Supplementary Information

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

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.

Contents

1 Materials and methods

1.1 Materials

2,4,6-Trihydroxybenzene-1,3,5-tricarbaldehyde (95%), 3,7-diaminodibenzo[b,d]thiophene 5,5-dioxide (97%), and [2,2'-bipyridine]-5,5'-diamine (97%) were obtained from BLD pharm Ltd. 3,7-diamino-2,8-dimethyldibenzo-thiophene sulfone (70%; contains 2'6-dimethyl isomer) was purchased from TCI chemicals. Methyl sulfoxide (99.7+%) was purchased from Thermo Scientific Chemicals. o-Dianisidine, tin (IV) oxide (nanopowder, ≤100 nm avg. part. size), zinc selenide (powder, 10 µm, 99.99% trace metals basis), tin(IV) oxide (≥99.99% trace metals basis), tungsten(VI) oxide (nanopowder, <100 nm particle size (TEM)), tungsten(VI) oxide (powder, ≤25 µm, ≥99% trace metals basis), vanadium(V) oxide (99.9% trace metals basis), iron(III) oxide (powder, <5 µm, ≥96%), iron(III) oxide (nanopowder, <50 nm particle size (BET)), titanium(IV) oxide, rutile (powder, <5 µm, ≥99.9% trace metals basis), titanium(IV) oxide, rutile (nanopowder, <100 nm particle size, 99.5% trace metals basis), titanium(IV) oxide, anatase (powder, 99.8% trace metals basis), titanium(IV) oxide, anatase (nanopowder, <25 nm particle size, 99.7% trace metals basis), chloroplatinic acid solution (8 wt. % in H_2O, liquid), hexadecyltrimethylammonium bromide (≥98%), sodium dodecyl sulfate (ACS reagent, ≥99.0%), acetic acid (glacial, ACS reagent, ≥99.7%), polyvinylpyrrolidone (powder, average mol wt 10,000), sodium dodecylbenzenesulfonate (technical grade), ammonium lauryl sulfate solution (~30% in H_2O), L-ascorbic acid (99%), L-cysteine (≥98.5% (RT), suitable for HPLC, BioUltra) and triethanolamine (reagent grade, 98%) were all purchased from Sigma-Aldrich Ltd. Materials were used as purchased, with exception of tin(IV) oxide (≥99.99% trace metals basis), which was ground using a pestle and mortar to increase flowability.

Material	Component Name
Iron(III) oxide (powder, <5 µm, ≥96%)	Fe_2O_3 sample 1
Iron(III) oxide (nanopowder, <50 nm particle size (BET))	Fe_2O_3 sample 2
Tin(IV) oxide (≥99.99% trace metals basis)	SnO_2 sample 1
Tin(IV) oxide (nanopowder, ≤100 nm avg. part. Size)	SnO_2 sample 2
Titanium(IV) oxide, anatase (powder, 99.8% trace metals basis)	TiO_2 anatase sample 1
Titanium(IV) oxide, anatase (nanopowder, <25 nm particle size, 99.7% trace metals basis)	TiO_2 anatase sample 2
Titanium(IV) oxide, rutile (powder, <5 µm, ≥99.9% trace metals basis)	TiO_2 rutile sample 1
Titanium(IV) oxide, rutile (nanopowder, <100 nm particle size, 99.5% trace metals basis)	TiO_2 rutile sample 2
Tungsten(VI) oxide (powder, ≤25 µm, ≥99% trace metals basis)	WO_3 sample 1
Tungsten(VI) oxide (nanopowder, <100 nm particle size (TEM))	WO_3 sample 2
Zinc selenide (powder, 10 µm, 99.99% trace metals basis)	ZnSe
Vanadium(V) oxide (99.9% trace metals basis)	V_2O_5

Table S1. Component names for the 12 inorganic semiconductors.

1.2 Synthesis of nanoCOF1–4

2,4,6-Trihydroxybenzene-1,3,5-tricarbaldehyde (TFP) (0.1 mmol) was dissolved in dimethylsulfoxide (DMSO) (1 mL). This TFP solution was added dropwise to a flask containing hexadecyltrimethylammoniumbromide (CTAB) solution (58 mL, 0.05 M in water). After ultrasonication, a sodium dodecyl sulfate (SDS) solution (1.8 mL, 0.05 M in water) was added to form solution A. Separately, the diamine monomer (**nanoCOF1**: [2,2'-bipyridine]-5,5'-diamine, **nanoCOF2**: 3,7-diaminodibenzo[b,d]thiophene 5,5-dioxide, **nanoCOF3**: 3,3'-dimethoxy-[1,1'-biphenyl]-4,4'-diamine (o-dianisidine), **nanoCOF4**: 3,7-diamino-2,8-dimethyldibenzo[b,d]thiophene 5,5-dioxide; 0.15 mmol in all cases) was dissolved in DMSO (0.5 mL). This solution was added dropwise to a flask containing CTAB (58 mL, 0.05 M in water). After ultrasonication, SDS solution (1.8 mL, 0.05 M in water) was added to form solution B. Finally, solutions A and B were mixed by a syringe pump with an addition speed of 1.5 mL min^{-1}, and acetic acid (5.8 mL) was added to the resultant solution using the same syringe pump an addition speed of 1.5 mL min^{-1}. The resulting solution was left to stir at room temperature for 72 h. The ideal concentrations of the synthesised nanoCOF suspensions, assuming complete conversion, are listed in the

Table S2. The nanoCOF colloids were diluted with water (20% v/v) before being used in the automated workflow.

Sample	Concentration (mg/mL)
nanoCOF1	0.343
nanoCOF2	0.414
nanoCOF3	0.412
nanoCOF4	0.447

Table S2. Ideal concentrations of nanoCOF colloidal solutions, assuming complete monomer conversion.

1.3 Batch information for nanoCOFs

Two independent batches of nanoCOF colloids were synthesised and used throughout the experimental campaign (Table S3). To ensure batch-to-batch consistency and material quality, UV–Vis spectroscopy (Figure S1) and dynamic light scattering (DLS) analyses (Figure S2) were performed as quality control measures. The results demonstrated comparable optical properties and hydrodynamic size distributions between the two batches.

nanoCOF batch	Experimental Phases
1	Phase 1
2	Phase 2 – Phase 3c

Table S3. Use of the nanoCOF batches in the experimental campaign.

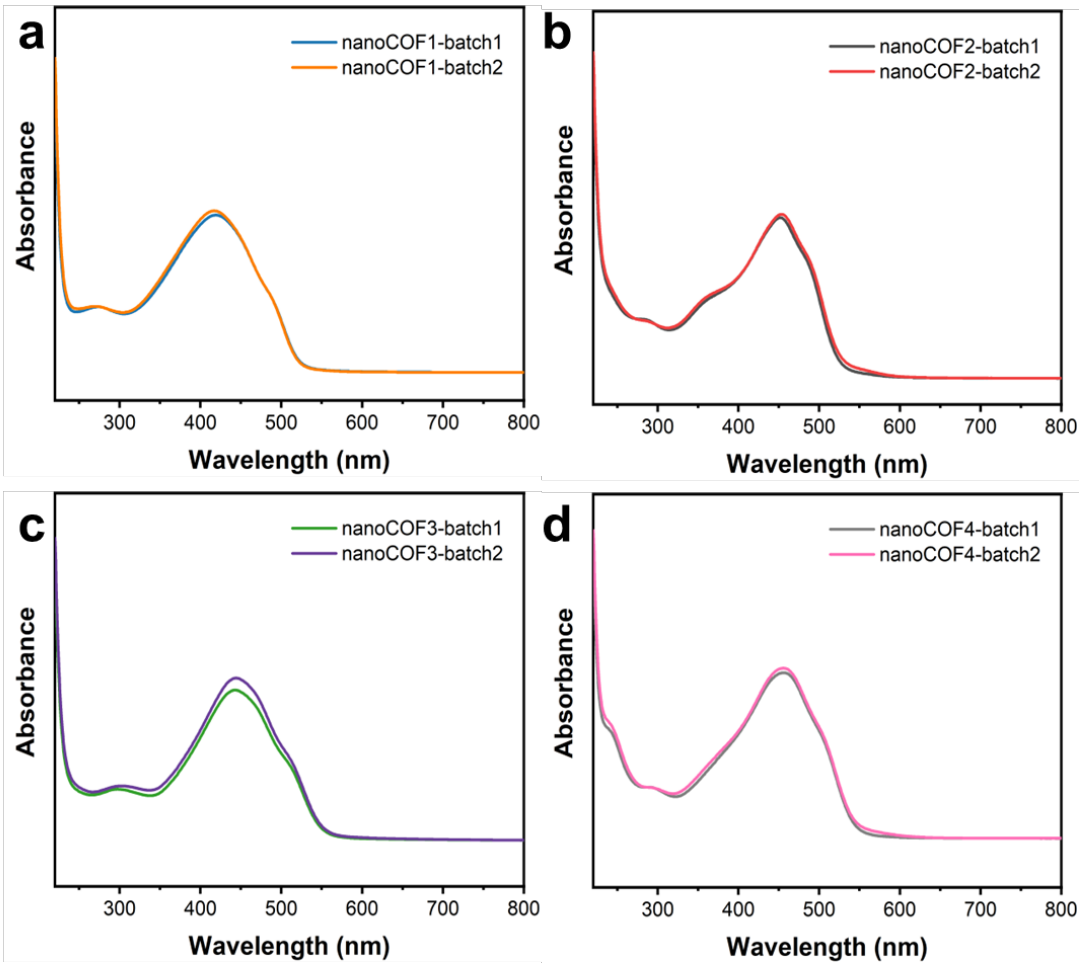

Figure S1. UV-vis absorbance data for batches 1 and 2 of (a) nanoCOF1; (b) nanoCOF2; (c) nanoCOF3 and (d) nanoCOF4 (0.05 mL of each nanoCOF colloids was diluted with water to a total volume of 3 mL for these spectra).

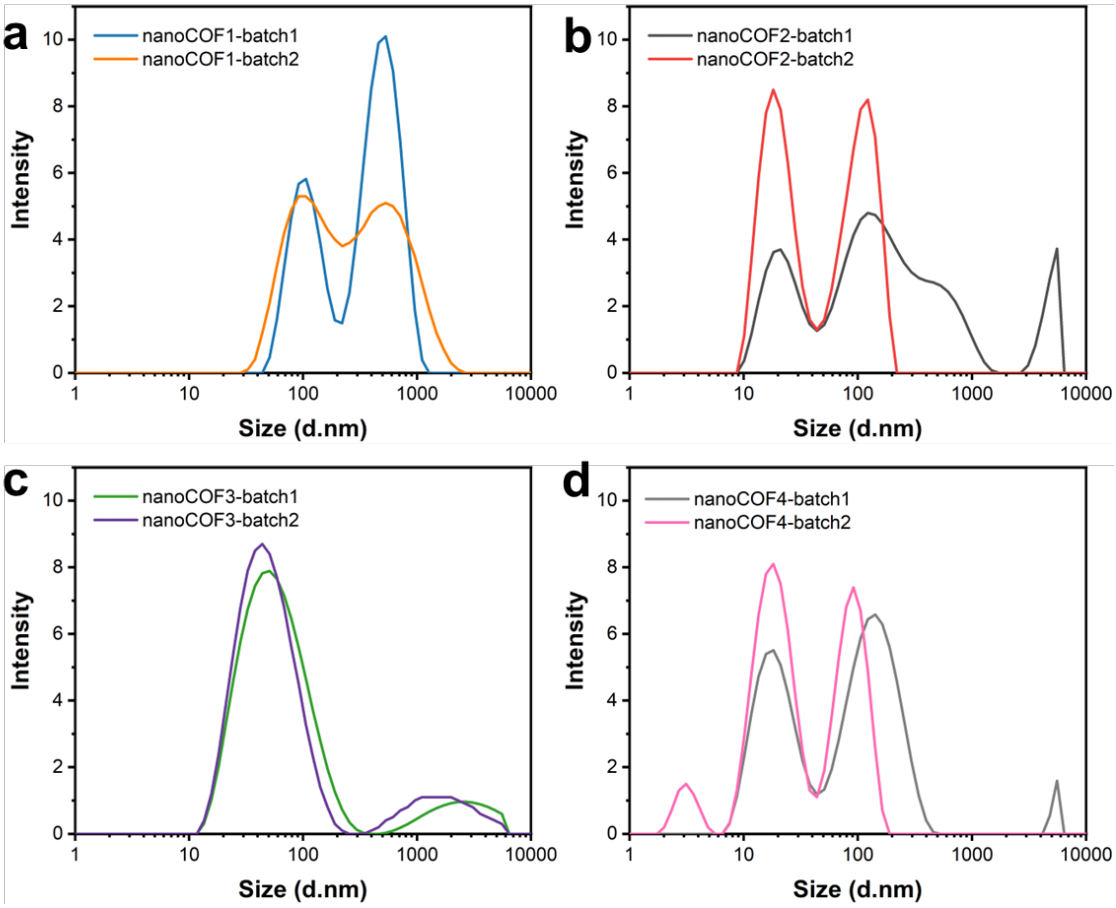

Figure S2. Particle size distribution data from dynamic light scattering measurements for batches 1 and 2 of (a) nanoCOF1, (b) nanoCOF2, (c) nanoCOF3, and; (d) nanoCOF4 (0.5 mL of each nanoCOF colloidal solution was diluted by water to a total volume of 5 mL before testing using a Zetasizer instrument.

1.4 Isolation of nanoCOFs

To isolate solid nanoCOF samples for characterisation, the colloidal nanoCOF solutions were neutralised with concentrated ammonia solution (6.8 mL) and ethanol was added (100 mL). The dispersions were centrifuged for 5 minutes at 5000 rpm. After removing the supernatant, the solids were washed with ethanol and centrifuged for 5 times. Finally, the solid samples were obtained by freeze-drying for FT-IR and UV-vis DRS tests.

1.5 Fourier transform infrared (FT-IR) spectra of the nanoCOFs

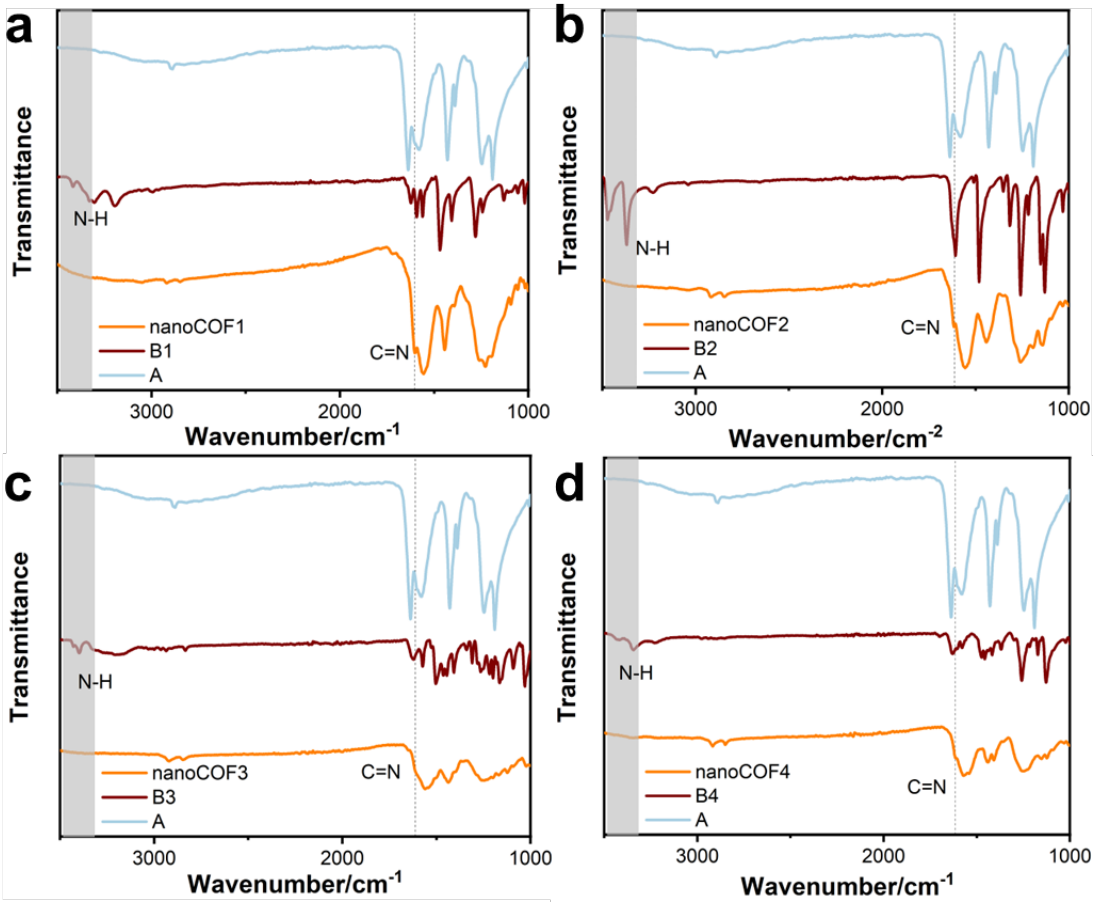

Figure S3. FT-IR results of monomers and their corresponding nanoCOFs. (a) nanoCOF1, (b) nanoCOF2, (c) nanoCOF3, and; (d) nanoCOF4. Monomer A = 2,4,6-trihydroxybenzene-1,3,5-tricarbaldehyde; Monomer B1 = [2,2'-bipyridine]-5,5'-diamine; Monomer B2 = 3,7-diaminodibenzo[b,d]thiophene 5,5-dioxide; Monomer B3 = o-dianisidine; Monomer B4 = 3,7-diamino-2,8-dimethyldibenzo-thiophene sulfone. Shaded region = N-H stretching region (monomers or terminal amines in nanoCOFs).

1.6 Stock solution preparation

Stock solution	Concentrations (M)	Water Volume (mL)	Mass (g)
L-cysteine	0.5	100	6.06
ascorbic acid	0.5	100	8.81
TEOA	0.5	100	7.46
CTAB	0.01	40	0.15
SDS	0.01	40	0.12
PVP	0.01	40	4.00
NaOH	0.01	40	0.02

Table S4. Stock solution concentrations for components added as aqueous solutions in the workflow (solutions prepared by weighing).

Stock solution	Concentration	Volume (mL)	Mother liquor concentration
nanoCOF1	0.0686 mg/mL	40	0.343 mg/mL
nanoCOF2	0.0828 mg/mL	40	0.414 mg/mL
nanoCOF3	0.0823 mg/mL	40	0.412 mg/mL
nanoCOF4	0.0894 mg/mL	40	0.447 mg/mL
H_2PtCl_6	0.0072 mg/mL	40	84 mg/mL
HCl	0.01 M	40	12 M

Table S5. Stock solution concentrations for components added as aqueous solutions in the workflow (solutions prepared by dilution).

Stock solutions of the nanoCOFs were prepared by taking the nanoCOF colloids (8 mL) and diluting with deionised water (32 mL). The solution was then sonicated for 15 minutes to fully disperse the nanoCOFs.

H_2PtCl_6 stock solution (0.0072 mg/mL) was prepared through a series of serial dilutions. An aqueous H_2PtCl_6 solution (84 mg/mL) was first diluted 8-fold (1 mL H_2PtCl_6 solution + 7 mL DI water) to obtain a 10 mg/mL solution. Subsequently, 0.1 mL of the 10 mg/mL solution was diluted with 0.9 mL DI water to obtain a 1 mg/mL solution. Then, 0.3 mL of the 1 mg/mL solution was mixed with 2.7 mL DI water to obtain a 0.1 mg/mL solution. Finally, 2.88 mL of the 0.1 mg/mL solution was diluted with 37.12 mL DI water to obtain the final H_2PtCl_6 stock solution (0.0072 mg/mL).

To prepare a 0.01 M HCl solution, concentrated HCl (33 µL, 12 M) was added to 40 mL of DI water and mixed thoroughly.

Stock solutions were changed every four days or upon running out of stock solution on the workflow (whichever came first). To minimize the effect of dissolved oxygen in water on the photocatalytic hydrogen evolution half-reaction, N_2 was continuously bubbled through the stock solutions.

2 Characterisation methods

2.1 Fourier transform infrared spectroscopy (FT-IR)

FT-IR spectra of the isolated nanoCOF solid samples were measured on a FT-IR Bruker Alpha instrument with ATR and KBr disk attachments with Windows 7 PC running Bruker OPUS instrument control and data analysis software.

2.2 UV-vis diffuse reflectance spectroscopy (UV-vis DRS)

The UV-visible absorption spectra of the samples were measured using a Shimadzu UV-2600i UV-vis spectrometer. For solid-state measurements, the reflectance of the powder samples was recorded. The optical band gaps were estimated from Tauc plots based on the Tauc relation[1]:

$$\alpha hv = A(hv - E_g)^n$$

where α is the absorption coefficient, hv is the photon energy, A is a proportionality constant, E_g is the optical band gap, and n depends on the nature of the electronic transition. For an allowed direct transition, $n = \frac{1}{2}$, for an allowed indirect transition, $n = 2$. Accordingly, the direct and indirect band gaps were obtained by plotting αhv^2 versus hv for a direct band gap, $\alpha hv^{0.5}$ versus hv for an indirect band gap and extrapolating the linear portion of each curve to $\alpha hv^{\frac{1}{n}} = 0$ to obtain E_g.

NanoCOFs have been reported to have a direct bandgap[2], anatase appears to have an indirect band gap, rutile has direct band gap[3], and ZnSe has a direct bandgap[4]. Bulk SnO_2 was reported to be a direct band gap semiconductor[5], WO_3 was reported to have an indirect bandgap[6], while for SnO_2 nanoparticles[7], V_2O_5[8], and Fe_2O_3[1], there is on-going discussion of their bandgap type. By fitting Tauc plots, we extracted both direct and indirect optical gaps of the materials due to lack of consensus in the literature around their classification.

The CB and VB of the materials were calculated based on the empirical formula[9]:

$$E_{CB} = \chi - E_e - \frac{1}{2}E_g$$

$$E_{VB} = E_{CB} + E_g$$

χ is the absolute electronegativity of the semiconductor, the χ of the inorganic materials were in Table S6.

Sample	χ
Fe_2O_3	χ = 5.83[9]
SnO_2	χ = 6.24[10]
TiO_2	χ = 5.81[11]
WO_3	χ = 6.59[11]
ZnSe	χ = 4.80[12]
V_2O_5	χ = 6.10[10]

Table S6. Absolute electronegativity (χ) of the inorganic materials used.

The calculated CB and VB of the materials (indirect bandgaps) are listed in Table S7.

Materials	In-direct bandgap (eV)	CB (eV)	VB (eV)
nanoCOF1	1.46	-0.72	0.74
nanoCOF2	1.43	-0.55	0.88
nanoCOF3	1.41	-0.78	0.63
nanoCOF4	1.43	-0.56	0.87
Fe_2O_3-1	1.91	0.37	2.28
Fe_2O_3-2	1.64	0.51	2.15
SnO_2-1	2.93	0.28	3.21
SnO_2-2	3.16	0.16	3.32
TiO_2 anatase-1	3.09	-0.24	2.86
TiO_2 anatase-2	3.05	-0.22	2.84
TiO_2 rutile-1	2.91	-0.15	2.77
TiO_2 rutile-2	2.89	-0.14	2.76
WO_3-1	2.57	0.81	3.38
WO_3-2	2.65	0.77	3.42
ZnSe	1.80	-0.60	1.20
V_2O_5	2.17	0.52	2.69

Table S7. The conduction band (CB), valence band (VB) and the calculated indirect bandgap of the semiconducting materials

The calculated CB and VB of the materials (direct bandgaps) are listed in Table S8.

Materials	Direct bandgap (eV)	CB (eV)	VB (eV)
nanoCOF1	1.96	-0.72	1.24
nanoCOF2	1.92	-0.55	1.37
nanoCOF3	1.75	-0.78	0.97
nanoCOF4	1.82	-0.56	1.26
Fe_2O_3-1	2.10	0.28	2.38
Fe_2O_3-2	2.16	0.25	2.41
SnO_2-1	3.75	-0.14	3.62
SnO_2-2	3.82	-0.17	3.65
TiO_2 anatase-1	3.35	-0.37	2.99
TiO_2 anatase-2	3.35	-0.37	2.99
TiO_2 rutile-1	3.05	-0.22	2.84
TiO_2 rutile-2	3.20	-0.29	2.91
WO_3-1	2.72	0.73	3.45
WO_3-2	3.02	0.58	3.60
ZnSe	2.53	-0.97	1.57
V_2O_5	2.33	0.44	2.77

Table S8. The conduction band (CB), valence band (VB) and the calculated direct bandgap of the semiconducting materials.

2.3 Dynamic light scattering (DLS)

The size distributions of the samples were measured by dynamic light scattering using a Zetasizer Nano ZS instrument from Malvern Instruments Nordic AB. Stock solutions were prepared by dispersing 5 mg of the inorganic materials in distilled water (5 mL), or 0.5 mL of the nanoCOFs colloidal solutions in distilled water (4.5 mL). Particle size distributions of nanoCOFs before the addition of SDS (0.5 mL nanoCOF2 (0.0828 mg/mL), 0.5 mL H_2PtCl_6 (0.0072 mg/mL), 2 mL AA (0.5M) plus 2 mL distilled water) were also analysed using the Zetasizer.

2.4 Particle size distribution (PSD)

Particle size distribution of nanoCOF after the addition of SDS (0.5 mL nanoCOF2 (0.0828 mg/mL), 0.5 mL H_2PtCl_6 (0.0072 mg/mL), 2 mL AA (0.5M), 0.5 mL SDS (0.01M) and 1.5 mL distilled water) was analysed using a Mastersizer 3000 laser diffraction particle size analyser (Malvern Panalytical, UK). The measurements were performed based on the Mie scattering theory. The characteristic particle diameters Dx10, Dx50 and Dx90 were obtained from the volume-based particle size distribution.

2.5 Zeta potential tests

The zeta potentials of the samples were measured by Zetasizer Nano ZS from Malvern Instruments Nordic AB. Stock solutions were prepared by dispersing 0.5 mg of the inorganic materials in distilled water (5 mL), or 0.5 mL of the nanoCOFs colloids in distilled water (4.5 mL). NanoCOFs before and after the addition of SDS were also analysed using the Zetasizer (before: 0.5 mL nanoCOF2 (0.0828 mg/mL), 0.5 mL H_2PtCl_6 (0.0072 mg/mL), 2 mL AA (0.5M) and 2 mL distilled water; after: 0.5 mL nanoCOF2 (0.0828 mg/mL), 0.5 mL H_2PtCl_6 (0.0072 mg/mL), 2 mL AA (0.5M), 0.5 mL SDS (0.01M) and 1.5 mL distilled water).

2.6 Oxidation potentials of the sacrificial agents

Sacrificial Agent	Oxidation Potential (V vs. NHE, pH=0)
Ascorbic acid (AA)	0.39[13]
Triethanolamine (TEOA)	1.21[14]
L-Cysteine	0.19[15]

Table S9. The oxidation potentials (vs NHE) of the sacrificial agents at pH 0 (includes source references for oxidation potential values).

2.7 Photoluminescence spectroscopy (PL)

Photoluminescence (PL) measurements were performed on an Edinburgh FLS1000 Photoluminescence Spectrometer. For all measurements, 100 µL of the liquid sample was deposited into a quartz solid-sample holder to prevent the sample from aggregating and then settling at the bottom of the liquid sample cell.

The excitation spectrum was first recorded by monitoring the emission at 630 nm using a 610 nm long-pass filter over the wavelength range of 250–600 nm. The corresponding emission spectrum was recorded with an excitation wavelength of 450 nm using a 475 nm long-pass filter over the wavelength range of 480–800 nm. Subsequently, an additional excitation spectrum was recorded by monitoring the emission at 450 nm using a 420 nm long-pass filter over the wavelength range of 250–400 nm, and the corresponding emission spectrum was recorded with an excitation wavelength of 375 nm using a 420 nm long-pass filter over the wavelength range of 430–800 nm. For all measurements, the excitation and emission bandwidths were set to 3 nm, and each spectrum was acquired with 10 accumulations.

Sample 1: 0.5 mL nanoCOF2 (0.0828 mg/mL), 0.5 mL H_2PtCl_6 (0.0072 mg/mL), 2 mL AA (0.5M), 0.5 mL SDS (0.01M) and 1.5 mL distilled water).

Sample 2: 0.5 mL nanoCOF2 (0.0828 mg/mL), 0.5 mL H_2PtCl_6 (0.0072 mg/mL), 2 mL AA (0.5M) and 2 mL distilled water).

3 Workflow specification

3.1 Workflow hardware

The workflow comprised six stations. 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 (see section 3.1.1 for more details). A bespoke instrument (Labman) was used to perform sample inertisation with nitrogen (to exclude oxygen) and cap crimping in one step. Sample illumination was carried out at a photolysis station (section 3.1.1). Gas chromatography measurements were performed with a 7890B GC and a 7697A Headspace Sampler from Agilent GC. Transfer between stations was performed by a LBR iiwa 14 R820 Kuka Mobile Robot mounted on a KMP 200 KUKA Mobile Platform base. Communication between the process management system and the stations was achieved using various communication protocols (TCP/IP over WIFI/LAN; RS-232).

3.1.1 Liquid station and source

The custom-built liquid dispenser was modified to reduce cross-contamination between the different liquid lines and to improve dispense accuracy in the vial. A custom dispense head was designed that replaced the existing dispense head (Figure S4), while keeping the rest of the hardware unchanged from that reported previously (ref. 9 in main text).

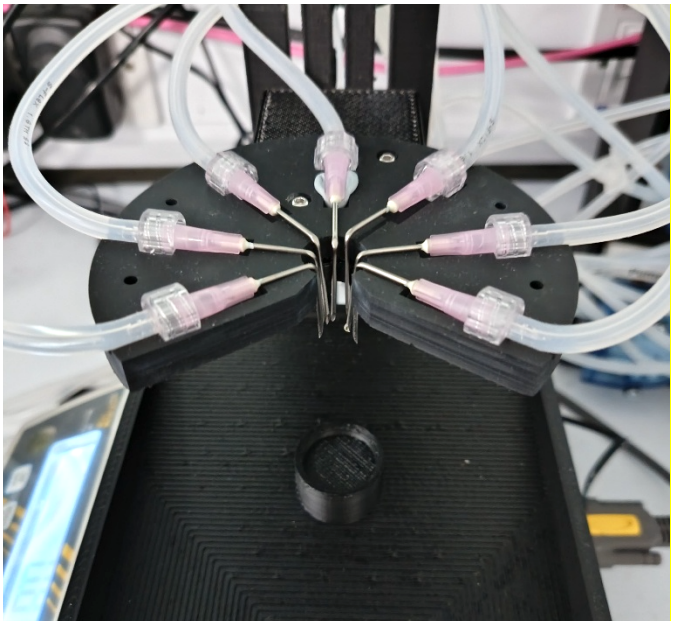

Figure S4. Newly designed 7-channel liquid dispense head. The needles are a friction fit into the slot to prevent them from moving during dispensing.

A custom light station was designed to illuminate the samples from the bottom. A vibrating sample rack using a vibratory motor (VIBTEC, M3/20) was mounted to the top of an aluminium frame. A G2V Sunbrick^{TM} AAA LED solar simulator was mounted upside down on the aluminium frame. The intensity of the LEDs of the solar simulator was tuned to match the AM 1.5 spectrum. The spectrum of the light source is shown below (Figure S5). The intensity of this light source and its fidelity to the solar spectrum was a major improvement on the light source that was reported previously (ref. 9 in main text).

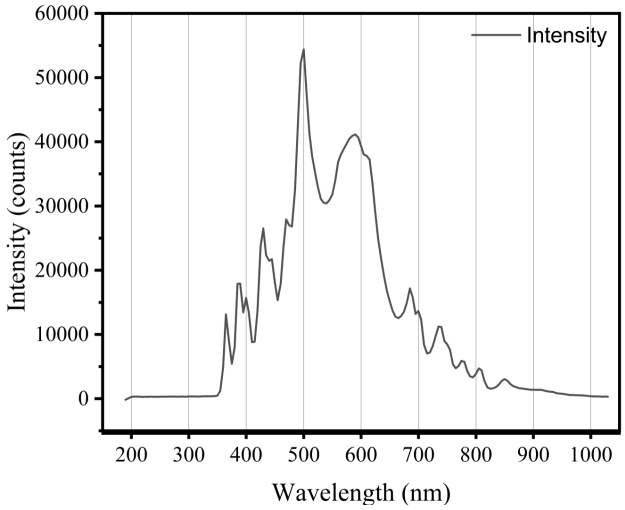

Figure S5: Light spectrum of G2V Sunbrick^{TM} AAA LED solar simulator

3.1.2 GC detector saturation in Phase 1

Visible peak splitting indicated saturation of the GC detector. This started to occur at a raw GC reading of 8.95 µmol/mL (Figure S6) (61 µmol h^{-1} on scale of Figure 1d). We set the cutoff for the GC detector saturation slightly below this at 8.75 µmol/mL (Figure S7).

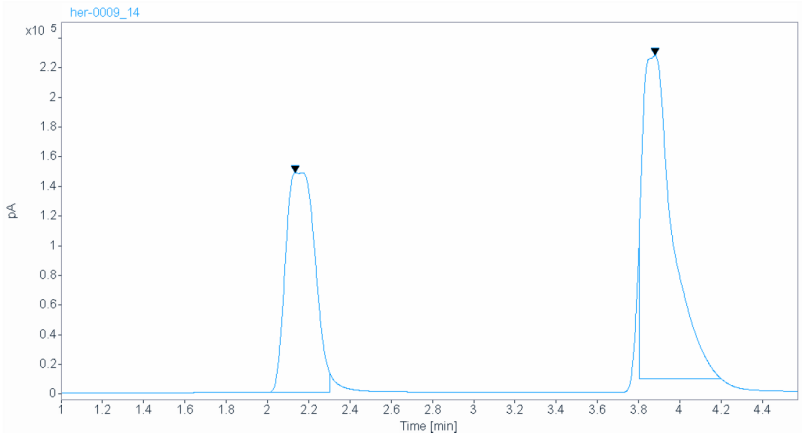

Figure S6. Raw GC spectra obtained from sample 142. First visible splitting of peaks with hydrogen production of 8.95 µmol/mL at ~2.2 min.

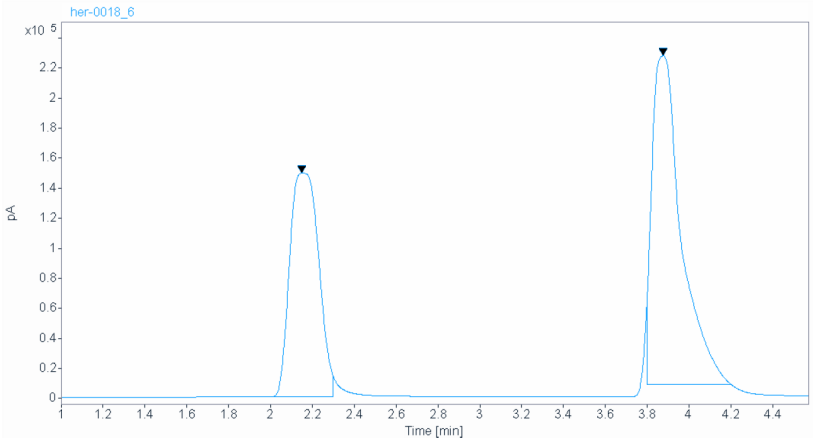

Figure S7. Raw GC spectra obtained from sample 278. GC peak corresponding to a hydrogen production of 8.75 µmol/mL (60 µmol/h) at ~2.2 min.

We note that the LLM was provided with the peak areas only – as such, it could not detect the peak splitting shown in Figure S6.

3.2 Web bridge and Opsight integration

The bridging software layer referenced in the Methods section of the main text was implemented as a small, standalone service running alongside the mobile-robot workflow management system in the lab. It was deliberately thin: it held no optimisation state, ran no chemistry logic, and was the only component that touched both the public Opsight REST API and the lab's local file system. Opsight owned the campaign state, optimiser model, and batch recommendations. The lab workflow system owned scheduling, the robot, and the instrument stations. The web bridge handled format translation, authentication, and synchronisation between the two software platforms, as shown in Figure S8. The details of the Opsight software platform architecture are summarised in Figure S9.

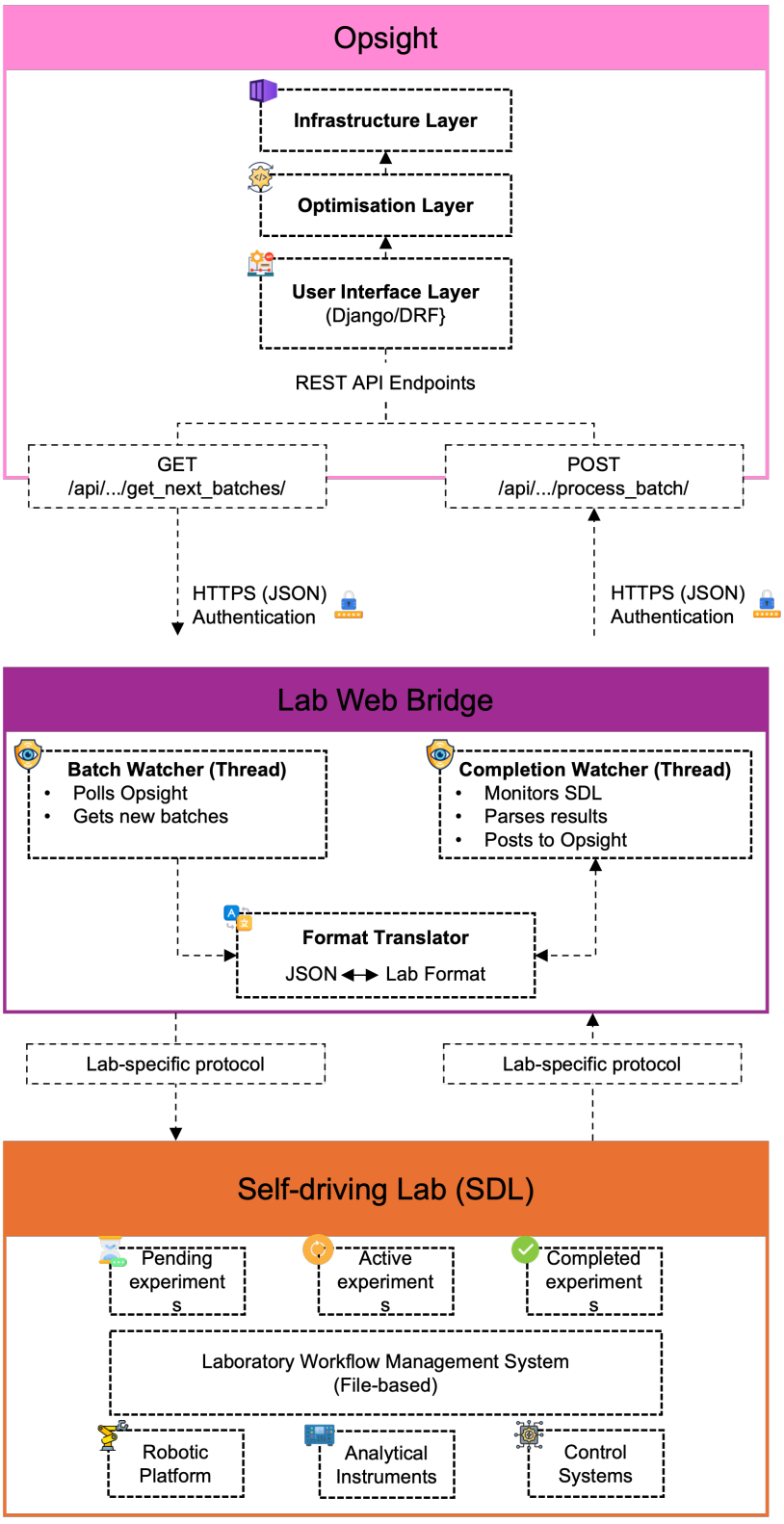

Figure S8. Integration architecture for connecting self-driving labs to Opsight. The Lab Web Bridge service polled Opsight for new batches via a REST API, translated them into laboratory-specific formats, monitored the lab workflow folders for completed experiments, and posted results back to Opsight to trigger the recommendation of the next batch. Graphics include elements from Flaticon.com.

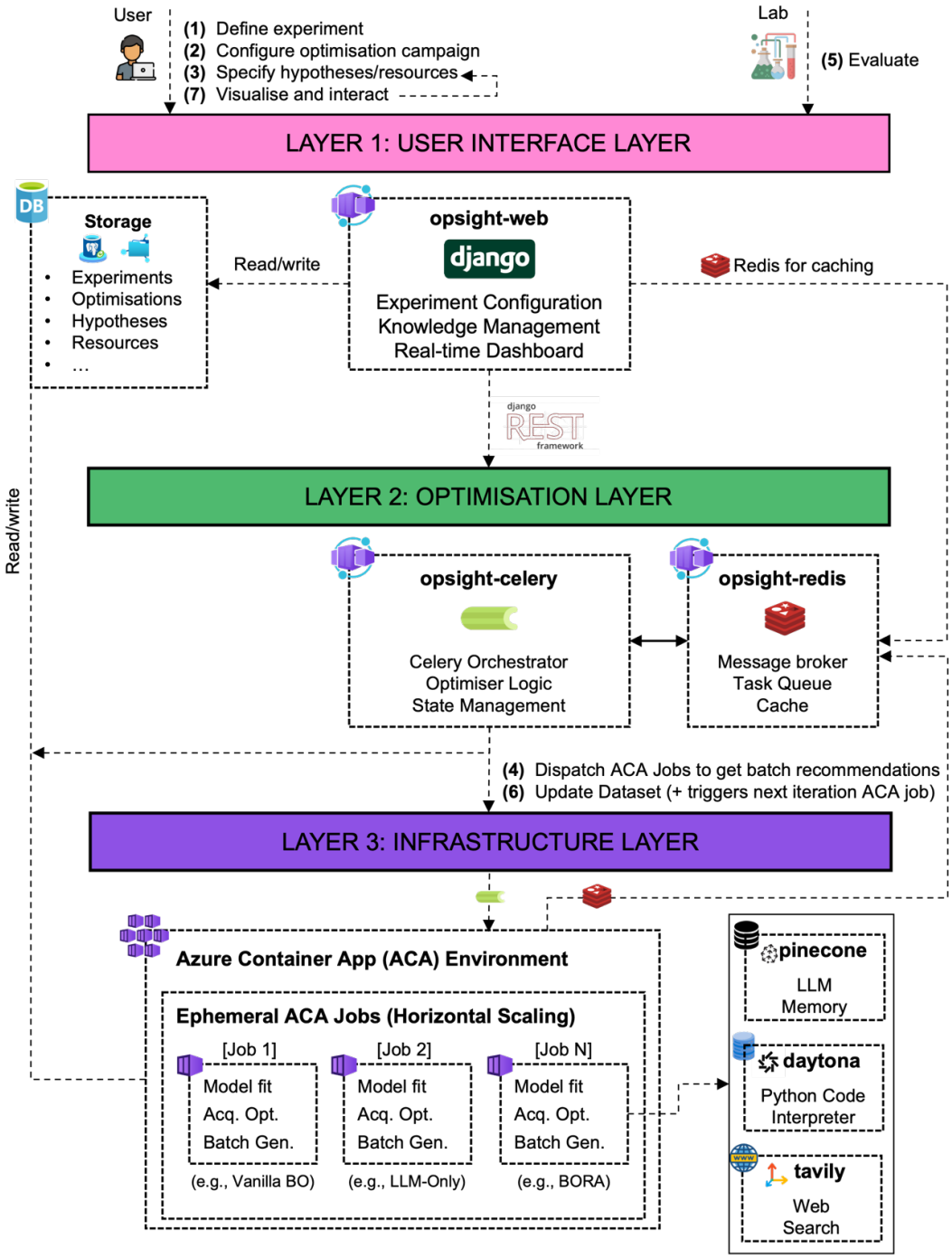

Figure S9. Details of Opsight software architecture for closed-loop experiments. The diagram illustrates the Azure Container App (ACA) services (opsight-web, opsight-celery, opsight-redis), their communication through a PostgreSQL database and a Redis message broker, external infrastructure services (Pinecone vector database, Daytona AI code interpreter, Tavily search) for LLM-based optimisers (*e.g.*, LLM-Only or BORA), and the seven-stage optimisation workflow: (1) experiment definition, (2) campaign configuration, (3) knowledge specification, (4) batch recommendation generation via distributed ACA jobs, (5) experimental evaluation, (6) dataset update and trigger of next iteration in an ACA job, and (7) real-time visualisation and user interaction. Arrows indicate Application Programming Interface (API) calls, database transactions, and asynchronous task flows forming the closed human-in-the-loop optimisation cycle. Graphics include elements from Flaticon.com.

3.2.1 Folder-based contract with the lab workflow system

The lab workflow system drove the Kuka mobile robot from plain-text ".run" files in three watch folders ("runqueue/", "running/", "completed/"), with one file per 16-vial batch, corresponding to the optimisation campaign batch size. The bridge plugged into this existing contract without modification: new batches appeared as files in "runqueue/", the lab system moved them to "running/" while executing and wrote back to "completed/" once gas-chromatography analysis was finished. The web bridge never moves files between folders itself, keeping responsibility for sample integrity entirely with the workflow system.

3.2.2 Control flow

The bridge ran two short polling loops against the specified live Opsight optimisation campaign.

Direction action	Trigger	Opsight call	Web bridge side
Opsight -> Web bridge	New batch from Opsight	GET /optimisations/{id}/get_next_batches/	Writes into "runqueue/"
Web bridge -> Opsight	New file appears in "completed/"	POST /optimisations/{id}/process_batch/	Parses results

Table S10. Web bridge and Opsight communication via Opsight endpoints, and corresponding lab folder actions.

In the forward direction (Opsight -> Web bridge), the web bridge received the generated batch of 16 catalyst formulations from Opsight, appended per-batch control samples, filled the water column to the configured total volume (5 mL), and rendered the batch 16 × 25 component matrix against a plain-text ".run" template that also carried the illumination time, GC method, and rack identifier.

In the reverse direction (Web bridge -> Opsight), the bridge parsed the batch results, written as a CSV block by the lab workflow system, and posted the formulation matrix and measured H_2 evolution targets to Opsight. In the same cycle, it wrote the actual as-dispensed values (plus batch/sample identifiers and measured target) into a "lab_data_logs" CSV file, treated as the 11^{th} resource file. The "lab_data_logs" file was first populated from batch 1, then updated after every subsequent completed batch and uploaded back to Opsight, so the LLM also saw the actual dispensed liquid and solid amounts on later batches, in addition to the target dispensing values. By using both resources, the LLM was able to make judgements about execution errors—that is, the difference between the dispense amounts and the target amounts. These differences were mostly low, particularly for liquid dispensing (Table S22, S23).

4 Opsight optimisation campaign state management

Throughout an optimisation campaign, a complete transactional deep copy of the up-to-date campaign state and linked records (samples, comments, hypotheses, resources, instruction prompt versions, and optimiser-state snapshots) was kept on a database. At any iteration point, a chosen per-batch snapshot could be saved under a different namespace. This facilitated the following functionalities that allowed us to study the LLM's stochasticity and its performance comparison against humans.

4.1 Duplication

At a chosen iteration checkpoint, the up-to-batch database was copied under a different namespace and new runtime task identifiers were generated so a new campaign could be run as a completely independent run. Using this campaign duplication, we could test the stochasticity of the LLM (*e.g.*, the 20 repeat initialisations in batches 1 and 24, see main text).

4.2 Rollback

A campaign rollback restores a selected batch snapshot while rolling resources and/or instructions back through their version histories, removing any post-target samples, logs and snapshots, and setting the run to a 'paused' status (see Section 6.1).

4.3 Intervention

At any batch checkpoint, the up-to-date campaign can be saved, assigned a paused status, and re-anchored under additional linked information, comments, hypotheses or other resources, offering a chance to provide additional context in the campaign. For example, it would have been possible to mention the phase behaviour changes that we observed that were not apparent to the LLM (Figure 4a, main text). Note that for the human/LLM comparisons in Phase 3 we did not do this, nor make any other interventions apart from those detailed in Extended Data Figure 8.

5 Experiment card and prompts

5.1 Experiment card

We attached the following JSON file to define the experiment. This included the experimental aim, formulation constraints, adjustable liquid and solid components, and the target measurement for hydrogen yield.


```
{
  "name": "Photocatalytic Hydrogen Production using nano-COFs and Inorganic Semiconductors",
  "domain": "Chemistry",
  "description": "The aim is photocatalytic H₂ evolution via two possible mechanisms. (i) A light-
absorbing semiconductor is paired with another light absorbing semiconductor, forming a heterojunction.
The formation of such heterojunction facilitates charge transfer from one semiconductor to another. A
variety of heterojunctions can be useful, including z-scheme, s-scheme and type II heterojunctions. As at
least one of the semiconductor has a conduction band positioned more negatively than the H⁺/H₂ potential
(0 V vs. NHE), molecular hydrogen is driven thermodynamically. Simultaneously, sacrificial electron donors
(ascorbic acid, TEOA, L-cysteine, etc.) quench photogenerated holes in the valence band of the second
semiconductor. The second semiconductor has a conduction band that straddles between the conduction band
and the valence band of the first semiconductor. (ii) Alternatively, both the electrons and holes can be
generated with a single semiconductor photocatalyst, provided that the conduction band of the
photocatalyst is more negative than the H⁺/H₂ potential and the valence band is more positive than the
redox potential of the sacrificial electron donor. By systematically varying the added volume of each
stock component—including co-catalyst (chloroplatinic acid), sacrificial agents, surfactants/dispersants,
pH/ionic-strength modifiers, and a library of organic and inorganic semiconductors—we seek the formulation
that maximizes the H₂ yield. The solution will be topped up with Water till the total solution reaches
5mL. The allowed mass for solid semiconductors are zero and all values from 0.4 up to 10.0 at intervals of
0.2mg. The liquid components are allowed from 0 to 5mL at intervals of 0.25mL.",
  "allow_duplicate_points": true,
  "constraints": {
    "intra-point": [
      {
        "name": "Total Liquid Volume",
        "description": "The sum of all liquid parameters should not exceed 5mL to fit inside the vial.",
        "operator": "<=",
        "weights": {
          "H2PtCl6_0.0072mgmL-1": 1,
          "nanoCOF1_0.0686mgmL-1": 1,
          "nanoCOF2_0.0828mgmL-1": 1,
          "nanoCOF3_0.0823mgmL-1": 1,
          "nanoCOF4_0.0894mgmL-1": 1,
          "AscorbicAcid_0.5M": 1,
          "L-Cysteine_0.5M": 1,
          "TEOA_0.5M": 1,
          "CTAB_0.01M": 1,
          "SDS_0.01M": 1,
          "PVP_0.01M": 1,
          "NaOH_0.01M": 1,
          "HCl_0.01M": 1
        },
        "value": 5
      },
      {
        "name": "Total Solid Mass",
        "description": "The sum of all solid parameters should not exceed 10mg",
        "operator": "<=",
        "weights": {
          "V2O5_mg": 1,
          "TiO2_rutile_sample1_mg": 1,
          "TiO2_anatase_sample1_mg": 1,
          "Fe2O3_sample1_mg": 1,
          "SnO2_sample1_mg": 1,
          "TiO2_rutile_sample2_mg": 1,
          "TiO2_anatase_sample2_mg": 1,
          "Fe2O3_sample2_mg": 1,
          "SnO2_sample2_mg": 1,
          "WO3_sample1_mg": 1,
          "WO3_sample2_mg": 1,
          "ZnSe_mg": 1
        },
        "value": 10
      },
      {
```

```
        "name": "Presence of Light Absorbing semiconductor",
        "description": "There must be at least one light absorbing semiconductor in the formulation.",
        "operator": ">=",
        "weights": {
          "nanoCOF1_0.0686mgmL-1": 1,
          "nanoCOF2_0.0828mgmL-1": 1,
          "nanoCOF3_0.0823mgmL-1": 1,
          "nanoCOF4_0.0894mgmL-1": 1,
          "V2O5_mg": 1,
          "TiO2_rutile_sample1_mg": 1,
          "TiO2_anatase_sample1_mg": 1,
          "Fe2O3_sample1_mg": 1,
          "SnO2_sample1_mg": 1,
          "TiO2_rutile_sample2_mg": 1,
          "TiO2_anatase_sample2_mg": 1,
          "Fe2O3_sample2_mg": 1,
          "SnO2_sample2_mg": 1,
          "WO3_sample1_mg": 1,
          "WO3_sample2_mg": 1,
          "ZnSe_mg": 1
        },
        "value": 0.2
      },
      {
        "name": "No 0_2mg Solid Components",
        "description": "The solid parameters in the formulation should not be 0.2mg.",
        "operator": "nonlinear"
      },
      {
        "name": "Maximum Non-Zero Components",
        "description": "The number of non-zero components in the formulation should not exceed 10.",
        "operator": "nonlinear"
      },
      {
        "name": "No Acid and Base Simultaneously",
        "description": "Either no NaOH and HCl or only one of them can be present in the formulation.",
        "operator": "nonlinear"
      }
    ]
  },
  "parameters": [
    {
      "name": "H2PtCl6_0.0072mgmL-1",
      "description": "Volume of chloroplatinic acid in water (0.0072 mg mL⁻¹) added.",
      "unit": "mL",
      "bounds": [
        0,
        5.0
      ],
      "step": 0.25
    },
    {
      "name": "nanoCOF1_0.0686mgmL-1",
      "description": "Volume of nanoscale-covalent organic frameworks (synthesized with 2,4,6-
triformylphloroglucinol and [2,2'-bipyridine]-5,5'-diamine monomer) suspension (0.0686 mg mL⁻¹) serving as
a light absorbing semiconductor.",
      "unit": "mL",
      "bounds": [
        0,
        5.0
      ],
      "step": 0.25
    },
    {
      "name": "nanoCOF2_0.0828mgmL-1",
      "description": "Volume of nanoscale-covalent organic frameworks (synthesized with 2,4,6-
triformylphloroglucinol and 3,7-diaminodibenzo[b,d]thiophene 5,5-dioxide monomer) suspension (0.0828 mg
mL⁻¹) serving as a light absorbing semiconductor.",
      "unit": "mL",
      "bounds": [
        0,
        5.0
      ],
      "step": 0.25
    },
    {
      "name": "nanoCOF3_0.0823mgmL-1",
```

```
      "description": "Volume of nanoscale-covalent organic frameworks (synthesized with 2,4,6-
triformylphloroglucinol and 3,3'-dimethoxy-[1,1'-biphenyl]-4,4'-diamine monomer) suspension (0.0823 mg
mL⁻¹) serving as a light absorbing semiconductor.",
      "unit": "mL",
      "bounds": [
        0,
        5.0
      ],
      "step": 0.25
    },
    {
      "name": "nanoCOF4_0.0894mgmL-1",
      "description": "Volume of nanoscale-covalent organic frameworks (synthesized with 2,4,6-
triformylphloroglucinol and 3,7-diamino-2,8-dimethyldibenzo[b,d]thiophene 5,5-dioxide monomer) suspension
(0.0894 mg mL⁻¹) serving as a light absorbing semiconductor.",
      "unit": "mL",
      "bounds": [
        0,
        5.0
      ],
      "step": 0.25
    },
    {
      "name": "L-Cysteine_0.5M",
      "description": "Volume of 0.5 M L-cysteine solution used as sacrificial donor.",
      "unit": "mL",
      "bounds": [
        0,
        5.0
      ],
      "step": 0.25
    },
    {
      "name": "AscorbicAcid_0.5M",
      "description": "Volume of 0.5 M ascorbic acid solution acting as sacrificial electron donor.",
      "unit": "mL",
      "bounds": [
        0,
        5.0
      ],
      "step": 0.25
    },
    {
      "name": "TEOA_0.5M",
      "description": "Volume of 0.5 M triethanolamine (TEOA) solution acting as hole scavenger.",
      "unit": "mL",
      "bounds": [
        0,
        5.0
      ],
      "step": 0.25
    },
    {
      "name": "CTAB_0.01M",
      "description": "Volume of 0.01 M cetyltrimethylammonium bromide (CTAB) surfactant solution for
dispersion stabilization.",
      "unit": "mL",
      "bounds": [
        0,
        5.0
      ],
      "step": 0.25
    },
    {
      "name": "SDS_0.01M",
      "description": "Volume of 0.01 M sodium dodecyl sulfate (SDS) anionic surfactant solution.",
      "unit": "mL",
      "bounds": [
        0,
        5.0
      ],
      "step": 0.25
    },
    {
      "name": "PVP_0.01M",
      "description": "Volume of 0.01 M polyvinylpyrrolidone (PVP) polymer stabilizer solution.",
      "unit": "mL",
```

```
    "bounds": [
      0,
      5.0
    ],
    "step": 0.25
  },
  {
    "name": "NaOH_0.01M",
    "description": "Volume of 0.01 M NaOH solution (base) used to adjust reaction pH.",
    "unit": "mL",
    "bounds": [
      0,
      5.0
    ],
    "step": 0.25
  },
  {
    "name": "HCl_0.01M",
    "description": "Volume of 0.01 M HCl solution (acid) used to adjust reaction pH.",
    "unit": "mL",
    "bounds": [
      0,
      5.0
    ],
    "step": 0.25
  },
  {
    "name": "V2O5_mg",
    "description": "Mass of V₂O₅ powder, a light absorbing semiconductor.",
    "unit": "mg",
    "bounds": [
      0,
      10.0
    ],
    "step": 0.2
  },
  {
    "name": "TiO2_rutile_sample1_mg",
    "description": "Mass of sample 1 of rutile TiO₂ powder, a light absorbing semiconductor.",
    "unit": "mg",
    "bounds": [
      0,
      10.0
    ],
    "step": 0.2
  },
  {
    "name": "TiO2_anatase_sample1_mg",
    "description": "Mass of sample 1 of anatase TiO₂ powder, a light absorbing semiconductor.",
    "unit": "mg",
    "bounds": [
      0,
      10.0
    ],
    "step": 0.2
  },
  {
    "name": "Fe2O3_sample1_mg",
    "description": "Mass of sample 1 of ferric oxide (Fe₂O₃) powder, a light absorbing semiconductor.",
    "unit": "mg",
    "bounds": [
      0,
      10.0
    ],
    "step": 0.2
  },
  {
    "name": "SnO2_sample1_mg",
    "description": "Mass of sample 1 of tin(IV) oxide (SnO₂) powder, a light absorbing semiconductor.",
    "unit": "mg",
    "bounds": [
      0,
      10.0
    ],
    "step": 0.2
  },
  {
```

```
      "name": "TiO2_rutile_sample2_mg",
      "description": "Mass of sample 2 of rutile TiO₂ powder, a light absorbing semiconductor.",
      "unit": "mg",
      "bounds": [
        0,
        10.0
      ],
      "step": 0.2
    },
    {
      "name": "TiO2_anatase_sample2_mg",
      "description": "Mass of sample 2 of anatase TiO₂ powder, a light absorbing semiconductor.",
      "unit": "mg",
      "bounds": [
        0,
        10.0
      ],
      "step": 0.2
    },
    {
      "name": "Fe2O3_sample2_mg",
      "description": "Mass of sample 2 of ferric oxide (Fe₂O₃) powder, a light absorbing semiconductor.",
      "unit": "mg",
      "bounds": [
        0,
        10.0
      ],
      "step": 0.2
    },
    {
      "name": "SnO2_sample2_mg",
      "description": "Mass of sample 2 of tin(IV) oxide (SnO₂) powder, a light absorbing semiconductor.",
      "unit": "mg",
      "bounds": [
        0,
        10.0
      ],
      "step": 0.2
    },
    { "name": "WO3_sample1_mg",
      "description": "Mass of sample 1 of tungsten(VI) oxide (WO₃) powder, a light absorbing
semiconductor.",
      "unit": "mg",
      "bounds": [
        0,
        10.0
      ],
      "step": 0.2
    },
    { "name": "WO3_sample2_mg",
      "description": "Mass of sample 2 of tungsten(VI) oxide (WO₃) powder, a light absorbing
semiconductor.",
      "unit": "mg",
      "bounds": [
        0,
        10.0
      ],
      "step": 0.2
    },
    { "name": "ZnSe_mg",
      "description": "Mass of zinc selenide (ZnSe) powder, a light absorbing semiconductor.",
      "unit": "mg",
      "bounds": [
        0,
        10.0
      ],
      "step": 0.2
    }
  ],
  "target": {
    "name": "Hydrogen Produced_umol",
    "description": "Amount of hydrogen evolved in the photocatalytic system, quantified by gas
chromatography after irradiation.",
    "unit": "µmol"
  }
}
```

An overview of the experimental campaign across different phases is shown in Figure . Phases 1 and 2 were shared across all Phase 3 experiments (Human Phase 3a; LLM Phase 3b1/3b2, and Phase 3c). Only Phase 1 was performed with an illumination time of 60 min; the rest of the experiments were performed with illumination time of 15 min. A noise-handing system prompt was used to guide the optimisation under noisy conditions. The different instruction prompts used for the different phases are detailed in section 5.3.

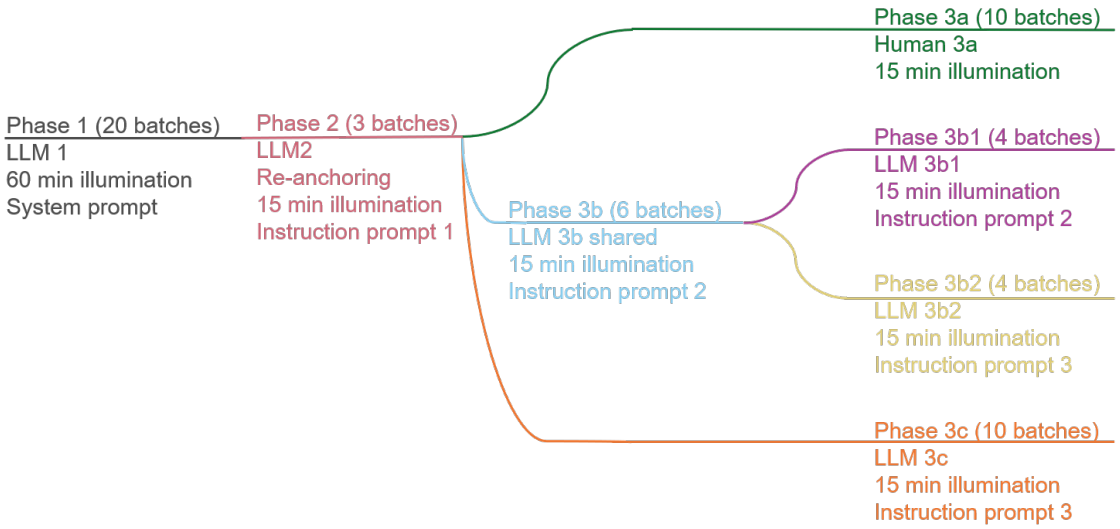

Figure S10. Overview of phases in the experimental campaign (Phases 1, 2, and 3).

5.2 System prompt – noise handling

A noise-handling prompt was used during closed-loop optimisation iterations to cater for a heterogeneous colloidal catalysis experiment that was expected to be noisy and, potentially, inconsistent across repeated measurements. Its purpose was to guide the optimiser to distinguish between true performance trends and variability arising from measurement noise, stochastic experimental effects, or differences between proposed and executed conditions. Our intent was to make each batch noise-aware by explicitly checking prior results, identifying duplicate or near-duplicate experiments, assessing observation and execution noise separately, and allocating new experiments across replicates, local exploitation, exploratory points, and, where needed, calibration anchors. This approach prioritised robust optima rather than single high-performing outliers and ensured that future suggested experiments were informed by both measured performance and data reliability. As discussed in the main text, it did also increase the experimental cost by introducing duplicates. The full noise-handling prompt is available in the accompanying Git repository (section 6.2): https://anonymous.4open.science/r/llm-vs-chemist-closed-loop-catalysis-experiments/

5.3 Instruction prompts

5.3.1 Instruction prompt 1

This prompt was used to initiate Phase 2 (Figure S10) and to renormalise the experiment to a shorter photolysis time.

```
    We have looked at the chromatograms used to calculate the amount of hydrogen evolved, and it seems
that the gas chromatography (GC) detector becomes saturated when the hydrogen production exceeds about
8.75 µmol. As a result, we have shortened the illumination time from 60 minutes to 15 minutes for all
future experiments. Batches 1–20 were conducted with 60-minute illumination time. Because of this detector
saturation, it is possible that the actual quantity of hydrogen evolved is underestimated for experiments
that produce more than about 8.75 µmol. The shorter illumination time of 15 minutes will result in lower
Hydrogen Produced_umol in subsequent batches. The relative performance of each formulation is expected to
remain comparable. If the hydrogen production kinetics were linear in all cases, with no induction time,
then one would expect the amount of hydrogen produced to be reduced by a factor of around 4 for the new 15
minutes irradiation time. We suggest spending batches 21-23 to anchor the new Hydrogen Produced values at
the new, shorter illumination time of 15 minutes, focusing on the higher-performing formulations that we
have tested so far. We will then provide some further input.
```

5.3.2 Instruction prompt 2

This prompt was used at the start of Phase 3b.

```
    This experiment will end after batch 33. Use batches 24 to 33 to maximise the hydrogen production at
the current illumination time (15 minutes). You should look for substantial improvements in hydrogen
production, rather than optimising too narrowly around the current maximum
```

5.3.3 Instruction prompt 3

This adapted prompt was used at the start of Phase 3c.

```
    We have looked at the chromatograms used to calculate the amount of hydrogen evolved, and it seems
that the gas chromatography (GC) detector becomes saturated when the hydrogen production exceeds about
8.75 µmol. As a result, we have shortened the illumination time from 60 minutes to 15 minutes for all
future experiments. Batches 1–20 were conducted with 60-minute illumination time. Because of this detector
saturation, it is possible that the actual quantity of hydrogen evolved is underestimated for experiments
that produce more than about 8.75 µmol. The shorter illumination time of 15 minutes will result in lower
Hydrogen Produced_umol in subsequent batches. The relative performance of each formulation is expected to
remain comparable. If the hydrogen production kinetics were linear in all cases, with no induction time,
then one would expect the amount of hydrogen produced to be reduced by a factor of around 4 for the new 15
minutes irradiation time. We have spent batches 21-23 to anchor the new Hydrogen Produced values at the
new, shorter illumination time of 15 minutes.

# IMPORTANT
Now, with the anchoring phase finished, the updated goal of the entire experiment is **now** to maximize
Hydrogen Produced_umol at the illumination time of 15 minutes.
We plan to end the experiment after batch 33. Use batches 24 to 33 to maximise the hydrogen production at
the illumination time of 15 minutes, focusing on looking for formulations that can lead to substantial
improvements rather than optimising too narrowly around previous formulations.
```

6 LLM-optimiser stochasticity and resource-pack ablations

The LLM optimiser was stochastic, meaning that identical inputs do not produce identical batches. We ran three sets of 20 replicate restarts to (i) characterise that stochasticity at two campaign checkpoints (batch 1 in Phase 1 and batch 24 in Phase 3b2), (ii) verify that the real experimental batches 1 and 24 were not outliers relative to their respective replicate distributions, and; (iii) isolate the effect of the 10 attached resource files.

Arm	Checkpoint	Resources	Prior in context	N proposed formulations
B1_nores	Phase 1, batch 1	No	None	20 × 16 = 320
B1_res	Phase 1, batch 1	Yes	None	20 × 16 = 320
B24_res	Phase 3b2, batch 24	Yes	Batches 1-23 of the full experimental run of Phase 3b2	20 × 16 = 320

Table S11. Repeat batch protocols.

B1_nores and B1_res were paired cold-start arms that differed only in whether the 10 resource files were attached; hence, any statistical difference between them should reflect the influence of those resources on the initial strategy (Extended Data Figure 5). B24_res uses the same instruction prompt setup that started Phase 3b2, so its proposals share the same in-context history as the real batch 24 of Phase 3b2.

6.1 Replicate-restart procedure

The three arms were generated as follows:

- B1_nores and B1_res: 20 fresh campaigns were created from the experiment card. For B1_nores, the 10 resource files were detached from the campaign setup before the first batch was requested. All other elements were left unchanged. The LLM optimiser was then started to propose batch 1.
- B24_res: 20 duplicates were created from the Phase 2 (batch 23 checkpoint). Thus, each replicate inherits the same 23 batches of state from Phases 1 and 2 from the experimental Phase 3b2 campaign. The instruction prompt that marked the start of Phase 3b2 was appended in all cases, and the LLM optimiser was then resumed to propose batch 24.

Across all 60 replicate campaigns, the experiment card, the LLM optimiser with its GPT-5.1-medium model, the batch size of 16, and the resource pack (where applicable) were held constant. The only source of variation between replicates of the same arm is stochastic sampling within GPT-5.1.

6.2 LLM stochasticity

Table S12 reports the centroid distance between the real experimental batch and each of the 20 virtual-repeated batches for each arm, using unit-normalised Euclidean distance in the 25-dimensional formulation space.

Arm	Experimental batch	Mean centroid distance, exptl. -> 20 repeats	Range	Mean repeat-to-repeat centroid distance
B1_nores	Phase 1, batch 1	0.26	0.12 to 0.44	0.28
B1_res	Phase 1, batch 1	0.17	0.08 to 0.24	0.18
B24_res	Phase 3b2, batch 24	0.05	0.03 to 0.14	0.07

Table S12. Experimental batch anchoring across the three repeat arms.

In every arm, the experimental-batch centroid sits within the spread of the replicate distribution: that is, the experimental-to-repeat distance is no larger than the repeat-to-repeat distance. As such, the executed batches were representative of the overall strategy averaged over 20 repeat batches. Removing the resource pack widened the strategy (B1_nores centroid spread was about 1.5× that of B1_res), and 23 batches of in-context history narrowed it further (B24_res is roughly a third of B1_res). This is consistent with a series of experiments where both targeted information (resources) and experimental data narrowed the search.

Per-replicate distances and the corresponding distance-profile figure are provided in Figure .

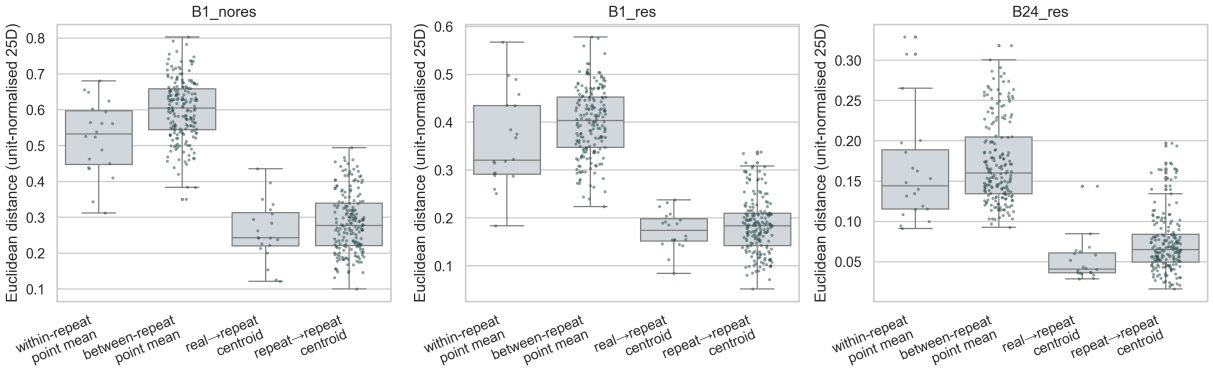

Figure S9. Distance-profile distributions for within-repeat, between-repeat, and real-to-repeat anchoring (“real” = experimental). The *within-repeat point* mean is the average pairwise distance among the 16 formulations in a virtual batch, summarised across the 20 repeats, capturing the spread within a single batch. The *between-repeat point* mean is the average cross-distance between formulations from two different repeats, capturing how far the restart clouds sit from each other.

In all three arms, the between value exceeds the within value by only 6 to 15% (15%, 12%, and 6% for B1_nores, B1_res, and B24_res), so the restart clouds overlap heavily in formulation space, rather than separating into distinct islands. The within-arm strategy structure is explored in section 6.3. Both means shrink together along B1_nores -> B1_res -> B24_res, showing that the resource pack and accumulated experimental context tighten each batch.

Extended Data Figure 4b shows that the eventual Phase 1 best categorical formulation [nanoCOF2 + AA + H_2PtCl_6 + SDS] was almost absent from the 20 B1_res repeats (2/320 proposals in 1/20 repeat runs). Table S13 extends analysis to other closely related categorical formulation families and to the B1_nores arm, again to test whether the relevant formulations were recoverable from latent prior knowledge alone.

Categorical formulation	B1_nores	B1_res	B24_res
LLM-best: [nanoCOF2, nanoCOF4, AA, Pt, SDS]	0 (0 reps)	0 (0 reps)	3 (3 reps)
Human-best: [nanoCOF4, AA, Pt, SDS]	0	3 (>= 1 rep)	2
[one of nanoCOF1/2/3/4, AA, Pt, SDS]	1 (1 rep)	15 (>= 1 rep)	200 (20 reps)
[nanoCOF2, AA, Pt, SDS]	1 (1 rep)	2 (1 rep)	194 (20 reps)

Table S13. Categorical formulation recovery across arms (n proposals out of 320).

Table S13 shows that, with a cold start, none of these formulation families was recovered by the LLM with any consistency, with or without resources. By contrast, by batch 24 (B24_res), the LLM exploits the SDS-including 4-component categorical formulation scaffold in 20/20 (100%) of runs.

6.3 Ablation studies with and without attached resource pack

The cold-start comparison B1_nores vs. B1_res isolates the effect of the 10 resource files on the LLM's initial strategy in batch 1. We characterised this effect along three axes: how concentrated the strategy was (diversity), which components shifted in usage (per-component prevalence), and how broadly across the repeats each shift was carried (replicate coverage). We then read these shifts against a manual inspection of the 20 B1_nores and 20 B1_res reasoning logs.

6.3.1 Strategy diversity

Attaching the resource pack reduced the diversity of the proposed batches without removing it (Table S14). On the pooled 320 proposals per arm, the Vendi score (the effective number of distinct directions covered by the cloud, on a cosine-similarity kernel) fell from 9.3 without resources to 5.8 with resources, and the median pairwise distance fell from 3.95 to 2.64. The same narrowing was present at the level of a single 16-point batch: the per-replicate median Vendi score was 4.55 without resources and 3.43 with resources (two-sided permutation test on the difference in medians across the 20 vs. 20 replicate-level values, n=10000 relabellings, $p < 0.001$). The resource pack therefore focused the strategy but did not lock it onto one single direction.

Arm	Pooled median pairwise distance	Pooled Vendi score	Per-replicate median Vendi score (n=20)
B1_nores	3.95	9.31	4.55
B1_res	2.64	5.78	3.43

Table S14. Strategy diversity at cold start, with and without attached resources.

6.3.2 Per-component prevalence and shifts

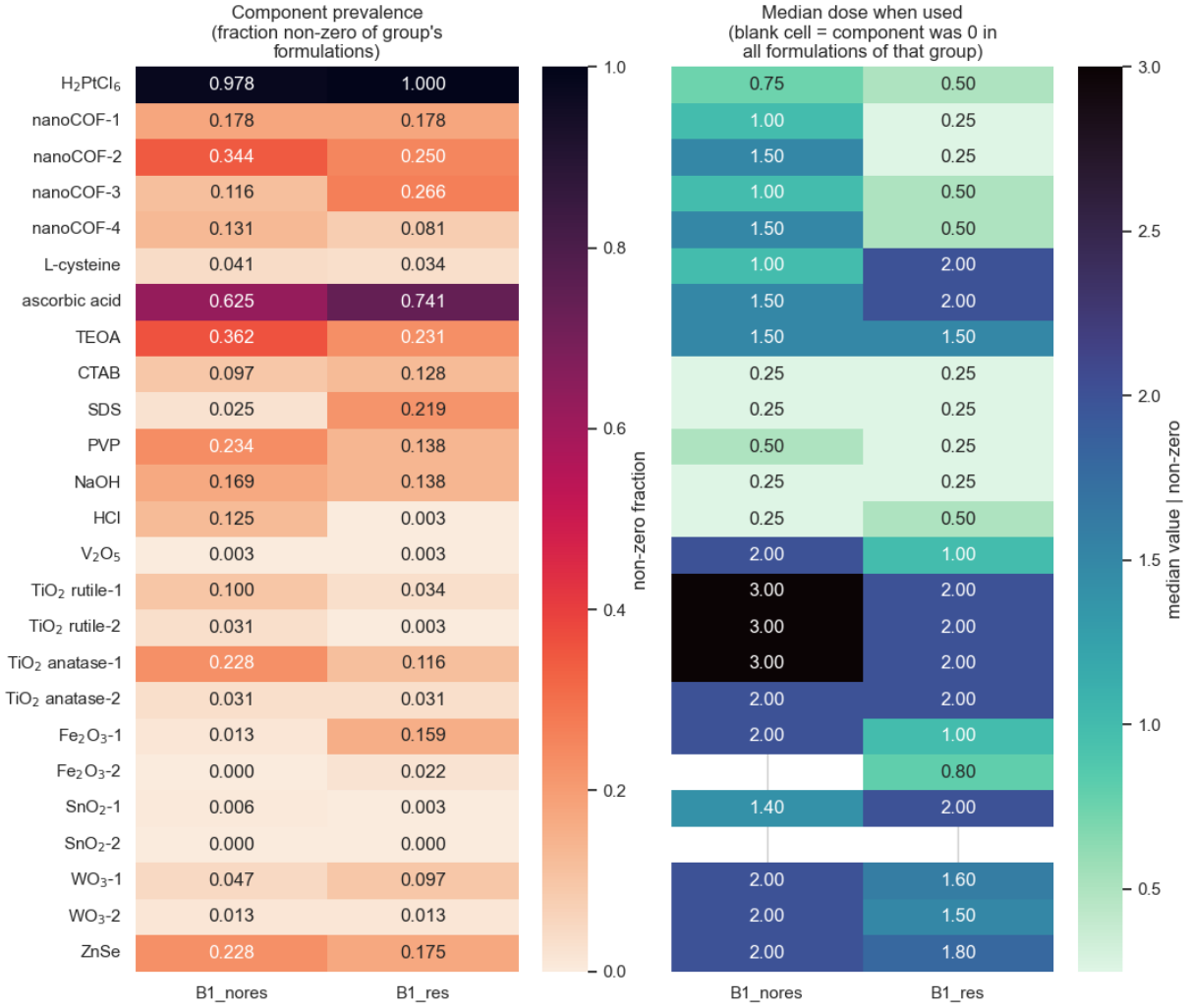

Figure S10. Component-usage profiles for the paired cold-start arms. Left heatmap: per-component prevalence, reported as the fraction of formulations with non-zero amount of each component (out of 320 proposals per arm), for B1_nores and B1_res. Right heatmap: median amount when the component was used (non-zero only) in each arm; blank cells indicate that the component was never selected in that arm. Restricting the view to B1_nores vs B1_res isolates the effect of the resource pack at batch 1.

Figure reports two-sided permutation tests on the 20 vs. 20 replicate-level mean prevalences for the components with the largest pooled differences, with Benjamini-Hochberg (BH) correction within the top-10 family. Five components are significant at $q < 0.05$. The signed shifts split cleanly into two groups: components that the resource pack promotes (SDS, Fe_2O_3-1, nanoCOF3, all used more with resources), and components that the LLM defaults to from general literature when those resources are removed (HCl, TiO_2 anatase-1, with TEOA borderline at $q = 0.10$).

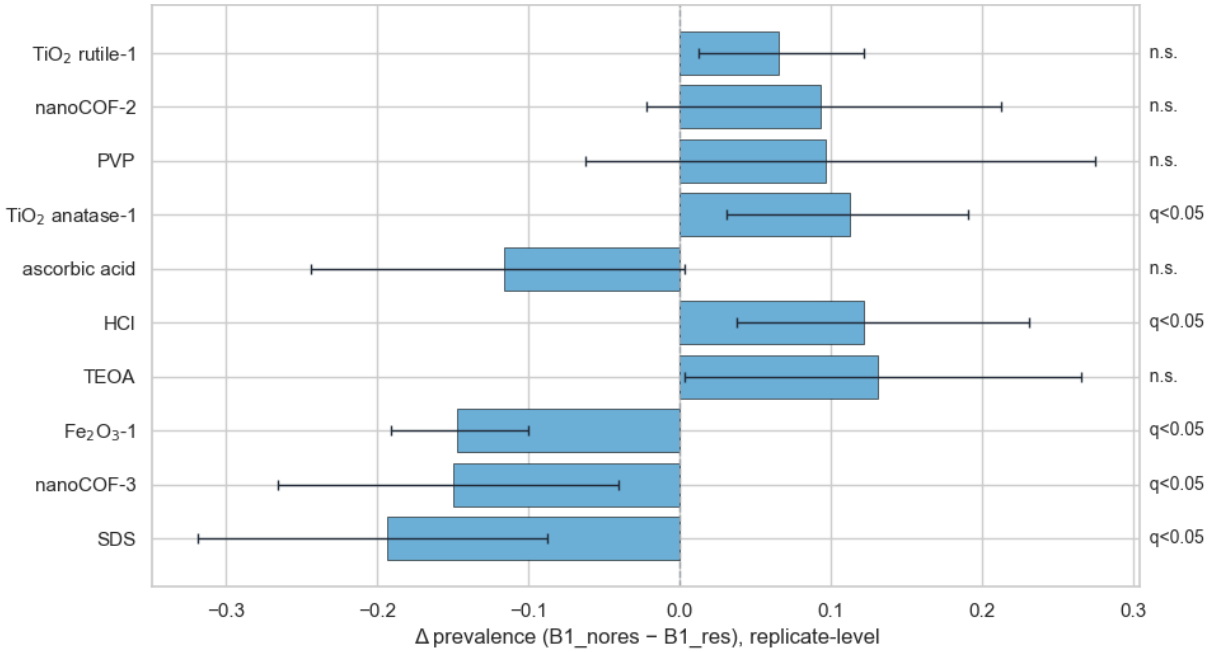

Figure S11. Top 10 component-prevalence shifts (B1_nores vs. B1_res), replicate-level permutation test with BH correction. “n.s” denotes “not significant”.

Table S15 disaggregates SDS and TEOA to the replicate level. The shifts are broad rather than driven by outliers: 14 of 20 B1_res replicates suggest SDS in at least one of their 16 proposals, against only 4 of 20 B1_nores replicates. TEOA was suggested by every single B1_nores replicate, and one B1_nores replicate proposes TEOA in all 16 points of its batch.

Arm	Component	Replicates with ≥ 1 hit	Replicates with all 16 points containing the component
B1_res	SDS	14 / 20 (70%)	1 / 20
B1_nores	SDS	4 / 20 (20%)	0 / 20
B1_res	TEOA	17 / 20 (85%)	0 / 20
B1_nores	TEOA	20 / 20 (100%)	1 / 20

Table S15. Replicate coverage of SDS and TEOA at cold start.

Close inspection of the 20 B1_res log files reveals the mechanism behind these shifts. The LLM consistently identified the key nanoCOF photocatalysis study[2] in 100% of batches and heavily favoured formulations derived from that work. This resulted in the elevated SDS use seen above, with 14/20 replicates suggesting SDS. SDS appearing as the combination [nanoCOF + AA + H_2PtCl_6 + SDS] in 7/20 replicates. The LLM justified SDS use through charge-stabilisation and colloidal-assembly arguments based on the supplied zeta-potential and DLS data. The LLM correctly calculated the nanoCOF loading (≈0.017 mg COF from 0.25 mL of a 0.0686 mg mL^{-1} suspension) and frequently selected nanoCOF2 and nanoCOF3 because of their favourable band structures, particularly the more negative conduction band of nanoCOF3 (Figure , nanoCOF3 being much more prevalent with resources). The model often proposed TEOA-containing formulations, particularly in oxide-based heterojunction systems, commonly paired with NaOH under classical Pt/TiO_2 alkaline conditions; inorganic-only formulations accounted for 22.5% of all B1_res proposals (72/320). HCl was suggested only once across the 20 B1_res replicates, with early COF-

only systems generally favouring near-neutral AA conditions (consistent with the suppression of HCl in Figure).

Pt-loading calculations were attempted in 35% of batches; the model consistently recognised the literature optimum of >=15 wt% Pt and referenced this threshold in 7 batches, although it often treated H_2PtCl_6 as elemental Pt rather than a Pt precursor, only occasionally accounting for the ~40% Pt mass fraction of the salt (the same conversion error discussed in the main text, Phase 0). This points to a general problem with LLM's handling of chemical stoichiometry; that is, the experimental run was not simply 'bad luck'.

For the 20 B1_nores log files, the key nanoCOF paper[2] was found in only 25% of batches, which heavily influenced the direction that the LLM took. For example, without finding the nanoCOF paper, the LLM tried to use nanoCOF loadings close to the literature values used for bulk COF systems (ca. 2–10 mg), leading to higher median component doses than in B1_res (Figure). It also resulted in HCl being used more frequently to lower the pH of ascorbic acid formulations (35% of batches); this was not observed in the with resource batches due to the strong adherence to the experimental conditions reported in the nanoCOF paper. The selection of nanoCOFs was also heavily influenced by the presence of the nanoCOF paper: without it, the LLM favoured nanoCOF2 and nanoCOF4 as they contained sulfone groups, a functionality found to be beneficial in the bulk COF literature. By contrast, when the nanoCOF paper was found, then no real preference for nanoCOF selection was stated, with the LLM instead tending to favour screening all four nanoCOFs.

The use of SDS was much lower when resources were not provided (4/20 vs 14/20). This appeared to be linked to the discovery of two papers. In 50% of batches, a paper was identified that states that SDS addition is strongly detrimental to H_2 production in an analogous photocatalyst system (*Small* **2024**, *20*, 2406236). Also, 50% of the time (25% of batches overall) a paper was identified stating the benefits of PVP in a photocatalyst set up (*Materials Letters* **2021**, *298*, 130025). Neither of these papers was used in the experiments with resources.

Taken together, these observations highlight that the LLM can find and prioritise relevant literature, but much less effectively than when it is provided up front as a resource (25% versus 100%).

7 Evolution of hypotheses

Figure 2c in the main paper tracks how the LLM's reasoning themes evolved across the full Phase 3b2 campaign (33 batches, 89 named hypotheses). This section documents the clustering methodology behind that analysis.

7.1 Clustering methodology

Each hypothesis was one named-rationale pair produced by the LLM, with a confidence label (high / medium / low). A single batch typically contained three to six hypotheses. We clustered only the 89 hypotheses produced over the full run of Phase 3b2 batches 1–33.

The clustering pipeline had four steps:

1. Text representation. Each hypothesis was concatenated as name + " - " + rationale and encoded with sentence-transformers/all-mpnet-base-v2 into a 768-dimensional vector, then L2-normalised.

2. Density clustering. HDBSCAN was run in the full 768-d embedding space. We tuned the hyperparameters cluster_selection_method (eom, leaf), min_cluster_size, min_samples, and a soft-reassignment threshold over a grid, and scored each candidate with a composite of embedding-space silhouette, agreement with a 2-D UMAP layout (silhouette and local kNN purity), TF-IDF keyword separation across clusters, and balance penalties against a single dominant cluster, excessive fragmentation, or residual noise.

3. Soft reassignment. With the chosen configuration, HDBSCAN-flagged noise points were reassigned to the nearest cluster centroid if their cosine similarity to that centroid exceeded the tuned threshold; otherwise, they remain labelled as noise. This rescued semantically plausible but locally sparse hypotheses without collapsing the density partition.

7.2 Themes and migration

The Phase 3b2 corpus partitioned into seven themes (Table S16). Theme labels follow the same c1-c7 ordering as in Figure 2c, with c5 the dominant exploitative theme around the [nanoCOF + Pt + AA + SDS] optimum.

Theme	N hypotheses	Clustered hypothesis
C1	13	nanoCOFn + oxide heterojunctions under AA + Pt
C2	7	nanoCOFn-Pt-ascorbate-SDS heterojunction optimum and local neighbourhood
C3	8	Pt-loading sweep for nanoCOFn + AA
C4	9	Robust ranking of the four nanoCOFs at fixed AA and Pt
C5	30	Replicated nanoCOFn-Pt-SDS-ascorbate optimum
C6	11	SDS loading sensitivity at fixed nanoCOFn, Pt, and AA
C7	11	Tuning redox environment with pH and mixed sacrificial donors

Table S16. Phase 3b2 hypothesis themes (89 hypotheses across 33 batches).

The full cluster contents are detailed in https://anonymous.4open.science/r/llm-vs-chemist-closed-loop-catalysis-experiments/

8 LLM vs human team comparisons

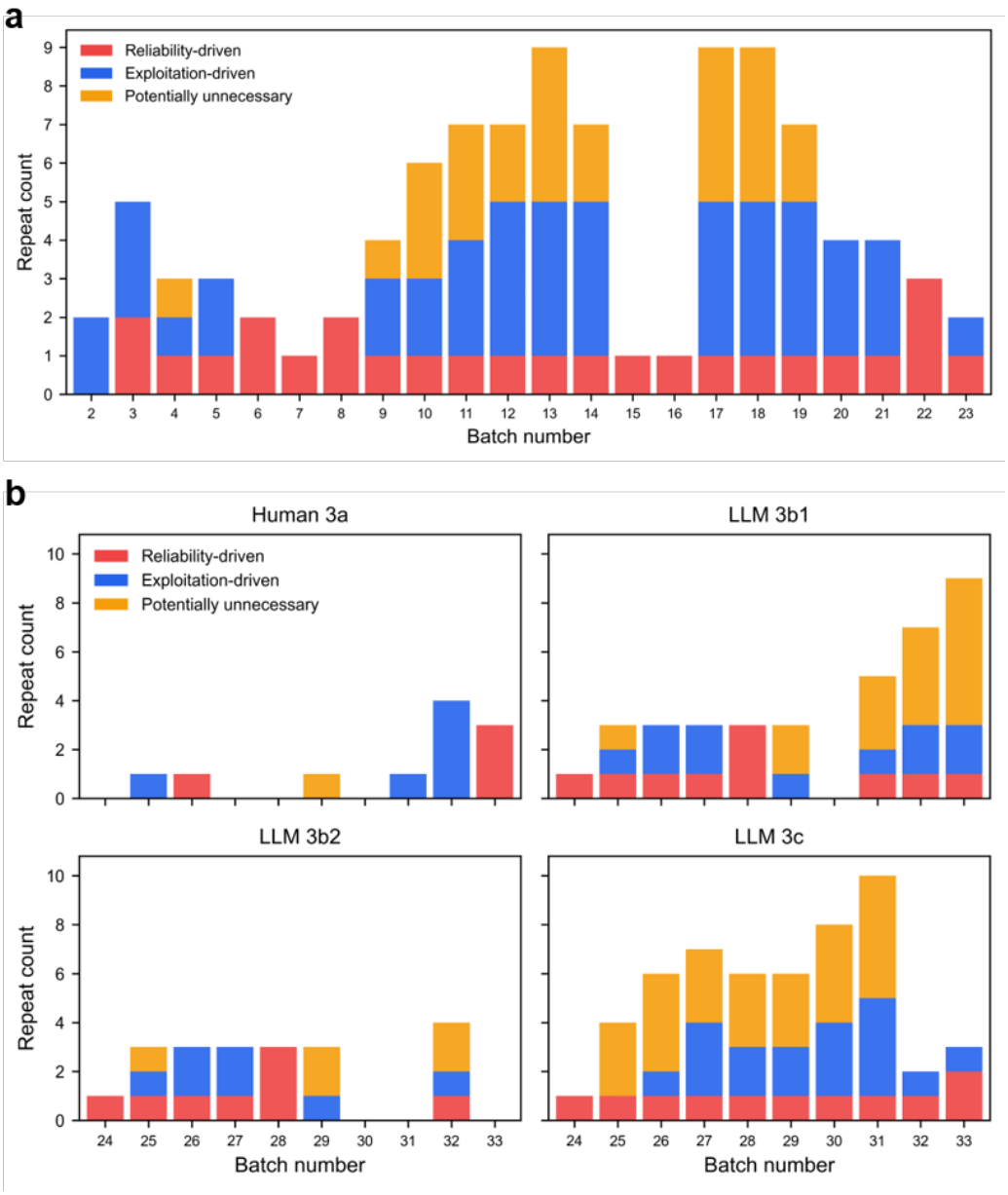

Figure S12. Phase-specific repeat counts and classification by batch. **a**. Shared LLM1 and LLM2 phases. **b**. Head-to-head Phase 3a (human) and Phase 3b1, 3b2, 3c (LLM) optimisations. Each repeated formulation was assigned using the previous-batch status as follows. Reliability-driven: repeats associated with failed/missing/invalid or unstable prior evidence, including high noise (cumulative noise in the top 20%, quantile = 0.8), high uncertainty (top 20%, quantile = 0.8), or replicate disagreement. Exploitation-driven: stable repeats of prior top performers, where top performer was defined as rank in top-5 or objective percentile in the top 20% (percentile ≥ 0.80). Potentially unnecessary: stable repeats of low/mid performers (not meeting top-5 and not in the top-20% percentile rule). Stacked bars show per-transition repeat counts by driver within each path.

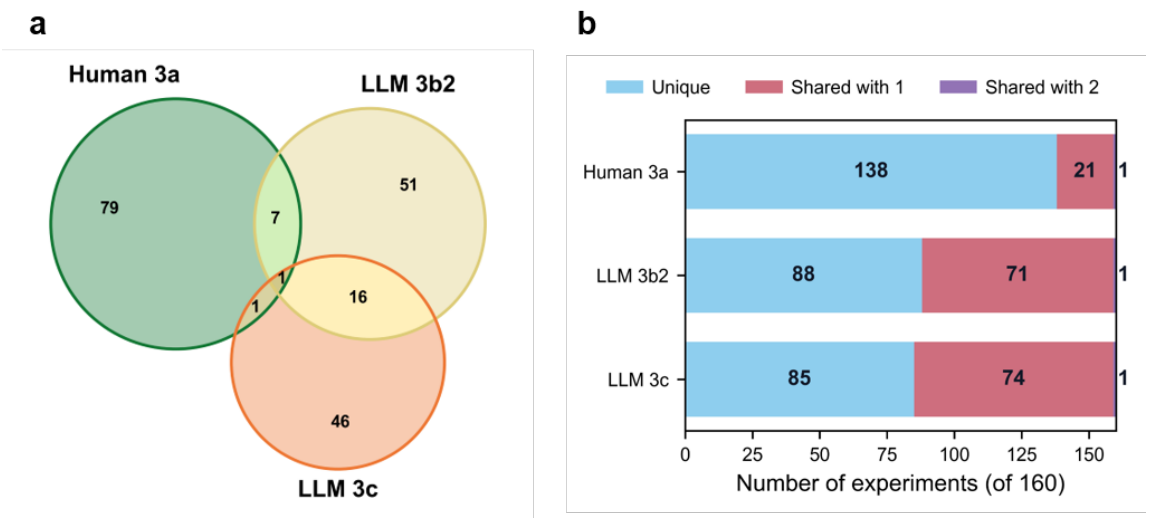

Figure S13. Unique formulations in Phase 3 by LLM and human team. **a**, Overlap of unique formulations by the human team and the LLM, ignoring repeats. A repeat formulation is one that has the same components, all at the same values, unlike a categorical formulation that has the same non-zero components irrespective of values. Circles proportional to number. In total, the human team explored 88 unique specific formulations in Phase 3a while the LLM explored 75 and 64, respectively, in Phase 3b2 and Phase 3c. **b**, Repeat aware comparison between human and LLM search strategies in Phase 3. An equivalent analysis of categorical formulations is available in Figure3 in the main text.

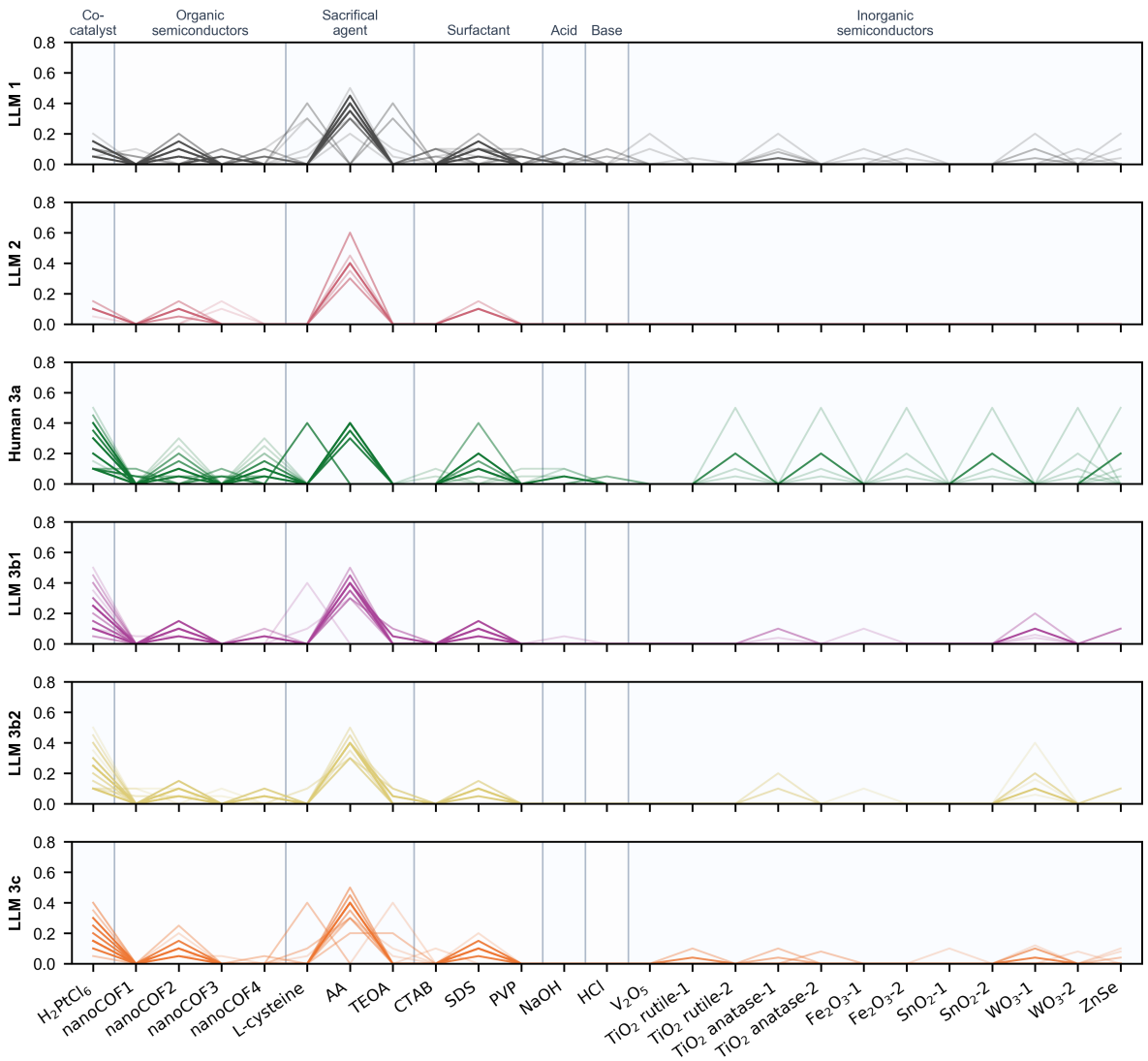

Figure S14. Parallel-coordinate map of unique formulation compositions by experimental phase. Panels correspond to Phases LLM1 (Phase 1, shared), LLM2 (Phase 2, shared), Human 3a, and LLM 3b1, 3b2, 3c. Each line is a distinct formulation profile with individual concentrations normalised to [0,1]. Features are ordered and banded by chemical family (organic semiconductors, sacrificial agents, etc.) to highlight how different branches sampled different regions of the shared formulation space.

9 Human team hypotheses

General logic for picking the experiments for the hypotheses

- Picked the smaller particle size for the inorganics.
- For batches 1–6, most samples were tested in triplicate to help isolate experimental noise.

Batches 1–6 Hypotheses

(Name, Rationale)

- **H1**: [nanoCOF1, nanoCOF3, nanoCOF4] + AA + Pt + SDS, It would be useful to also test the effects of SDS with the other three nanoCOFs to see whether their performance improves, as for nanoCOF2. **H1 Samples:** will test nanoCOF1, 3 & 4 under the current best conditions – 3 nanoCOFs tested in triplicate **(9 samples)**
- **H2**: nanoCOF2 + AA + Pt + SDS + [NaOH, HCl], Test the effects of pH on the best performing formulation; it is possible that sodium ascorbate (AA + modest NaOH) might function as a more effective sacrificial donor (https://pmc.ncbi.nlm.nih.gov/articles/PMC12814356); we also note some evidence from the experiments that HCl addition can be destructive (*e.g.*, experiments 45-48). **H2 Samples:** will test best conditions for nanoCOF2 + 0.25 mL of HCl/NaOH, 2 conditions, tested in triplicate **(6 samples)**
- **H3**: nanoCOF2 + AA + Pt + higher SDS, The highest SDS addition that has been explored so far is 1 mL (samples 135 and 304); higher SDS addition has not been tested. **H3 Samples:** will test nanoCOF2 with 1 and 2 mL of SDS (0.5 mL in best formulation), 2 conditions, tested in triplicate **(6 samples)**
- **H4**: NanoCOF + cysteine + NaOH + Pt, We found previously that cysteine plus a base was a good sacrificial donor combination for an organic polymer semiconductor (P10) and this hypothesis tests whether this learning is transferable to nanoCOFs. **H4 Samples:** Test cysteine as the sacrificial agent instead of ascorbic acid – keeping volume the same at 2 mL in the presence of 0.25 mL NaOH. Will test with and without SDS, 2 conditions, tested in triplicate **(6 samples)**
- **H5**: nanoCOF2 + AA + Pt + SDS, The concentration of AA used in the solution is possibly saturated, therefore try doubling the concentration of all other components. **H5 Samples:** maximise concentration of best formulation (1 mL Pt, 1 mL nanoCOF2, 1 mL SDS), 1 condition in triplicate **(3 samples)**
- **H6**: Explore combinations of [nanoCOF1, nanoCOF2, nanoCOF3, nanoCOF4] + AA + Pt + SDS, Investigate whether the nanoCOFs can form heterojunctions with each other or reach a synergistic combination via some other mechanism. **H6 Samples:** will test all combinations of 2 nanoCOFs using the current best formulation but instead of 0.5 mL of one nanoCOF, it will be 0.25 mL of each nanoCOF. 6 combinations (1/2, 1/3,1/4, 2/3, 2/4, 3/4) ran in triplicate **(18 samples)**
- **H7**: nanoCOF2 + [WO3, TiO2, ZnSe] + AA + Pt + SDS, WO_3, TiO_2 and ZnSe addition correlated with higher hydrogen production under certain conditions without surfactant (*e.g.*, a small increase in HER with ZnSe addition in sample 31 compared to sample 33); it would therefore be useful to study these inorganic additives with nanoCOF2 in the presence of SDS to see whether surfactant addition magnifies this effect.

- **H8**: [nanoCOF1, nanoCOF2, nanoCOF3, nanoCOF4]+ [Fe_2O_3_sample_1, Fe_2O_3_sample_2, SnO_2_sample1, SnO_2_sample2, TiO_2_anatase_sample2, TiO_2_rutile_sample_1, TiO_2_rutile_sample_2, WO_3_sample_2] + AA + Pt + SDS, The eight inorganic additives Fe_2O_3_sample_1, Fe_2O_3_sample_2, SnO_2_sample1, SnO_2_sample2, TiO_2_anatase_sample2, TiO_2_rutile_sample_1, TiO_2_rutile_sample_2, WO_3_sample_2 have not been tested as possible heterojunction partners with nanoCOFs.
- **H9**: [High inorganic solids, high SDS] + nanoCOF2 + AA + Pt, The maximum inorganic solid addition explored so far is 2 mg (samples174, 176, 237 and 239), which is much lower than the 10 mg upper limit; this hypothesis tests the regime of higher inorganic solid addition to try to 'force' Z-scheme formation; it is unclear whether light "shading" occurs at 2 mg addition (*e.g.*, experiments 174 and 176 gave moderate HERs). Hypotheses 7, 8 and 9 are addressed by the following samples: **H7/8/9 Samples:** will test the best nanoCOF2 formulation with inorganic addition. 6 inorganics WO_3, TiO_2 (anatase and rutile), ZnSe, Fe_2O_3 and SnO_2 at the following amounts 0.5, 1, 2 and 5 mg (5 mg samples specifically addresses hypothesis **9**). When there was a choice of inorganic particle size (all inorganics except ZnSe), the smaller particle size was selected **(24 samples)**
- **H10**: Inorganic light absorbers + sacrificial agent + Pt + (optionally) surfactant, most formulations explored so far have involved nanoCOFs; this hypothesis explores formulations where an inorganic solid is the sole light absorber in combination with Pt and a suitable sacrificial donor and (optionally) a surfactant. **H10 Samples:** will test TiO_2 (anatase and rutile), ZnSe and SnO_2 as they have the correct band gap to perform sacrificial HER on their own. Chose 2 mg as the amount to test in this round. When there was a choice of inorganic particle size (true for all inorganics except ZnSe) the smaller particle size was selected. Tested with and without SDS, 4 inorganics, tested under 2 conditions (with/without SDS) in triplicate **(24 samples)**

Batch 7 Hypotheses

- **H11**: Explore combinations of [nanoCOF2 and nanoCOF4] + AA + Pt + SDS under higher concentrations, This hypothesis combines the best results from hypotheses **5** and **6** and investigates whether a mixture of nanoCOF2 and nanoCOF4 outperforms a formulation containing only nanoCOF2 or nanoCOF4, using a higher formulation concentration. **H11 Samples:** Focus on exploring different ratios of nanoCOF2/4 (3:1, 1:1 and 1:3) using the higher concentration formulation – 1.0 mL nanoCOF + 2.0 mL AA + 1.0 mL Pt + 1.0 mL SDS **(4 samples)**
- **H12**: Varying nanoCOF4 amount, Investigate nanoCOF4 loading around current optimal loading of 0.5 mL. **H12 Samples:** Test nanoCOF4 + 2.0 mL AA + 0.5 mL Pt + 0.5 mL SDS formulation with 0.25 and 0.75 mL nanoCOF4 **(2 samples)**
- **H13**: nanoCOF4 + AA + higher Pt + SDS, Test the effects of increasing Pt loading using the current best formulation (Pt/nanoCOF4 ratio: 7.5%, 15%). **H13 Samples:** Test 0.5 mL nanoCOF4 + 2.0 mL AA + 0.5 mL SDS formulation with 0.75 and 1.50 mL Pt **(2 samples)**
- **H14**: nanoCOF4 + AA + Pt + SDS + NaOH, This hypothesis builds on hypothesis **2** and will test the effects of basic conditions with nanoCOF4. **H14 Samples:** Testing the addition of 0.25 mL NaOH the current best formulation for nanoCOF4: 0.5 mL nanoCOF + 2.0 mL AA + 0.5 mL Pt + 0.5 mL SDS **(1 sample)**

- **H15**: Explore combinations of [nanoCOF2 and nanoCOF4] + AA + Pt + SDS under nanoCOF loading > 1 mL, This hypothesis builds on the best-performing combination from hypothesis **6** and investigates a higher nanoCOF loading. **H15 Samples:** Total volume of nanoCOF added to each sample is 1.5 mL. The concentration of AA used in the solution is possibly saturated so is reduced to 1.5 mL to allow the nanoCOF loading to be increased. 7 different volume ratios of nanoCOF2 and nanoCOF4 were tested: 1:0, 5:1, 2:1, 1:1, 1:2, 1:5 and 0:1 **(7 samples)**

Batch 8 Hypotheses

- **H16**: An extension of hypothesis **5** and **13**, where AA is still saturated, replacing AA with more nanoCOF4 and a higher Pt loading of 15%. **H16 Sample**: 0.75 mL nanoCOF4 + 1.5 mL AA + 2.25 mL Pt + 0.5 mL SDS **(1 sample)**
- **H17**: nanoCOF4 + AA + higher Pt + SDS, This hypothesis builds on hypothesis **13** and tests the effects of increasing Pt loading using the current best formulation (Pt/nanoCOF4 ratio: 15, 17.5 and 20%). **H17** Samples: Test 0.5 mL nanoCOF4 + 2.0 mL AA + 0.5 mL SDS formulation with 1.5, 1.75 and 2.0 mL Pt **(3 samples)**
- **H18**: Replace SDS with NaOH + surfactant mixture, Test whether the performance enhancement provided by SDS is due to increasing the sodium concentration by adding NaOH and with surfactants PVP and CTAB. **H18** Samples: Test 0.5 mL nanoCOF4 + 2.0 mL AA + 1.5 mL Pt formulation with (0.25 and 0.5 mL) PVP or CTAB + (0.25 and 0.5 mL) NaOH **(5 samples)**
- **H19**: Increase Pt loading for mixture of nanoCOF2 and nanoCOF4. This hypothesis takes the best result from hypotheses **6** and **15** and combines it with hypothesis **17** to investigate higher Pt loadings on nanoCOF2/4 mixtures. **H19 Samples**: 0.25 mL nanoCOF2 + 0.25 mL nanoCOF4 with Pt/nanoCOF ratios of 15, 17.5 and 20% **(3 samples)**
- **H20**: HER scales with product of Pt concentration and nanoCOF concentration at high Pt. AA is still saturated, reduce AA to allow larger amounts of Pt and nanoCOFs to be investigated. **H20 Samples:** nanoCOF2/4 mixture 1.25 mL with 1.5 mL Pt, and single nanoCOF 2 and 4 formulations at 0.75 ml nanoCOF with 2 mL Pt. For all samples AA 1.75 mL + SDS 0.5 mL **(4 samples)**

Batch 9 Hypotheses

- **H21**: nanoCOF4 + AA + higher Pt + SDS, Test the effects of further increasing Pt loading using the current best formulation (Pt/nanoCOF4 ratio: 22.5%, 25%). This hypothesis builds on hypothesis **17**. **H21 Samples:** Test 0.5 mL nanoCOF4 + 0.5 mL SDS and for Sample 1: 1.75 mL AA + 2.25 mL Pt, Sample 2: 1.5 mL AA + 2.5 mL Pt **(2 samples)**
- **H22**: Fine tuning SDS amount (0.25, 0.50 and 0.75 mL) on best performing formulation so far. **H22 Samples:** 0.5 mL nanoCOF4 + 1.75 mL AA + 2.0 mL Pt + SDS, SDS amounts tested are 0.25, 0.5 and 0.75 mL **(3 samples)**
- **H23**: Retest top 3 points from batch 8 (hypothesis 17) to see the spread/noise. **H23 Samples:** 0.5 mL nanoCOF4 + 2.0 mL AA + 0.5 mL SDS, Pt amounts tested are 1.5, 1.75 and 2.0 mL **(6 samples)**
- **H24**: Perturb around nanoCOF2 and nanoCOF4 mixtures where the total nanoCOF volume is 1 mL. This hypothesis builds on hypothesis **11** and investigates nanoCOF2/4 mixtures with higher Pt loading (1.5 mL up from 1.0 mL) and lower SDS loading (0.75 mL down from 1.0 mL) than in hypothesis **11**. **H24 Samples:** 1.0 mL [nanoCOF2 +

nanoCOF4] + 1.75 mL AA + 1.50 mL Pt + 0.75 mL SDS, nanoCOF2/4 volume ratios: 3:1, 1:1, 1:3 **(3 samples)**

- **H25:** Perturb Pt and AA loadings for the nanoCOF2 and nanoCOF4 1:1 mixtures where the total nanoCOF volume is 0.5 mL. This hypothesis builds on hypothesis **6** investigates fine tuning the Pt and AA amounts at higher Pt loadings, nanoCOF/Pt ratio ≥ 20%. **H25 Samples:** 0.25 mL nanoCOF2 + 0.25 mL nanoCOF4 + AA + Pt + 0.5 mL SDS. Sample 1 = 1.75 mL AA + 2.25 mL Pt, Sample 2 = 2.0 mL AA + 2.0 mL Pt. **(2 samples)**

Batch 10 Hypotheses

- **H26**: Retest top 3 points from batches 8 and 9 to see the spread/noise. **H26 Samples:** Pt loadings of 1.5 mL (6 samples), 1.75 mL (5 samples) and 2.0 mL (5 samples) + nanoCOF4 0.5 mL + AA 2.0 mL + SDS 0.5 mL **(16 samples)**

9.1 Human reasoning time tracking

Activity	Time (h)	People Involved
23-3-2026 Initial review of data and generate initial 5 hypothesis	2	QJY, ZX
24-3-2026 Review Opsight data	1.5	QJY, ZX, CB
24-3-2026 Hypothesis discussion in Andy's office	4	QJY, ZX, CB, AIC
24-3-2026 Going through data and looking through optimisation logs	2.5	AIC
25-3-2026 Lunchtime discussion	1.5	QJY, ZX, AIC
25-3-2026 Reviewing data used to create PPT summary circulated previously	1.5	AIC
26-3-2026 Generate Batches from hypotheses	3	QJY, ZX, CB
26-3-2026 Reviewing Batches from hypotheses	2	QJY, ZX, CB, AIC
27-3-2026 Reviewing data from Batch 1/2 + reviewing data from Batch 3/4 + reviewing data from batch 5	0.6	QJY, ZX
27-3-2026 Reviewing human Batches 1 and 2	0.5	AIC
28-3-2026 Reviewing / thinking about Batches 3/4	0.75	AIC
29-3-2026 Reviewing human Batches 1-6	1.5	AIC, CB
29-3-2026 Reviewing human Batches 1-6, coming up with new hypothesis, generate new points for Batch 7	2	QJY, ZX
30-03-2026 Reviewing all hypotheses / expts	0.5	AIC, QJY, ZX
30-03-2026 Reviewing all hypotheses / expts	0.3	CB
06-04-2026 Reviewing Batch 7	0.3	QJY
07-04-2026 Reviewing Batch 7	0.25	ZX
07-04-2026 Reviewing Batch 7 and thinking about next expts	1	AIC, CB
07-04-2026 Reviewing Batch 8 hypotheses	1	AIC
07-04-2026 Discussing Batch 7 and coming up with experiments	2	QJY, ZX, CB
08-04-2026 Discussing Batch 8 and coming up with hypothesis + experiments	0.75	QJY, ZX, CB
09-04-2026 Reviewing all batches so far and suggesting additional points for Batch 9	1	AIC
09-04-2026 Discussing points for Batch 9	2	AIC, QJY, ZX, CB
09-04-2026 Slack huddle to discuss Batch 9 points	1.5	AIC, QJY, ZX
09-04-2026 Discussing Batch 10 points	0.2	QJY, ZX
10-04-2026 Reviewing and thinking about Batch 10 suggestions	0.5	AIC
TOTAL	34.65	person hours

Table S17. Human reasoning time log. This includes all individual and group time spent in producing the hypotheses and points for batches 1–10 (320 experiments) in Phase 3a.

10 Additional off-line experiments

10.1 Repeatability tests for best formulations found in human Phase 3a and LLM Phases 3b and 3c

The most active catalytic formulations found in Phase 3a, 3b and 3c, respectively, were not repeated an equal number of times (Table 1, main text). To further test the repeatability of each 'best' formulation composition, the hydrogen evolution rate was measured 15 times for each composition. The results are shown in Figure S17 and Table S18, below.

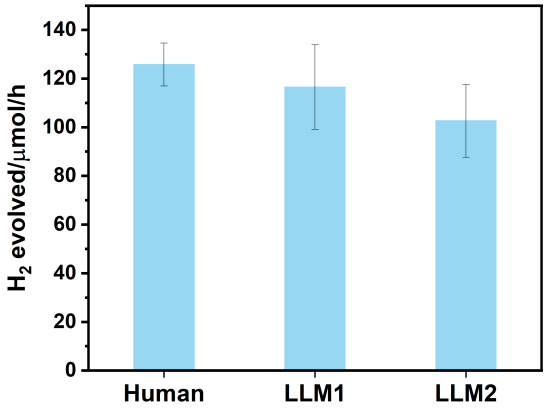

Figure S15. Verification of catalytic activity for best formulations from Human Phase 3a, LLM Phase 3b2 (LLM1) and LLM Phase 3bc (LLM2). 15 repeats for each of the 3 formulations.

Phase	Average HER (μmol/h)	S.D. (μmol/h)
Human – Phase 3a	126	8
LLM1 – Phase 3b2	117	17
LLM2 – Phase 3c	103	15

Table S18. Average HER value and standard deviations of the best three formulations found, averaged over 15 repeats. Note that the same best formulation was found in Phase 3b1 and Phase 3b2.

10.2 Surfactant effects

Prompted by the observed phase behaviour changes (Figure 4a, main text) and the increase in catalytic activity associated with SDS (B in Figure 1d, main text), we also explored the effect of other anionic surfactants on the best formulation found in Phase 1. The results are shown in Figure S18 and Table S19; two other anionic surfactants were also found to provide the same benefits.

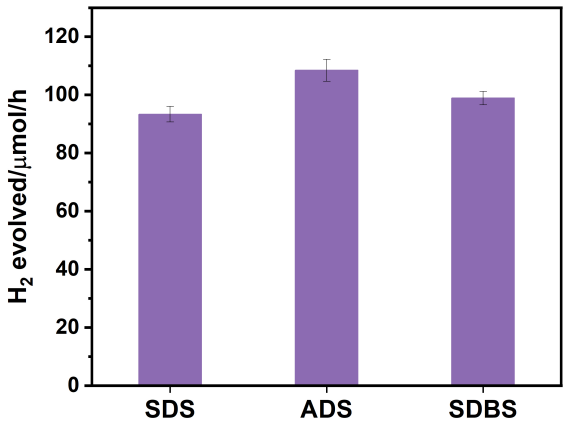

Figure S16. Hydrogen evolution rate for nanoCOF2 with the addition of different anionic surfactants. Formulation: 0.5 mL nanoCOF2 (0.0828 mg/mL), 0.5 mL H_2PtCl_6 (0.0072 mg/mL), 2 mL AA (0.5M) and 0.5 mL surfactant (0.01M). SDS: sodium dodecyl sulfate; ADS: ammonium lauryl sulfate; SDBS: sodium dodecylbenzenesulfonate. Five repeats of each formulation (averages shown).

Surfactant	Average HER (μmol/h)	S.D. (μmol/h)
SDS	93	3
ADS	108	4
SDBS	99	2

Table S19. Average HER values and standard deviations of the best formulation found in Phase1 with different anionic surfactants (averages of 5 repeats each).

10.3 Particle size change for nanoCOF2 after the addition of SDS

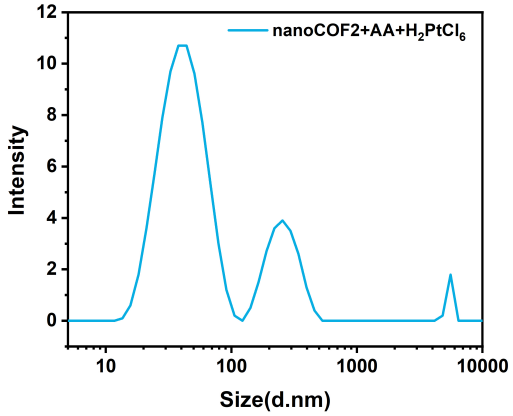

Figure S17. DLS results of nanoCOF2 with AA and H_2PtCl_6 (0.5 mL nanoCOF2 (0.0828 mg/mL), 0.5 mL H_2PtCl_6 (0.0072 mg/mL), 2 mL AA (0.5M), 2mL H_2O) – no SDS added.

Peak	Size(d.nm)	Intensity%	St Dev(d.nm)
Peak1	43	78	17
Peak2	262	20	71
Peak3	5490	2	219

Table S20. DLS peak distribution for nanoCOF2 with AA and H_2PtCl_6

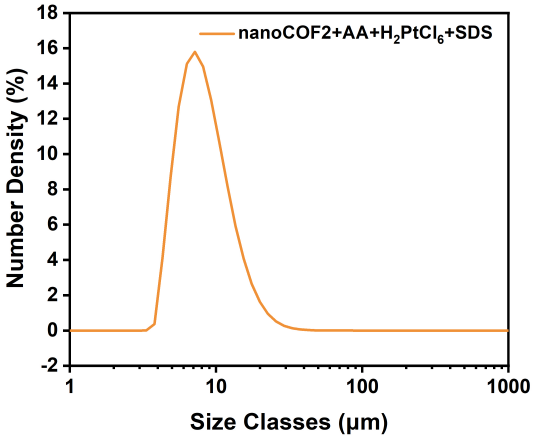

Figure S18. Mastersizer results for nanoCOF2 with AA, H_2PtCl_6 and SDS (0.5 mL nanoCOF2 (0.0828 mg/mL), 0.5 mL H_2PtCl_6 (0.0072 mg/mL), 2 mL AA (0.5M) 0.5 mL SDS (0.01M), 1.5mL H_2O) – SDS added.

Sample Name	Dx (10) (µm)	Dx (50) (µm)	Dx (90) (µm)	Mode (µm)
nano COF2 + SDS	5.1	7.8	13.9	7.1

Table S21. Particle diameters of nanoCOF2 with AA, H_2PtCl_6 and SDS from Mastersizer measurements. Dx (10, 50, 90): The diameter which accounts for 10, 50 and 90% of the total volume of the sample; respectively. Mode: The most common diameter of the sample.

11 Error logs - LLM errors and workflow errors

11.1 LLM errors

Below is a summary of the key LLM errors encountered during the experimental campaign. Errors are classified as either:

- Major – they affected future experiments
- Moderate – the error was corrected in the next batch

Phase 0: Initialisation

Major Error – LLM incorrectly calculates the Pt loading from the literature, as it assumes all of the mass of the 0.0072 mg/mL H_2PtCl_6 stock solution is Pt. This is incorrect and causes the LLM to underestimate the Pt loading for all samples.
Reason for error: LLM fails to distinguish the difference between Pt and H_2PtCl_6, even though the chemical formula was clearly provided. As discussed above in section 6.3.2, this occurred in most repeat initialisations where such calculations were attempted (that is, the experimental run was not an outlier).

Phase 1

Moderate Error - Batch 1, H1: points match rationale, though not a direct comparison between nanoCOFs 1-4 as the, nanoCOF, Pt and AA amounts are not fixed, particularly for nanoCOF4:

- 0.25 mL nanoCOF**1** + 0.5 mL H_2PtCl_6 + 2 mL AA
- 0.25 mL nanoCOF**2** + 0.5 mL H_2PtCl_6 + 2 mL AA
- 0.25 mL nanoCOF**3** + 0.5 mL H_2PtCl_6 + **1.5 mL** AA
- **0.5 mL** nanoCOF**4** + **0.25 mL** H_2PtCl_6 + **1 mL** AA

Classified as moderate as subsequent hypotheses in Batch 2 and Batch 3 dis perform a direct comparison between the different nanoCOFs (0.5 mL H_2PtCl_6 + 2 mL AA + 0.25 mL nanoCOF)
Reason for error: Hard to identify why the LLM made this inconsistent comparison. There were far more instances where the LLM *was* able to test a hypothesis by varying one parameter at a time while keeping others parameters constant, so this error was a minority case.

Moderate Error - Batch 9 summary: LLM states we are only up to batch 8. Other analysis seems appropriate. Model states that it has not tested AA and SDS perturbations, but these were trials occurred in Batch 8. Quite a lot of duplication between batches 9 and 10.
Reason for error: Failure of data analysis tool (code interpreter)

Moderate Error - Batch 13: Batch 13 is identical to batch 12
Reason for error: Failure of data analysis tool (code interpreter)

Major Error - Batch 15 summary: Analysis reasonable, except for the statement: “Nearby perturbations within ±0.25 mL in Pt and SDS show a relatively flat optimum plateau”. In fact, inclusion of 0.25 mL SDS consistently reduced performance to 4–5 umol, which is almost a 50% reduction in performance compared to 0.5 mL SDS. This is not a “flat optimum plateau”.

LLM takes until Batch 19 summary to identify that 0.25 mL SDs is “strongly detrimental” (with respect to higher SDS concentrations).
Reason for error: Possible that the LLM is hallucinating without evidence to back it up

Phase 2 (Instruction added to explain illumination time is reduced from 60 to 15 min)

The summary indicates that LLM is aware of the change in illumination time and will retest previous high performing samples from batches 1–20. No major/moderate errors observed.

All batches in this phase (batches 21-23) were reasonable, with no obvious failures. We noted that batch 23, H3, used 3 mL AA, which had never been used before that point, but the LLM provided no explicit reasoning for this (4 repeats of formulation *SDS 0.5 mL, nanoCOF 0.5 mL, AA **3** mL, H_2PtCl_6 0. 5 mL*)

Phase 3

Phase 3b1 – Budget aware LLM

LLM was aware that batch 24 was the first optimisation-focused batch at 15 min with a finite, defined experimental budget.

Batch 24 summary seemed appropriate; model correctly learns that higher Pt loading is beneficial. Model also shows that it is budget aware: “..positioning later batches (25-33) to explore…”

Major Error - Batch 29 summary: defaulted to best formulation at 60 min with 0.5 mL Pt, while best at 15 min used 1.25 mL Pt. This memory failure effected all remaining batches, with the Batch 31 summary also incorrectly stating there where three batches left, when in fact there were only two.
Reason for error: Combined effect of context-window limitations and unsuccessful execution of the data analysis tool (code interpreter).

Phase 3b2 – New prompt for batches 30–33

Summary correctly identified difference between 60 and 15 min illumination times (Phase 1 versus Phase 2 & 3). It was also budget aware, stating only batches 30b2-33b2 remain.

Moderate Error - Batch 30b2: Points in this batch did not match the stated rationale as they used the Pt volumes from the best Phase 1 60-min illumination time formulations (0.5 mL H_2PtCl_6) rather than the 1.25 mL Pt concentrations that were found to be superior 15 min illumination times. This memory failure appeared to resolve itself at the end of the batch since the Batch 30b2 summary correctly identified a 1.25 mL Pt volume in the best formulation for 15 min illumination.
Reason for error: Context-window limitations and unsuccessful execution of the data analysis tool (code interpreter)

Moderate Error - Batch 32b2 summary: Had the same memory failure as batch 30b2. All points in Batch 33b2 used the Pt volumes from the best 60 min illumination time formulations (0.5 mL H_2PtCl_6) instead of the 1.25 mL Pt.

Reason for error: Context-window limitations and unsuccessful execution of the data analysis tool (code interpreter)

Phase 3c – New prompt / human monitoring (batches 24c-33c)

Initialisation of batch 24 correctly identified [0.75 mL Pt + 0.5 mL nanoCOF2 + 2 mL AA + 0.5 mL SDS] as the best formulation so far at 15 min illumination and in the Batch 30c summary, the LLM correctly identified that there were only 3 batches left. Analysis failures occurred at batches 28c, 31c, 32c and 33c with human intervention reinitialising the batch to overcome this issue (Extended Data Figure 8).

11.2 Workflow errors

We characterised execution errors by comparing the actual dispensed values recorded in the experimental logs (noted as lab_data_logs_x.csv in https://anonymous.4open.science/r/llm-vs-chemist-closed-loop-catalysis-experiments/, where x refers to the batch number) with the target formulations originally proposed by the LLM. This was important because the LLM reasoned about the formulation that was physically executed in the lab, together with the proposed (intended) formulation. The model also explicitly compared target and actual amounts of substances dispensed to assess the significance of execution errors, to estimate the noise level, and to determine whether repeat experiments were needed. The following is an example taken from the optimisation log of Phase 1:

“Based on batch 1, the system appears low in measurement noise but shows modest execution noise (up to ~0.03 mL in liquids, ~0.1–0.2 mg in most solids, with one larger deviation in an Fe_2O_3 solid that did not produce H_2). There are no true replicated conditions yet, so we will explicitly include replicates this round to quantify variance. Ground-truth comes from lab_data_logs; suggested vs actual settings are very close in the active region, so we can treat the commanded grid as a good design basis while attributing outcomes to actuals.”

The dispense error summary per component is shown in

Table S22 for liquids and

Table S23 for solids across 912 experiment campaign (LLM 1, 2, 3b shared, 3b1, 3b2, and 3c, as well as Human 3a). The error for each dispensing operation was defined as the absolute difference between the proposed dispense amount and the actual dispensed amount. The interpretation of this error was based on the dispensing resolution (discretisation step) of each component type: 0.25 mL for liquid components and 0.20 mg for solid components.

For liquid dispenses, an absolute error below 0.125 mL is less than half of the dispensing resolution. Such deviations can be interpreted as noisy realisations of the intended formulation, since the executed value remains closer to the target formulation than to a neighbouring dispense level. These errors were therefore considered relatively minor. Errors in the range 0.125–0.25 mL were large enough that the executed value became difficult to distinguish from a neighbouring formulation, but they still fell within one dispensing resolution

of the intended value. These errors were therefore considered moderate. Errors of 0.25 mL or greater correspond to at least one full dispensing resolution, meaning the executed sample may be better interpreted as a distinct formulation rather than as execution noise around the intended one. These errors were therefore considered as major.

For solid dispenses, the same interpretation was applied using the solid dispensing resolution of 0.20 mg. An absolute error below 0.10 mg remained within half of the dispensing resolution (minor). Errors between 0.10 mg and less than 0.20 mg were close to a neighbouring dispense level but remained within one dispensing resolution of the intended value and were therefore considered moderate. Errors of 0.20 mg or greater correspond to at least one full solid dispensing resolution (major).

Component	Total number of dispense	Number of dispense error (e, mL) in each range		
		$0 \leq e < 0.125$	$0.125 \leq e < 0.25$	$e \geq 0.125$
H_2PtCl_6	910	903	0	7
ascorbic acid	894	894	0	0
water	863	862	0	1
nanoCOF2	790	785	1	4
SDS	779	777	2	0
nanoCOF4	129	129	0	0
TEOA	37	36	0	1
nanoCOF3	32	32	0	0
NaOH	22	22	0	0
L-Cysteine	20	20	0	0
nanoCOF1	20	20	0	0
PVP	12	12	0	0
CTAB	11	11	0	0
HCl	7	7	0	0
Total	4526	4510	3	13
Percent	100%	99.6%	0.1%	0.3%

Table S22. Per-component dispense error summary for liquids based on absolute execution error across 912 experiments. For each component, “Total number of dispense” counts the number of experiments in which that component was dispensed. Dispense errors are reported as the number of occurrences within each absolute error range: $0 \leq e < 0.125$ mL, $0.125 \leq e < 0.25$ mL, and $e \geq 0.125$ mL, where e is the absolute difference between the proposed and actual dispensed amounts.

Component	Total number of dispenses	Number of dispensing errors (e, mg) in each range		
		$0 \leq e < 0.10$	$0.10 \leq e < 0.2$	$e \geq 0.20$
WO_3-1	39	34	2	3
TiO_2 anatase-1	22	8	4	10
ZnSe	21	12	6	3
TiO_2 anatase-2	12	1	6	5
SnO_2-2	10	8	0	2
TiO_2 rutile-1	10	1	2	7
TiO_2 rutile-2	10	5	2	3
WO_3-2	7	2	1	4
Fe_2O_3-2	6	2	1	3
Fe_2O_3-1	3	3	0	0
V_2O_5	2	0	2	0
SnO_2-1	1	1	0	0
Total	143	77	26	40
Percent	100%	54%	18%	28%

Table S23. Per-component dispense error summary for solids based on absolute execution error across 912 experiments. For each component, “Total number of dispense” counts the number of experiments in which that component was dispensed. Dispense errors are reported as the number of occurrences within each absolute error range: $0 \leq e < 0.10$ mg, $0.10 \leq e < 0.2$ mg, and $e \geq 0.20$ mg, where e is the absolute difference between the proposed and actual dispensed amounts.

The total dispense error summary, including both liquids and solids, are shown in

Table S24. Overall, the majority of dispense errors, combining liquid and solid dispenses, were minor, accounting for 98% of all dispensing operations. Moderate and major errors accounted for 0.6% and 1.1% of dispenses, respectively. Liquid dispenses were almost exclusively minor, with 99.6% falling below 0.125 mL, while moderate and major errors accounted for only 0.1% and 0.3%, respectively. By contrast, solid dispenses showed a higher proportion of larger errors: 54% were minor, 18% were moderate, and 28% were major.

State	Total number of dispenses	Minor	Moderate	Major
Liquid	4526	4510	3	13
Solid	143	77	26	40
Total	4669	4587	29	53
Percent	100%	98%	0.6%	1.1%

Table S24. Total dispense error summary. Dispense error was calculated as the absolute difference between the proposed and actual dispensed amounts. Errors were considered minor, moderate, and major at absolute differences of <0.125 mL, 0.125–<0.25 mL, and ≥0.25 mL for liquids, and <0.10 mg, 0.10–<0.20 mg, and ≥0.20 mg for solids, respectively.

12 Appendices

All additional information can be found in the following repository:
https://anonymous.4open.science/r/llm-vs-chemist-closed-loop-catalysis-experiments/

Repository Structure

```
│   │   └── Optimisation log and summary raw from Opsight/
│   ├── 4.4 Phase 3c (independent LLM run, batches 24c-33c)/
│   │   ├── Lab data logs/
│   │   └── Optimisation log and summary raw from Opsight/
│   ├── 4.5 Phase 3b2 stochasticity/
│   │   └── B24_res/
│   │       └── logs/
│   └── 4.6 Phase 3 LLM reasoning costs/
├── 5. Comparison of performance and strategy for human and LLM reasoning/
│   ├── 5.1 Best-formulation verification/
│   └── 5.2 Reasoning time and cost comparison/
├── 6. Methods/
│   ├── 6.1 Datasheet/
│   └── 6.2 LLM-driven closed-loop optimisation/
├── 7. Annotated logs/
│   ├── Phase 1_annotated.pdf
│   ├── Phase 2_annotated.pdf
│   ├── Phase 3b1_annotated.pdf
│   ├── phase3b2_annotated.pdf
│   └── phase3c_annotated.pdf
├── 8. Workflow video/
│   └── video.mp4
├── LICENSE
└── README.md
```

13 Supplementary References

1 Khurram, R. *et al.* Synthesis and characterization of an alpha-Fe_2O_3/ZnTe heterostructure for photocatalytic degradation of Congo red, methyl orange and methylene blue. *RSC Adv.* **10**, 44997-45007 (2020). https://doi.org:10.1039/d0ra06866g

2 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

3 Zhang, J., Zhou, P., Liu, J. & Yu, J. New understanding of the difference of photocatalytic activity among anatase, rutile and brookite TiO_2. *Phys. Chem. Chem. Phys.* **16**, 20382-20386 (2014). https://doi.org:10.1039/c4cp02201g

4 Sharma, Y. C., Ansaria, P., Sharma, R., Mathur, D. & Dard, R. A. Bandgap tuning of optical and electrical properties of zinc selenide. *Chalcog. Lett.* **18**, 183-189 (2021). https://doi.org:10.15251/cl.2021.184.183

5 Li, Y. *et al.* Realizing a SnO_2-based ultraviolet light-emitting diode via breaking the dipole-forbidden rule. *NPG Asia Mater.* **4**, 30 (2012). https://doi.org:10.1038/am.2012.56

6 Syrek, K. *et al.* Band gap engineering of tungsten oxide-based nanomaterials. *J. Photochem. and Photobiol. C: Photochem. Rev.* **62**, 100681 (2025). https://doi.org:10.1016/j.jphotochemrev.2024.100681

7 Mounkachi, O. *et al.* Band-gap engineering of SnO_2. *Solar Energy Mater. Solar Cells* **148**, 34-38 (2016). https://doi.org:10.1016/j.solmat.2015.09.062

8 Lee, S., Wang, Q. & Chakrapani, V. Electronic vs optical bandgap of stoichiometric V_2O_5: Effects of surface transfer doping and electron accumulation layer. *J. Appl. Phys.* **136**, 205101 (2024). https://doi.org:10.1063/5.0233493

9 She, X. *et al.* High efficiency photocatalytic water splitting using 2D α-Fe_2O_3/g-C_3N_4 Z-scheme catalysts. *Adv. Energy Mater.* **7**, 1700025 (2017). https://doi.org:10.1002/aenm.201700025

10 Sadeghzadeh-Attar, A. Enhanced photocatalytic hydrogen evolution by novel Nb-doped SnO_2/V_2O_5 heteronanostructures under visible light with simultaneous basic red 46 dye degradation. *J. Asian Ceram. Soc.* **8**, 662-676 (2020). https://doi.org:10.1080/21870764.2020.1773621

11 Jiang, L., Gao, X., Chen, S., Ashok, J. & Kawi, S. Oxygen-Deficient WO_3/TiO_2/CC Nanorod arrays for visible-light photocatalytic degradation of methylene blue. *Catalysts* **11**, 1349 (2021). https://doi.org:10.3390/catal11111349

12 Ray, S. & Tarafder, K. Investigation of CdSe and ZnSe as potential back surface field layers for CdTe-based solar cells: a study from first principles calculations. *Adv. Theory Sim.* **6**, 2200718 (2023). https://doi.org:10.1002/adts.202200718

13 Harris, D. C. in *Quantitative Chemical Analysis* AP21 (W. H. Freeman and Company, 2011).

14 Li, Y. *et al.* Experimental and theoretical studies of the standard oxidation potentials of triethylamine and triethanolamine. *Canad. J. Chem.* **103**, 480-492 (2025). https://doi.org:10.1139/cjc-2024-0272

15 Zhang, H., Lau, W. C. & Hollenberg, P. F. Formation of the thiol conjugates and active metabolite of clopidogrel by human liver microsomes. *Mol. Pharmacol.* **82**, 302-309 (2012). https://doi.org:10.1124/mol.112.079061

16 The inorganic structure images shown in Figure 1a were reproduced using the Materials Project resource: Horton, M.K., Huck, P., Yang, R.X. *et al.* Accelerated data-driven materials science with the Materials Project. *Nat. Mater.* **24**, 1522–1532 (2025). https://doi.org/10.1038/s41563-025-02272-0