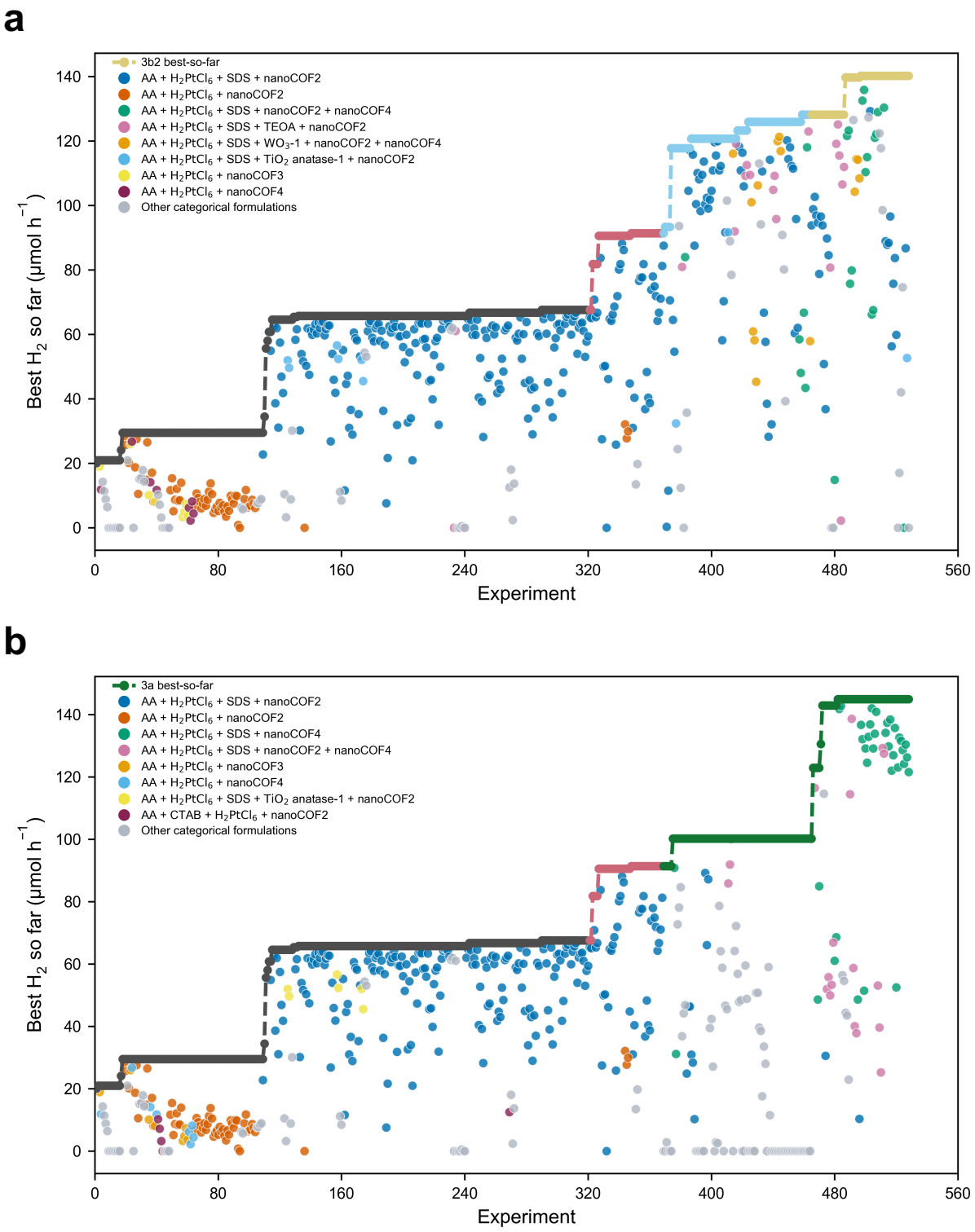
Extended Data Fig. 4 | The best categorical formulation at the end of Phase 1 was discovered by iterative experiment, not recovered from the latent prior knowledge in the LLM. **a**, Per-component usage frequency, defined as the percentage of experiments in which each component was non-zero, compared for the 20 repeat batches with resources (n = 320 total points) and the first experimental batch in the experimental Phase 1 (n = 16). Components are ordered by their frequency in the restart dataset. SDS was not explored in the first experiment batch. **b**, Rolling proportion (16-experiment window) of the final best categorical formulation from Phase 1 (nanoCOF2 + ascorbic acid (AA) + H_2PtCl_6 + SDS) and its pre-SDS three-component analogue (nanoCOF2 + AA + H_2PtCl_6) across the 320-experiment Phase 1 sequence. The horizontal dotted line marks the occurence of the final best categorical formulation in the 20-repeat prior (0.6%; that is, 2/320 suggested formulations). The first SDS-containing categorical formulation, [nanoCOF2 + AA + Pt + SDS], appeared at experiment 109.



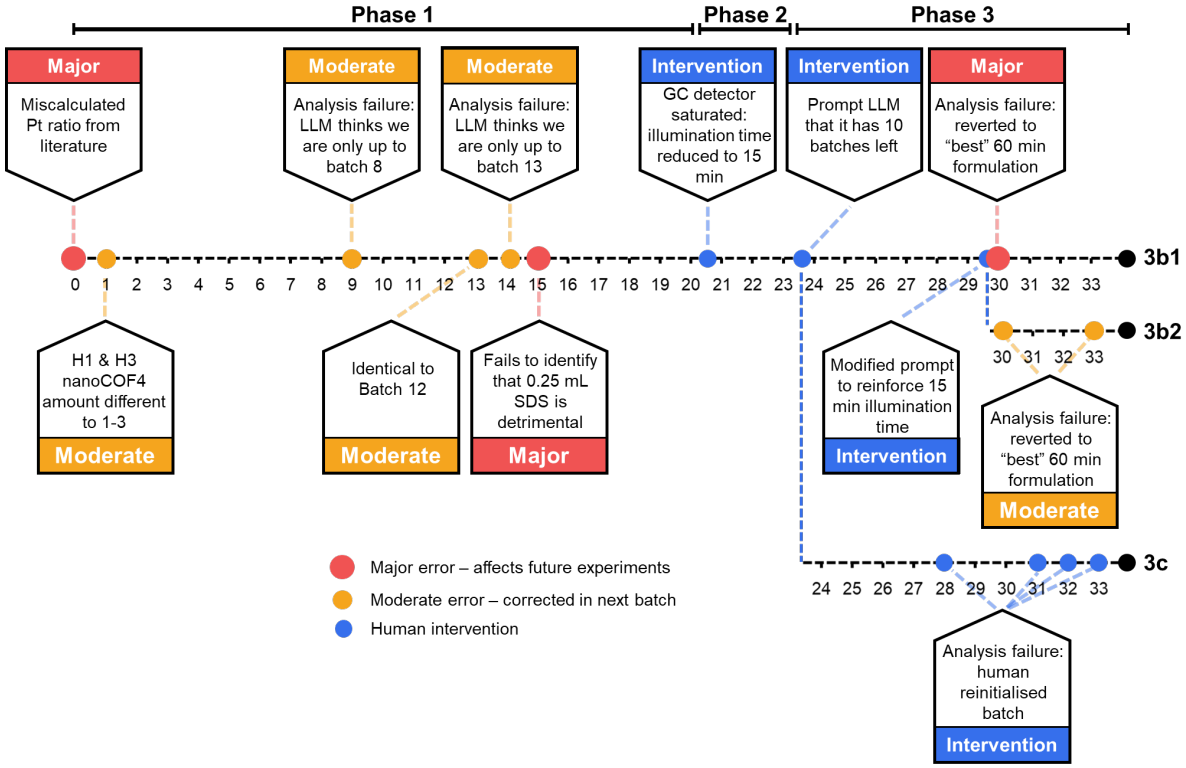
Extended Data Fig. 5 | Effect of omitting resources on strategy in first batch (Phase 1). **a**, Co-occurrence of the 25 formulation components over 20 repeat runs for the initial 16 experiments in Phase 1 (batch 1) without attaching any resources (*i.e.*, references, bandgap data, zeta potentials, etc.). Equivalent plot with resources is shown in Fig. 2a. Co-occurrence is defined as both components having non-zero values in the formulation (off-diagonal points); the on-diagonal points show the relative occurrence of the 25 components across the 20 repeats. **b**, Comparison of the occurrence of the 25 components in the 20 repeat runs (320 suggested formulations) with and without resources. Individual cells show the number of formulations where that component was non-zero. Beyond differences in non-zero preferences for individual components, there were also differences in the average amounts of individual the components suggested. On average, the volume of nanoCOF solution suggested was higher without resources. For example, in run 20 (no resources), the model reasoned that: “even filling the vial entirely with a single nanoCOF suspension gives ≤ ~0.45 mg COF. This is far below the 3–10 mg COF mass commonly used in batch COF HER experiment.” This latter figure (3–10 mg) referred to the more dominant literature on bulk COFs, rather than nanoCOFs, because the LLM located the nanoCOF literature much less often without the relevant publication resource (25% vs 100%).



Extended Data Fig. 6 | Human reasoning explored more low-performing formulations in Phase 3 while still finding higher performing formulations than the LLM. a, Plots show the count of formulations in each performance range (raw H_2 production). **b**, Co-occurrence of the 25 formulation components over Phase 3 for the human team (Phase 3a). **c**, Same plot for the LLM in Phase 3 (Phase 3b2). Co-occurrence is defined as both components having non-zero values in the formulation (off-diagonal points); the on-diagonal points show the relative occurrence of the 25 components across the 20 repeats. Human reasoning explored more points outside of the most heavily explored formulations in Phase 3, as indicated by the greater number of non-zero points outside of the top left quadrant in **b**.



Extended Data Fig. 7 | Optimisation of catalyst activity in a 25-dimensional photocatalyst formulation space over 528 experiments using LLM reasoning and LLM/human hybrid reasoning. a, LLM reasoning in Phase 3 (Phase 3b2). **b**, Human domain expert reasoning in Phase 3 (Phase 3a). Each point is an experiment, colour-coded by categorical formulation (Phase 1 and Phase 2 are common to both plots).



Extended Data Fig. 8 | Early LLM reasoning errors impacted optimisation performance. A timeline showing reasoning errors (red/yellow) and human interventions (blue). One of the most impactful errors occurred before any experiments (Phase 0), where a miscalculation of Pt precursor stoichiometry led to initial Pt concentrations that were two times lower than would have otherwise been extrapolated from previous literature (ref. 8). In batch 1, the LLM proposed to directly compare the 4 nanoCOFs but the nanoCOF4 experimental conditions (Pt and AA concentrations) were different to the nanoCOF1-3 experiments, thus disadvantaging nanoCOF4 (moderate error). At batch 15, the LLM stated that perturbations within ±0.25 mL SDS showed a relatively flat optimum plateau, but 0.25 mL SDS consistently reduced performance by almost 50% compared to 0.5 mL SDS; it was not until batch 19 that the LLM recognized that lower SDS concentrations were strongly detrimental (major error). The analysis failure in batch 30 is discussed in the main text. Interventions in Phase 3c (blue points) refer to reinitialising the batch if an analysis failure occurred. The GC saturation intervention was not an LLM oversight: we did not provide the model with any information about the linear measurement range or GC peak shapes. Likewise, the intervention at batch 24 (10-batch termination condition) was a team decision, not a correction to the LLMs actions. Hence, all human interventions, for example to correct memory or analysis errors, occurred in Phase 3 of the campaign.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf