Research Square

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

A cold-start self-evolving multi-agent intelligence for closed-loop prediction from materials synthesis to performance

Kunlei Zhu

zhyq9188@126.com

Qufu Normal University https://orcid.org/0009-0006-7178-2996

Tao Li

Qufu Normal University

Xueyang Cheng

Qufu Normal University

Shaoqi Mao

Qufu Normal University

Yingxue Ren

Qufu Normal University

Shimiao Sun

Qufu Normal University

Physical Sciences - Article

Keywords:

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11771581/v1

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

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

A cold-start self-evolving multi-agent intelligence for closed-loop prediction from materials synthesis to performance

Abstract

Large language models (LLMs) are increasingly used to propose materials and plan experiments[1–3], but whether they can predict previously unreported syntheses and the structures and performance of the resulting products before experimental validation remains unclear. Here we test this capability using a multi-agent framework of general-purpose, unfine-tuned LLMs that enables cross-disciplinary collaboration, with an AI scientist coordinating chemistry, materials science and electrochemistry agents to performance closed-loop predictions of the reaction process, crystal phase and battery performance of TiO_2 anodes synthesized by a previously unreported rapid Joule-heating route. The agents predicted 64 conditions blind in eight rounds and, after each round, rewrote natural-language Skills from the released measurements. A task-matched team predicted the reaction process and phase in 91.7% and 82.8% accuracy, with errors concentrated near sharp transitions, whereas quantitative predictions of rate-capability and long-cycle performance achieved accuracies of 59.2% (72.3% for anatase-rich products) and 68.1%, within ±30% of experimental values. Revised Skills raised accuracy on identical conditions in most matched tests, and rewriting Skills without new measurements did not. A complete prediction chain cost ~US$0.27 per condition. These results indicate that LLM predictions could replace exploratory trials under conditions far away from sharp transitions, pointing to a prediction-first research paradigm.

Introduction

Traditional materials discovery has long relied on experimental trial-and-error, with synthesis, characterization and performance testing forming a resource-intensive and time-consuming iterative cycle[4–6]. Each iteration demands specialized equipment, skilled labor, and significant financial investment, while success is often unpredictable and dependent on prior domain expertise[7]. As societal and economic demands for advanced materials in areas such as energy storage[8] and catalysis[9] continue to grow, this manual approach has become a major bottleneck, limiting the pace of innovation and scalability[4,10].

The advent of *ab initio* computational methods, particularly density functional theory (DFT)[11,12], brought a paradigm shift by enabling the prediction of crystal structures, stability, and electronic properties without prior experiments[13]. High-throughput screening campaigns

established open databases such as the Materials Project[14], OQMD[15], and AFLOW[16], cataloguing hundreds of thousands of crystal phases and guiding experimentalists toward promising candidates[13]. Landmark achievements include the design of hydrogen-evolution catalysts[9] and Li-ion cathode materials[17], demonstrating that computational materials science can narrow the search space and accelerate discovery. However, these methods remain resource-intensive. For example, DFT calculations are computationally expensive, and their accuracy depends on the choice of functional and approximations[18,19]. More importantly, the knowledge barrier remains substantial because *ab initio* approaches require deep expertise in electronic-structure theory and access to supercomputing infrastructure[20], making them inaccessible to many experimental researchers and limiting their applicability in synthesis-process design, where empirical process variables (e.g. temperature, time, solvent, and precursor chemistry) are often poorly captured by *ab initio* models[21,22].

Machine learning (ML) emerged as a complementary approach, capable of learning structure–property relationships from empirical datasets and generalizing across chemical spaces with lower computational cost than *ab initio* methods[23,24]. Graph-based neural networks such as crystal graph convolutional neural networks (CGCNN)[25] and MEGNet[26] achieved notable success in property prediction, while autonomous laboratories closed the loop by integrating robotic synthesis with ML-guided optimization[10,27,28]. Nevertheless, even these models present significant barriers. Traditional ML requires substantial expertise in data preprocessing, feature engineering, and model architecture selection[24]. Small and noisy datasets, which are common in experimental materials science, lead to overfitting or unreliable predictions[29,30]. Furthermore, synthesis routes are often context-dependent and not straightforwardly encoded as input features, making accurate yield and process prediction difficult for models trained on curated data alone[22]. For chemistry and materials researchers without formal training in computer science or data science, these methods remain out of reach, perpetuating the gap between computational prediction and laboratory practice.

Large language models (LLMs)[31–33] have recently introduced a qualitatively different mode of scientific reasoning. Built on natural language as their primary interface, they offer an intuitive and broadly accessible means of interacting with scientific knowledge and tools, enabling researchers to propose hypotheses, interpret experimental data, and generate synthesis recipes without specialized programming or explicit reprogramming[34]. LLMs have been applied to extract knowledge from materials databases[35], predict synthesis conditions for metal–organic

frameworks[36,37], make mathematical discoveries[38], and autonomously design experimental protocols[1,39]. In materials science, recent works have demonstrated LLM-guided structure generation[40], property prediction[41,42], and synthesis-feasibility assessment[43]. However, most LLM-based systems focus on discrete subtasks, such as materials screening[41] or extraction of synthesis recipes from literature[34–36], rather than spanning the entire discovery cycle. Critically, no prior work has demonstrated a closed-loop, end-to-end prediction system that autonomously maps from synthesis parameters to structure to performance, and then iteratively refines its internal reasoning with experimental feedback, without task-specific model retraining.

Here, we present a cold-start, self-evolving multi-agent framework based on general-purpose LLMs, including Gemini, GPT, DeepSeek, Claude, GLM and Grok. The framework autonomously predicts closed-loop, end-to-end synthesis-to-performance outcomes across all three stages (reaction kinetics, crystal phase formation and electrochemical behavior) and iteratively improves through experimental feedback without additional model-weight training. Starting from zero-shot predictions with role-specific prompts and no pre-existing domain Skills, the system distributes 64 predefined synthesis conditions across eight rounds of eight samples each, using a randomized stratified design. We validate the predictive capability of the framework using TiO_2 synthesized by rapid Joule heating as a Li-ion battery anode, a synthesis route not previously reported for TiO_2. After each prediction round, experimental results drive the autonomous revision of natural-language Skills, which are distilled reasoning documents that encode process-specific knowledge. A chief AI scientist coordinates four specialized agents (a chemistry agent for reaction kinetics processes, a materials agent for crystal phases, and two electrochemistry agents for rate capability and long-cycle performance), each refining its Skill independently. We evaluate the framework across three independent runs and find task-dependent performance: the Hybrid team achieves 91.7% and 82.8% pooled accuracy for two-class reaction-process classification and five-class crystal-phase classification, respectively, while numerical rate-capability and long-cycle performance predictions meet a ±30% relative-error tolerance in 59.2% (rising to 72.3% for anatase-rich products) and 68.1% of evaluated fields, respectively. A complete eight-round workflow of the Hybrid team uses an average of 22.96 million tokens and an estimated model-call cost of US$17.13, equivalent to only approximately US$0.27 per condition for a complete prediction chain. These gains come from eight new measurements per round, including conditions that failed to trigger self-propagating synthesis, rather than from a large training database, showing that a small set of samples that includes negative results can raise prediction accuracy. Our framework allows

drop-in model updates, offering a scalable pathway toward generalized, interpretable and cost-efficient materials discovery beyond domain-expert reliance. It also points to a prediction-first research paradigm in which LLM agents could replace part of exploratory experimentation.

Results

Cold-start, blind full-chain prediction

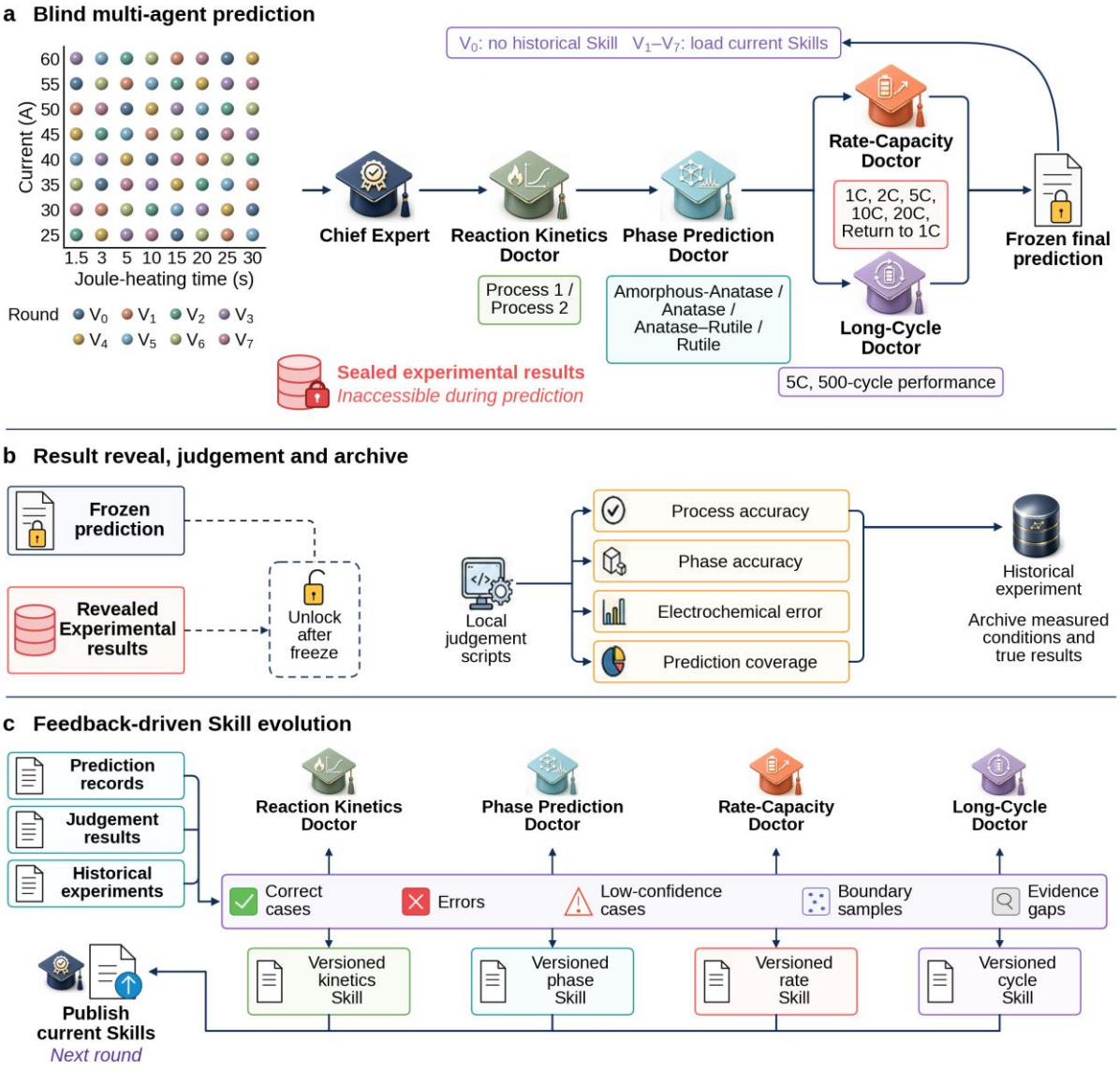


Fig. 1 | Cold-start, blind prediction of the full synthesis–performance chain for Joule-heated TiO_2. a, Blind multi-agent prediction. Left, the 8 × 8 prediction matrix of constant current and Joule-heating time; colours mark the eight prediction rounds (V_0–V_7) produced by randomized Latin-hypercube-like stratified sampling, each containing every current level and every heating time once. The AI scientist (Chief Expert) dispatches each round's conditions to the chemistry agent (Reaction Kinetics Doctor; Process 1, Joule heating alone; Process 2, Joule heating with self-propagating high-temperature synthesis), the materials agent (Phase Prediction Doctor) and two electrochemistry agents (Rate-Capacity Doctor: 1C, 2C, 5C, 10C, 20C and return to 1C; Long-Cycle Doctor: 5C, 500 cycles). Downstream agents receive the conditions and the upstream predictions. Predictions are frozen while the round's experimental results remain sealed. V_0 starts without Skills; later rounds load the latest published Skills. **b,** After freezing, the results are unlocked, scored by local scripts for process accuracy, phase accuracy, electrochemical error and prediction coverage, and archived. **c,** Feedback-driven Skill evolution. Using prediction records, judgements and archived experiments, each agent analyses correct cases, errors, low-confidence cases, boundary samples and evidence gaps, writes a versioned Skill and publishes it for the next round. Model weights are never updated. Solid arrows, task or data flow; red locks, inaccessible results; curved arrow, Skill evolution between rounds.

The framework predicts what a Joule-heating condition will produce, from the reaction pathway to battery performance, before any measurement of that condition is available (Fig. 1). We defined the prediction task as an 8 × 8 matrix of constant currents (25–60 A in 5-A steps) and heating times (1.5, 3, 5, 10, 15, 20, 25 and 30 s), giving 64 predefined Joule-heating conditions (Fig. 1a, left). Randomized Latin-hypercube-like stratified sampling divided the matrix into eight prediction rounds, V_0–V_7, of eight conditions each, so that every round contained each current level and each heating time exactly once and all 64 combinations were covered without repetition (Supplementary Table S1). For every condition, the agents received only the synthesis inputs (current, heating time, voltage limit and precursor mass in Supplementary Fig. S1a–d) and had to predict the complete outcome chain. The 64 conditions were experimentally validated by heating a butoxy- and hydroxy-containing TiO_2 precursor heated on carbon felt in air using a direct-current power supply, thereby evaluating the prediction performance of the agents through a synthetic route not previously reported for TiO_2. Because experimental procedures, including materials characterization and electrochemical testing, are time-consuming, we conducted these experiments in advance. The resulting experimental data were the withheld from the prediction agents and kept sealed until the predictions were finalized, after which they were used as independent reference to evaluate the frozen predictions and guide subsequent knowledge revision.

Within each round, an AI scientist agent (Chief Expert in Fig. 1a) dispatched the eight conditions to four specialist agents in causal order. A chemistry agent (Reaction Kinetics Doctor) first predicted the reaction process, distinguishing Joule heating alone (Process 1) from Joule heating that triggers self-propagating high-temperature synthesis (Process 2). A materials agent (Phase Prediction Doctor) then used the conditions and the predicted process to assign the product to one of five phase classes, from amorphous through anatase to rutile. Finally, two electrochemistry agents used the conditions and both upstream predictions to predict, in parallel, the rate capability (capacities at 1C, 2C, 5C, 10C and 20C and on returning to 1C, 1C=170 mAh g^{-1}; Rate-Capacity Doctor) and the 5C, 500-cycle performance (Coulombic efficiency, discharge capacity and capacity retention; Long-Cycle Doctor). Because the products obtained from Process 1 retained precursors, they were not tested electrochemically. Accordingly, the electrochemistry agents predicted performance only for conditions assigned to Process 2. The AI scientist then assembled and froze the final prediction for all eight conditions while that round's experimental results remained sealed. In V_0 the agents had no Skills and no history, so these first predictions relied only on the task description and the models' pretrained knowledge;

from V1 onwards, each agent loaded the latest published version of its Skill.

Once the predictions were frozen, the round's results were unlocked and scored by local judgement scripts rather than by the agents (Fig. 1b). Process and phase predictions were correct when they matched the experimental class. Numerical predictions were correct when they fell within ±30% of the measured value. Fields that were not measured or not predicted were excluded from accuracy denominators and recorded as prediction coverage. The scripts then archived the round's conditions, frozen predictions, judgements and measured results. The measured results comprised the process class and phase, processing quantities such as power, Joule heat and temperature, phase fractions and sizes, and rate and long-cycle performance (Supplementary Figs. S1–S3).

Each specialist agent then revised its own Skill (Fig. 1c). Using the prediction records, the judgements and the archived results of all previous rounds, it sorted the evidence into correct cases, errors, low-confidence cases, boundary samples and evidence gaps, and wrote a new Skill version. Each version is a structured text that states the agent's current best relationships between conditions and outcomes, their range of validity and boundary cases, the reasons for keeping, revising or deleting earlier rules, and hypotheses that still lack evidence. The AI scientist published the four updated Skills as the current versions, and the next round started in fresh agent sessions in which these Skills were the only knowledge carried over from earlier rounds. Model weights were never updated, so all knowledge carried between rounds resides in readable, versioned text. The prompts of all five agents and every published Skill version are provided in Supplementary Note 1 and Supplementary Data 1.

Accurate classes, approximate capacities

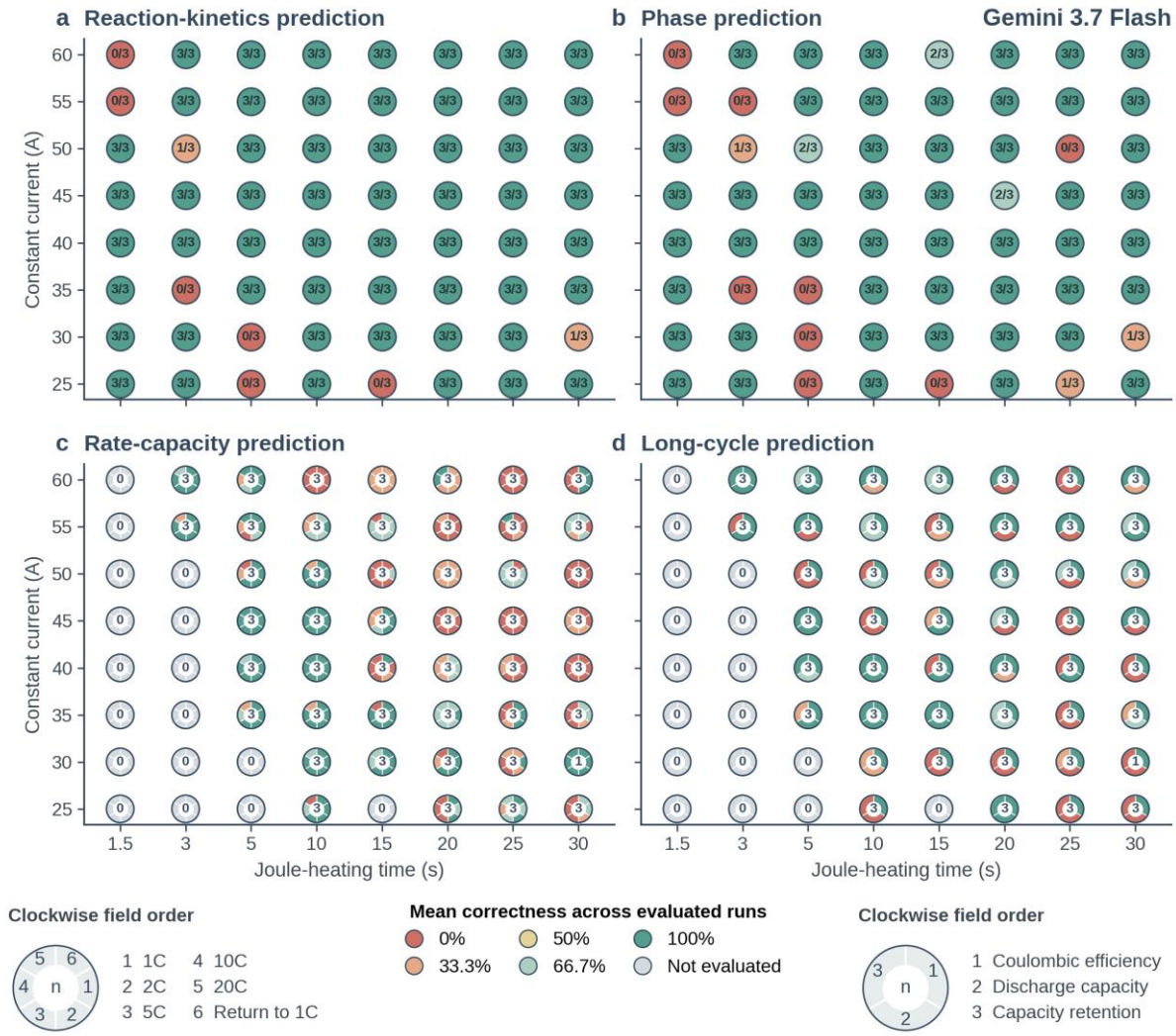


Fig. 2 | Condition-resolved accuracy of blind full-chain predictions. Three independent runs of the complete framework with Gemini 3.7 Flash on the 64 conditions. Heating times are plotted at evenly spaced positions. **a,b,** Reaction-process (**a**) and phase (**b**) classification; labels show the number of correct judgements out of three runs. **c,** Rate capability, with six sectors for the capacities at 1C, 2C, 5C, 10C, 20C and on returning to 1C. **d,** Long-cycle performance, with three sectors for Coulombic efficiency, discharge capacity and capacity retention after 500 cycles at 5C. Sector colours show the fraction of evaluated runs whose prediction lay within ±30% of the measured value; central labels give the number of evaluated runs. Grey sectors with a central zero mark fields without an evaluable prediction–measurement pair, either because the product was not tested electrochemically (Process 1) or because no performance prediction was made after an upstream Process 1 assignment; they do not denote zero error. Unevaluated fields are excluded from accuracy denominators. Panels **a,b** contain 192 judgements each, and **c,d** contain 834 and 417 evaluated fields. The runs are repeated executions on the same conditions, not additional physical samples; run-level accuracies are given in Supplementary Fig. S31.

Using Gemini 3.7 Flash for all five agents, we ran the complete eight-round framework three times on the same 64 conditions (Fig. 2). Each run is an independent execution of the same workflow, not an additional physical sample. A complete run used on average 16.99 million tokens and US$16.39 in estimated model-call costs, or about US$0.26 per condition for the full prediction chain (Supplementary Fig. S31). Pooled across the three runs, reaction-process and phase classifications were correct in 88.5% (170/192) and 81.2% (156/192) of judgements, and

53.1% (443/834) of rate-capability and 64.3% (268/417) of long-cycle predictions fell within ±30% of the measured values (Fig. 2). Fig. 2 pools the three runs and the accuracy of each run is shown in Supplementary Fig. S31. Run-to-run variation was small for the classifications (87.5–89.1% for the reaction process and 81.2% for phase in all three runs) and larger for rate capability (50.4–54.7%) and long-cycle performance (61.0–66.7%).

The classification errors were concentrated rather than scattered (Fig. 2a,b). Half of the 22 process errors occurred in the cold-start round V_0. The remaining 11 occurred at four short-pulse conditions (25 A and 30 A for 5 s, 50 A for 3 s and 60 A for 1.5 s), each adjacent to the measured current–time boundary at which self-propagating synthesis begins. Process 2, which formed at 50 of the 64 conditions, was recognized in 92.7% (139/150) of judgements and Process 1 in 73.8% (31/42) (Extended Data Fig. 1a). For phase, the dominant anatase/rutile mixture was assigned correctly in 93.9% (107/114) of judgements, rutile in 83.3% (15/18) and amorphous/anatase, the product of every Process 1 condition, in 73.8% (31/42) (Extended Data Fig. 1b). Pure anatase was the weak point. It formed at only six conditions, five of which bordered conditions that did not trigger self-propagating synthesis (Supplementary Fig. S2), and it was recognized in 16.7% (3/18) of judgements, with 14 of its 15 misassignments assigned to a neighbouring mixed class and none to rutile. No product was fully amorphous, so that class could not be tested.

Performance predictions were approximate rather than exact. Within ±30%, capacity accuracy fell steadily with rate, from 64.0% at 1C to 48.2% at 5C and 36.7% at 20C, and recovered to 64.0% on returning to 1C (Extended Data Fig. 1c–h). To show how these results depend on the tolerance, we recomputed accuracy for tolerances from ±10% to ±50% (Supplementary Table S2). Pooled rate-capability accuracy rose from 22.9% at ±10% to 39.4% at ±20%, 65.7% at ±40% and 72.2% at ±50%, and accuracy decreased from 1C to 20C at every tolerance. The predictions therefore captured the magnitude and rate dependence of the capacity more reliably than its exact value. In the long-cycle task, every Coulombic-efficiency prediction met the criterion at all tolerances, but this endpoint did not discriminate between predictions: measured values lay between 99.1% and 100.1%, and no prediction deviated by more than 0.6%. Without it, pooled long-cycle accuracy was 46.4% at ±30% and ranged from 18.3% at ±10% to 65.8% at ±50%. Discharge capacity after 500 cycles was the hardest long-cycle endpoint (43.2% at ±30%; Extended Data Fig. 1i–k).

Much of the remaining numerical error reflected the materials rather than the predictions. The

six conditions at 50, 55 and 60 A for 25 and 30 s produced pure, coarse rutile (crystallite sizes of 159–243 nm; Supplementary Fig. S2), which stores little lithium at room temperature[44]. Their electrodes delivered 2.5–7.6 mAh g^{-1} at 1C and 0.2–1.0 mAh g^{-1} at 20C, one to two orders of magnitude less than anatase-rich products (Supplementary Figs. S3 and S37). At such capacities, the ±30% band narrows to ±0.06–0.3 mAh g^{-1} at 20C, comparable to the 0.1 mAh g^{-1} resolution of the reported values, and the 500-cycle records of these electrodes show Coulombic efficiencies of 58–144% in individual early cycles and capacity retentions of 106–148% that result from changes of less than 0.5 mAh g^{-1} (Supplementary Fig. S37). Prediction accuracy therefore followed the phase composition (Supplementary Fig. S38a,c). For products containing at least 80% anatase, 75.9% (305/402) of rate-capability predictions fell within ±30%, rising to 83.3% (45/54) for the six pure-anatase products, whereas the value was 26.8% (53/198) for products containing at least 80% rutile. The 500-cycle discharge capacity showed the same contrast (52.2% versus 21.2%). The low pooled accuracy of the performance predictions thus reflects, to a large extent, the near-zero and unstable capacities of rutile-rich electrodes rather than the predictive ability of the models.

Errors also propagated along the chain. When the chemistry agent assigned a Process 2 condition to Process 1, the electrochemistry agents, as designed, made no performance prediction for it. This removed 11 of the 150 possible prediction–measurement pairs per endpoint, all from four conditions (25 A for 15 s, 30 A for 30 s, 30 A for 5 s and 60 A for 1.5 s). Counting these omissions as errors would lower pooled ±30% accuracy to 49.2% for rate capability and 59.6% for long-cycle performance, so the accuracy of the first link sets the coverage of the last. Because each round contained different conditions, these pooled scores cannot show whether feedback improved the Skills. We therefore examined what the Skills recorded and tested successive versions on identical conditions.

Feedback-revised Skills raise accuracy

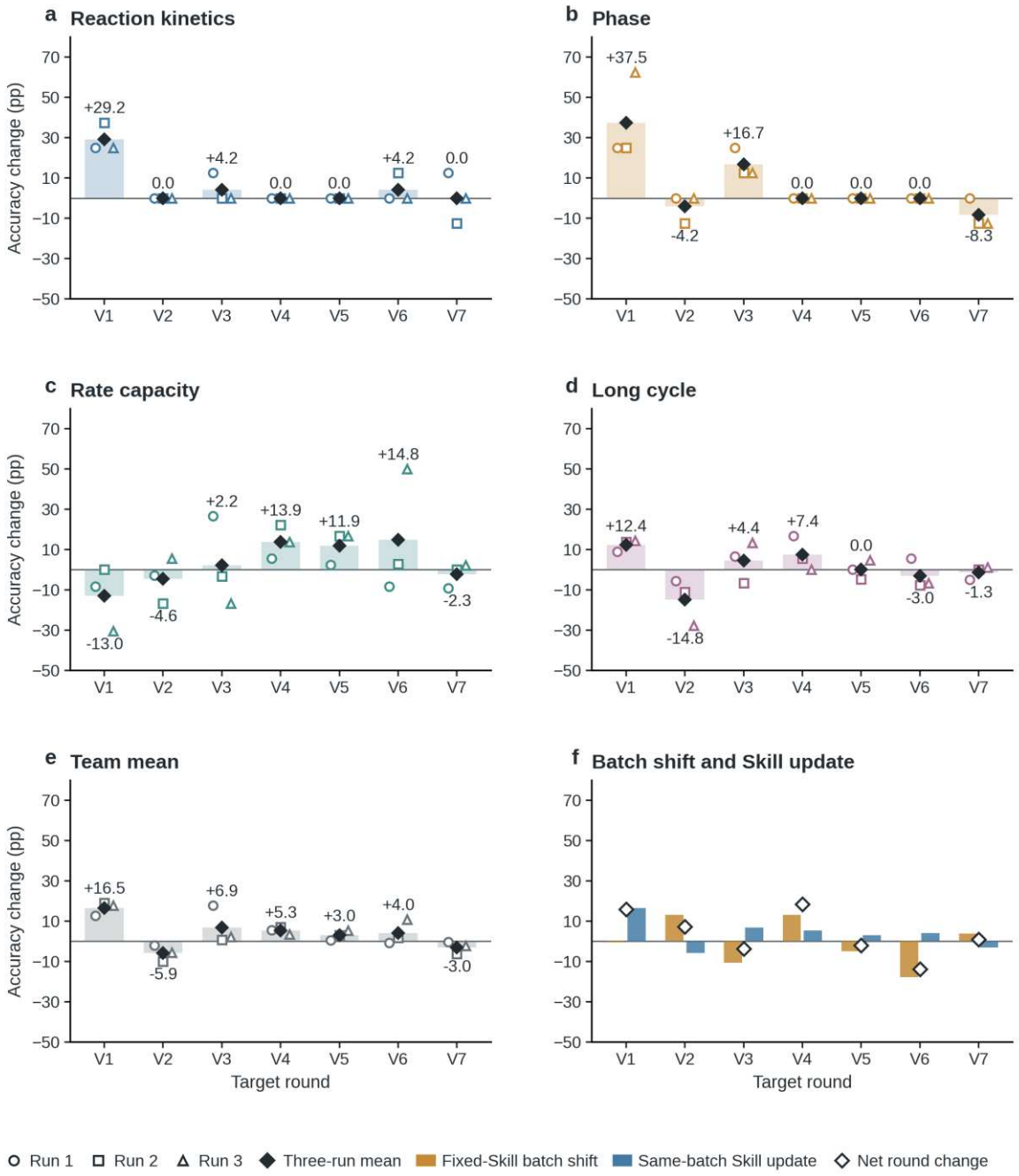


Fig. 3 | Matched-condition tests of Skill revision. a–d, Change in accuracy (newer minus older Skill) for reaction-process (**a**), phase (**b**), rate-capability (**c**) and long-cycle (**d**) predictions. For target rounds V_2–V_7, the two most recent Skills, both published before feedback on the target round, predict the same eight conditions, with only the corresponding agent's Skill replaced. Upstream predictions were not held fixed, so downstream changes include upstream propagation (Methods). V_1 compares the V_0 Skills with separate runs without Skills; its markers index separate runs, not paired histories. Open symbols, three dedicated comparison runs (distinct from the runs in Fig. 2); bars and filled diamonds, means. **e,** Team mean, weighting the four tasks equally. **f,** Decomposition of the round-to-round change into a fixed-Skill batch shift (older Skill on the target versus the preceding conditions), a same-batch Skill update and their sum. Accuracy is computed over measured fields that the corresponding version answered, so the two versions can have different denominators; numerical predictions are scored at ±30%. Changes are in percentage points (pp), calculated before rounding.

Experimental feedback guided the update of the Skills, with the rationale for each update traceable to the corresponding measurements (Extended Data Fig. 2 and Supplementary Data 1), rather than being determined by opaque black-box reasoning. Across the three runs, the four agents published 96 Skill versions. In the cold-start round, errors arose both for low-current, long pulses that triggered self-propagating synthesis and for short pulses that did not, and the first kinetics Skill accordingly made heating duration the leading criterion. Later versions constrained the process jointly by current and duration and, after V_7, placed the 3-s transition between 50 and 55 A, where the measurements show that self-propagating synthesis begins. The phase Skills developed time-dependent branches, widened the conditions assigned to pure rutile and extended an early window for pure anatase. The rate-capability Skills set separate capacity baselines for different phase regimes. For example, the 40-A, 30-s product, contained 88.2% rutile, delivered only 33.9 mAh g^{-1} at 1C and anchored a low-capacity regime. The long-cycle Skills treated Coulombic efficiency, capacity and retention as separate endpoints and identified an activation regime in which the 50-A, 10-s and 30-A, 30-s products showed capacity retention above 120% after 500 cycles. These statements are empirical working relationships with stated ranges of validity, not established mechanisms.

Skills revised from feedback improved predictions of identical conditions (Fig. 3). On the V_1 conditions, agents using the Skills written after V_0 were, as a team mean, 16.5 percentage points (pp) more accurate than separate runs without Skills (Fig. 3e), with gains of 29.2 pp for the reaction process and 37.5 pp for phase (Fig. 3a–b). For V_2–V_7, three dedicated comparison runs replaced one agent's Skill at a time with its previous version and scored both configurations on the same eight conditions. Upstream predictions were not held fixed and differed in 21 of 54 downstream comparisons, so these tests compare team configurations that include changes propagated from upstream agents (Methods). The newer Skills raised the team mean in four of six comparisons, by 3.0–6.9 pp, and lowered it at V_2 (−5.9 pp), mainly through the long-cycle Skill (−14.8 pp), and at V_7 (−3.0 pp), mainly through the phase Skill (−8.3 pp) (Fig. 3e). Gains were task-dependent (Fig. 3a–d). Reaction-process and phase Skills contributed mainly at V_1, and phase also at V_3 (+16.7 pp). Most of their later comparisons showed no change, whereas the rate-capability Skills raised accuracy by 11.9–14.8 pp at V_4–V_6. Accuracy in these tests is computed over the fields that each version answered, so a version can score lower while making more correct predictions. For example, at V_7, the newer rate-capability Skill answered 102 rather than 90 fields and made 58 rather than 52 correct predictions, yet its mean accuracy was 2.3 pp lower.

The same tests explain why round-level scores did not rise monotonically (Fig. 3f). With the Skills held fixed, moving to the next round's conditions changed team accuracy by up to about 18 pp, more than any same-condition Skill update after V_1. At V6, for example, the V_4 Skills scored 83.1% on the V_5 conditions but 65.3% on the V_6 conditions, and the V_5 Skills scored 69.3% on the same V_6 conditions. As a result, the 13.8-pp drop between rounds thus combined a −17.8-pp shift between the two sets of conditions with a +4.0-pp gain from the Skill update. Together, the Skill records and these matched tests demonstrate that experimental feedback rewrote the agents' working knowledge and, in most comparisons, made it more predictive.

Gains require Skills and new feedback

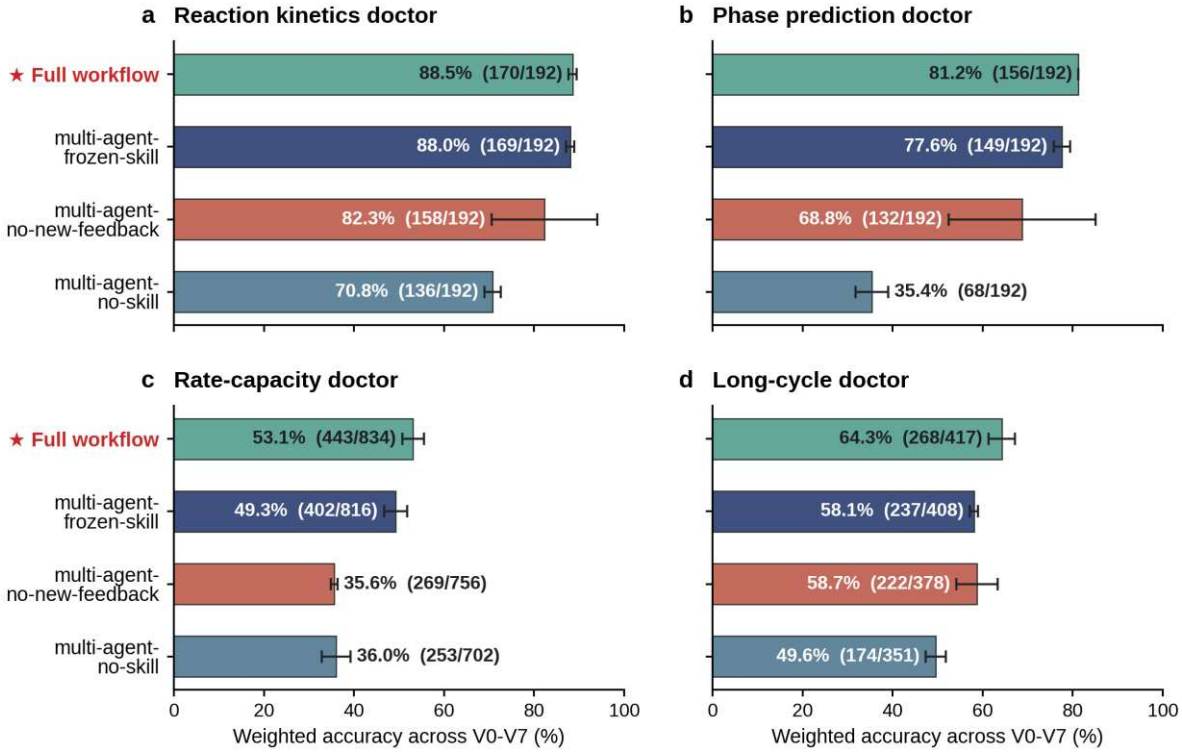


Fig. 4 | Accuracy gains require persistent Skills and new experimental feedback. a–d, Pooled accuracy of reaction-process (**a**), phase (**b**), rate-capability (**c**) and long-cycle (**d**) predictions with Gemini 3.7 Flash. The full workflow (star) is compared with three ablations: Skills written after V0 and never revised (frozen Skill); Skills revised in every round after a single disclosure of the V0 measurements and judgements, without any later feedback (no new feedback); and no Skills in any round (no Skill). Bars show correct predictions divided by evaluated fields over three runs and eight rounds, with counts given in the labels; numerical predictions are scored at ±30%. Error bars, sample standard deviation of the three run-level accuracies. Evaluated denominators of the performance tasks differ between workflows because unpredicted fields were excluded.

We next asked which parts of the loop produced these gains, comparing the full workflow with three ablations, each run three times on the same 64 conditions with Gemini 3.7 Flash (Fig. 4; Supplementary Figs. S25–S27). In the frozen-Skill workflow, the agents kept the Skills written after V_0 for all later rounds. In the no-new-feedback workflow, the agents received the V_0

measurements and judgements once and used them for the first Skill revision, as in the full workflow. They then continued to rewrite their Skills in every round from their own predictions alone, without any further measurements or judgements. In the no-Skill workflow, they predicted every round without Skills.

Removing Skills lowered accuracy in all four tasks, to 70.8% (136/192) for the reaction process, 35.4% (68/192) for phase, 36.0% (253/702) for rate capability and 49.6% (174/351) for long-cycle performance, compared with 88.5%, 81.2%, 53.1% and 64.3% for the full workflow (Fig. 4a–d). The largest loss, 45.8 pp for phase, indicates that knowledge carried between rounds mattered most for structural prediction. Freezing the V0 Skills retained the reaction-process accuracy (88.0%, 169/192) but lowered the phase (77.6%), rate-capability (49.3%) and long-cycle (58.1%) accuracies. Rewriting the Skills without new data performed worse than freezing them in three of the four tasks (82.3%, 68.8% and 35.6% for the reaction process, phase and rate capability), and for rate capability it fell to the no-Skill level. In the tested workflow, continued revision therefore improved predictions only when it was driven by new measurements, whereas without such measurements it eroded knowledge that the frozen Skills retained.

In addition, requesting the maximum reasoning setting for the full workflow did not improve all tasks together. Reaction-process results rose from 170/192 to 173/192, whereas phase, rate-capability and long-cycle results fell from 156/192, 443/834 and 268/417 to 152/192, 407/840 and 266/420 (Supplementary Fig. S28).

Model choice is task-dependent

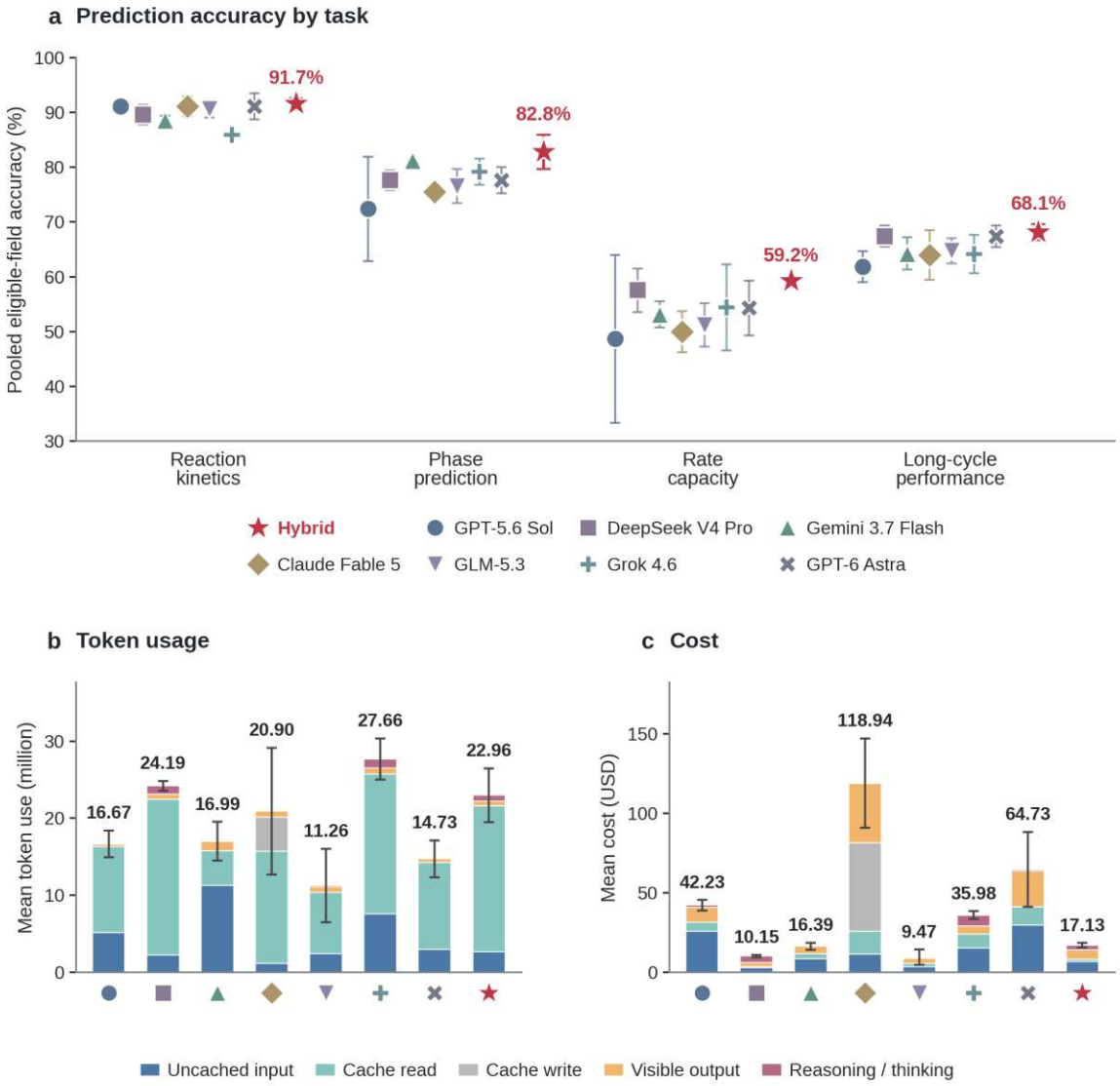


Fig. 5 | Task-dependent model performance and a task-matched Hybrid team. a, Pooled accuracy of reaction-process, phase, rate-capability and long-cycle predictions for seven single-model teams, in which one LLM serves all five agents, and for the Hybrid team (red stars), which uses GPT-6 Astra for the chemistry agent, Gemini 3.7 Flash for the materials agent and DeepSeek V4 Pro for the electrochemistry agents and the AI scientist. Symbols, pooled accuracy over three runs of V_0–V_7; error bars, sample standard deviation of run-level accuracies; labels give the Hybrid values. Numerical predictions are scored at ±30%. Hybrid counts are 176/192, 159/192, 487/822 and 280/411. The Hybrid results come from complete runs of the assembled team, not from selecting among existing predictions, while the same benchmark was used for model selection and evaluation . **b,c,** Mean token usage (**b**) and estimated model-call cost in US dollars (**c**) per run, partitioned into uncached input, cache read, cache write, visible output and reasoning tokens; numbers above the bars give the totals. Teams on the *x* axes are identified by the symbols defined in **a**. Error bars, s.d. (n = 3 runs). All teams predict the same 64 conditions.

To test whether one model is best suited to every agent role, we ran seven single-model teams, in which one LLM served all five agents, three times each (Fig. 5a; Supplementary Figs. S4–S9, S11–S16, S18–S23 and S29–S35). No model led every task. Reaction-process accuracy varied little, from 85.9% for Grok 4.6 to 91.1% for GPT-5.6 Sol, Claude Fable 5 and GPT-6

Astra. The other tasks spread further: phase accuracy ranged from 72.4% (GPT-5.6 Sol) to 81.2% (Gemini 3.7 Flash), rate-capability accuracy from 48.7% (GPT-5.6 Sol) to 57.5% (DeepSeek V4 Pro) and long-cycle accuracy from 61.8% (GPT-5.6 Sol) to 67.4% (DeepSeek V4 Pro and GPT-6 Astra).

Considering both accuracy and cost (Fig. 5), we therefore assembled a Hybrid team that assigned each role to a model that led, or nearly led, the corresponding task. GPT-6 Astra for the chemistry agent, Gemini 3.7 Flash for the materials agent, and DeepSeek V4 Pro for the two electrochemistry agents and the AI scientist. The Hybrid was run through the complete workflow, rather than assembled by selecting answers from existing outputs. Over three complete runs it reached pooled accuracies of 91.7% (176/192), 82.8% (159/192), 59.2% (487/822) and 68.1% (280/411) for the reaction process, phase, rate capability and long-cycle performance (Fig. 5a; Supplementary Figs. S10, S17, S24 and S36). These were the highest values among all configurations, 0.5–1.7 pp above the best single-model team in each task. Because model selection and evaluation used the same benchmark, the Hybrid identifies a task-matched configuration within this benchmark rather than a validated optimum.

The Hybrid retained the class-specific weaknesses of its components. It recognized pure anatase in 22.2% (4/18) of judgements, compared with 91.2% (104/114) for anatase/rutile mixtures, and its discharge-capacity and retention predictions met the ±30% criterion in 44.5% and 59.9% of pairs (Supplementary Fig. S17). Its electrochemical predictions were nevertheless most accurate for anatase-rich products. Rate-capability predictions met the ±30% criterion in 81.2% (39/48) of evaluated fields for the six pure-anatase products and in 72.3% (282/390) for products containing at least 80% anatase, compared with 47.0% (93/198) for products containing at least 80% rutile (Supplementary Fig. S38b,d).

Cost did not follow accuracy (Fig. 5b,c). Among the single-model teams, mean model-call costs per run spanned a 13-fold range, from US$9.47 for GLM-5.3 to US$118.94 for Claude Fable 5, whereas pooled accuracies differed by less than 10 pp in every task. The Hybrid used 22.96 million tokens and US$17.13 per run, only about US$0.27 per condition for the complete prediction chain. In addition, because each role can be served by a different LLM without changing the workflow, individual agents can adopt newer or cheaper models as they become available.

Discussion

Our results give a bounded answer to whether general-purpose LLMs can take over part of the trial-and-error cycle of materials research. In the Joule-heated TiO_2 system, agents without task-specific training can forecast the reaction process, crystal phase and electrochemical performance of unseen conditions and improved by rewriting natural-language Skills after each round of eight measurements (Fig. 1). How far these forecasts can replace experiments depended on the link in the chain. Process and phase classifications were reliable except near sharp transitions, and electrochemical values were approximate (Fig. 2). Electrochemical predictions were most accurate for anatase-rich products, the practically relevant TiO_2 anodes, and least accurate for rutile-rich products whose capacities approach zero (Supplementary Figs. S37 and S38). Skills revised from feedback predicted identical conditions more accurately in most comparisons, with the largest gain after the first round (Fig. 3), and neither frozen Skills nor Skills rewritten without new data reproduced the full loop (Fig. 4). Measurements stored as text were thus sufficient to adapt general-purpose LLMs to a specific synthesis–performance problem without updating any model weights.

After the cold-start round, errors clustered where outcomes change abruptly with the conditions (Fig. 2): at the onset of self-propagating synthesis and in the narrow window in which pure anatase formed. A threshold that lies between sampled conditions can be located only once measurements fall on both sides of it. For example, the Skills placed the 3-s transition between 50 and 55 A only after V_7, when a 50-A, 3-s condition was measured and found not to trigger self-propagating synthesis. Conditions near such thresholds are therefore where predictions are least reliable, and probably where each new experiment adds the most knowledge. Electrochemical performance was harder still, because predicting an exact capacity is intrinsically more demanding than assigning a class. The capacity of a TiO_2 electrode depends jointly on phase composition, crystallite size and morphology, and on electrode and cell properties that the synthesis inputs do not specify; modest changes in any of these can move a capacity beyond ±30%, most of all at high rates. This sensitivity was greatest for rutile-rich products. Their electrodes stored almost no lithium, so changes of a few tenths of a mAh g^{-1} between early cycles already exceeded the tolerance (Supplementary Fig. S37), and these products gave the lowest accuracies, whereas about three in four rate-capability predictions for anatase-rich products met the criterion (Supplementary Fig. S38). The pooled performance accuracy is therefore limited mainly by the intrinsic instability of these near-inactive electrodes and does not directly measure the predictive ability of the models. Structural information alone did not remove this difficulty. From V_1 onwards, the agents revised their Skills with the

measured phase fractions and crystallite sizes of all disclosed conditions, yet round-level rate-capability accuracy still ranged from 36.1% to 65.9% without a steady rise (Extended Data Fig. 2). The present numerical predictions therefore capture the magnitude and rate dependence of the capacity and belong to the screening stage of materials development rather than to final characterization.

Two alternative explanations do not fit the data (Figs. 3 and 4 and Extended Data Fig. 2). If the agents had recalled published outcomes, they would have been accurate from the start. Instead, the synthesis route had no precedent for TiO_2, the cold-start round produced half of all reaction-process errors and accuracy on identical conditions rose once Skills were revised. If the gains had come from computation alone, rewriting Skills without new measurements should have helped, yet it reduced accuracy in most tasks, and requesting maximum reasoning did not improve all tasks together. Pretraining still supplied general chemistry, as illustrated by the cold-start predictions that already invoked the amorphous–anatase–rutile crystallization sequence. The task-specific gains, however, came from the measurements. This mirrors findings for language agents that improve by storing verbal reflections on feedback without weight updates[45] and for LLMs that fail to correct their own reasoning without an external signal[46]. Here the external signal was a physical measurement, and the Skills improved only when one arrived.

This mode of use also differs from most LLM agents in chemistry and materials science, which propose molecules, materials or experimental plans that are then executed and confirmed[1,39,47,48], in some cases within closed optimization loops[49,50] or autonomous robotic laboratories[28]. Here the complete outcome chain is predicted first and experiments only score and correct the forecasts (Figs. 1 and 2), so reliable predictions could replace exploratory trials outright. The framework learned from only eight measurements per round, a regime in which conventional ML usually struggles, extending evidence that LLMs can learn from small chemical datasets[41] and that failed syntheses carry predictive value[6]. The knowledge reached later rounds as readable condition–outcome statements with stated ranges of validity, so every change in a prediction can be traced to a revised rule and checked against the measurements that prompted it, rather than relying on opaque black-box reasoning.

Two properties make this approach practical (Figs. 1 and 5). A complete prediction chain cost the Hybrid team about US$0.27 per condition in model calls, whereas each measured condition required synthesis, phase analysis and electrochemical testing that included 500 charge–

discharge cycles. The framework also operates entirely in natural language. Experimentalists supply conditions and results and can audit the Skills without curating training sets or running electronic-structure calculations, which lowers the barrier that first-principles and conventional ML methods impose on synthesis research. Within one workflow, agents for AI, chemistry, materials science and electrochemistry worked from shared conditions and each other's predictions, connecting tasks that are usually handled by separate specialists.

All evidence comes from one material system, one Joule-heating apparatus and a fixed 64-condition matrix (Fig. 1). The forecasting capability is therefore established within this space, whereas its transfer to other chemistries, synthesis routes and independently chosen conditions remains untested. The Hybrid team was selected and evaluated on the same benchmark, so its composition is task-matched rather than independently validated, and whether Skills written by one model stay useful when another model takes over the same role is untested. The most decisive next test is prospective: applying the framework to a new condition space, synthesizing only the conditions it ranks as uncertain or promising and counting the experiments avoided at a target accuracy. Such a test would measure directly how many trials LLMs can replace. The results point towards a prediction-first mode of materials research, in which experiments are spent mainly where predictions are least certain.

Methods

Study design and sealed experimental reference

We tested whether a multi-agent LLM framework can predict the outcome chain of a synthesis, from reaction process to crystal phase to electrochemical performance, before the outcome is known, and whether revising natural-language knowledge files, termed Skills, from experimental feedback improves those predictions (Fig. 1). The benchmark was TiO_2 prepared by rapid Joule heating. It contained 64 unique conditions in a complete 8 × 8 current–time grid: 25, 30, 35, 40, 45, 50, 55 and 60 A, each combined with 1.5, 3, 5, 10, 15, 20, 25 and 30 s. A predefined stratified allocation divided these conditions into eight prediction rounds, V_0–V_7, with every current and every duration represented once per round (Supplementary Table S1). Conditions were fixed in advance and were not selected adaptively from model outputs.

The logic of the study is prediction first, experiment second. In every round, the agents predicted the complete outcome chain of eight conditions, the predictions were frozen, and only then were the corresponding measurements released to score the predictions and to revise the Skills. Experiments therefore served solely as the validation reference and as the source of feedback. Because synthesis, X-ray diffraction and electrochemical testing, including 500 charge–discharge cycles, take far longer than a prediction round, all 64 conditions were in practice synthesized and characterized before the prediction study began. The resulting dataset was sealed, ensuring that the agents had no access to the outcomes of the current or subsequent rounds during prediction, and each round's results released only after its predictions had been frozen. Round labels therefore refer to the prediction and disclosure sequence rather than to the chronology of physical synthesis. All repeated runs and model configurations used the same sealed dataset. The workflow automated prediction, evaluation and Skill revision, whereas synthesis and characterization were carried out by the authors outside the workflow, as described in the final experimental sections of the Methods.

Agent roles and prediction inputs

The framework comprised one chief agent and four specialists for reaction-process classification, phase classification, rate capability and long-cycle performance. Each input record contained the voltage limit, current, heating duration and measured precursor mass, together with the common experimental background. The voltage limit was the power-supply upper bound rather than a constant measured sample voltage. The reaction-process task predicted Process 1 or Process 2. The phase task selected one of five labels: Amorphous, Amorphous/Anatase, Anatase, Anatase/Rutile and Rutile.

The chief first dispatched reaction-process prediction. Its predicted labels were then supplied to the phase specialist. After phase prediction, the rate-capability and long-cycle specialists received both upstream prediction tables and could operate in parallel. The rate specialist produced six capacity predictions, and the long-cycle specialist produced three numerical endpoints. Both performance specialists were instructed to withhold numerical predictions when the upstream predicted reaction class was Process 1. Those fields were recorded as “not predicted”, while a Process 2 prediction required all corresponding numerical outputs. Upstream classification therefore affected both the information supplied downstream and numerical answer coverage.

No additional pretraining or task-specific fine-tuning was performed. At V_0, models received the task instructions and experimental background but no project-specific Skills. In later rounds, specialists could read the current condition file, their authorized current Skill and the upstream predictions required by their task. Access to experimental-result files and to the cumulative experimental archive was prohibited during prediction, so historical experience reached the predictions only through the Skills. Disclosed measurements could be read directly during Skill revision. The complete prompts of all agents are given in Supplementary Note 1.

Prediction freezing and Skill revision

Specialists submitted full prediction records and structured summaries through an artifact-submission tool that checked the role, stage, round, required fields and all eight sample identifiers. Validated artifacts were written atomically and assigned SHA-256 digests, cryptographic fingerprints of the file contents that change if any character is altered, so that later modification of a frozen file would be detectable. A local aggregation script checked their consistency with the conditions and upstream summaries before assembling the prediction tables. The chief coordinated submissions and repairs but could not recompute or rewrite specialist predictions. Artifact validation permitted an initial check and at most two specialist repair attempts before a repeated gate failure stopped progression.

Release of experimental results was permitted only after final prediction assembly and freezing. After release, prediction artifacts could not be rewritten, and recovery was restricted to evaluation, archiving, Skill revision and publication. Local scripts scored the frozen predictions and added the disclosed measurements to a cumulative archive, and judgement files were retained separately. Each specialist then received its prediction record, summary, evaluation, disclosed archive and previous Skill. Revision required examining correct, incorrect and uncertain predictions, retaining, amending, merging, downgrading or deleting prior relationships, identifying supporting and conflicting observations, and separating untested hypotheses from relationships used for prediction. A suspected experimental error could be recorded as a hypothesis but did not permit alteration of a measured value.

Skills recorded the task scope, evidence sources, current predictive relationships, changes from the preceding version, unresolved hypotheses and instructions for future use. Publication preserved historical versions and registered the new version for the next round. Denoting the Skill published after round V_x by K_x, predictions in V_x used K_{x-1} for $x \geq 1$, and V0 used no Skill. A round's prediction score therefore preceded production of its newly published Skill, and the final K_7 had no subsequent V_8 test. Four specialists publishing eight versions produced 32 Skills per run and 96 across the three Gemini 3.7 Flash runs (Supplementary Data 1). The inspected primary run used local tools only, and its recorded calls contained no external literature search or webpage retrieval.

Execution environment and model configuration

The primary evaluation used Gemini 3.7 Flash for the chief and all specialists. The inspected run used OpenCode 1.18.18 and the model identifier gemini-3.7-flash, with a configured context limit of 1,048,576 tokens and output limit of 65,536 tokens per role. These limits were request budgets rather than observed token consumption. The default request variant was used. Temperature, top-p and a sampling seed were not explicitly specified in the inspected primary configuration.

An external supervisor ran the eight rounds sequentially and started a separate chief session for each round. Prediction and Skill-revision tasks were separate specialist sessions. Cross-round state was conveyed through versioned Skills and the disclosed archive, while within-round recovery could resume the same round session. The inspected primary run contained eight chief sessions and sixteen sessions per specialist, corresponding to prediction and revision in all eight rounds. The artifact-repair attempts described above were distinct from application programming interface (API) transport retries.

Prediction scoring and answer coverage

Reaction-process and phase predictions were correct only when they exactly matched the experimental category. Confusion-matrix percentages were normalized within each experimental class and classes with no experimental observations were marked as not applicable. Numerical predictions used a nominal inclusive ±30% tolerance. The implemented scorer rounded acceptance bounds outward to the decimal precision of the submitted prediction. For measured value y and a prediction written with d decimal places, the bounds were

$$L = \frac{\lfloor 0.70y \times 10^d \rfloor}{10^d},$$

$$U = \frac{\lceil 1.30y \times 10^d \rceil}{10^d}.$$

A prediction was correct when $L \leq \hat{y} \leq U$. This outward rounding could admit deviations slightly beyond an unrounded 30% boundary. For $y = 0$, the bounds were both zero and only a zero prediction passed. Percentage endpoints were evaluated on the numerical percentage scale.

Electrochemical fields were unevaluated if a measurement was unavailable or no numerical prediction was supplied, including outputs withheld following a predicted Process 1 label. Unevaluated fields were excluded from the accuracy denominator. Missing measurements and withheld model predictions were retained as distinct record states. Performance accuracy was consequently conditional on measurement availability and a numerical answer. Answer coverage was the fraction of measured endpoint fields for which a numerical prediction was supplied.

For each task, accuracy was 100% × correct evaluated fields/evaluated fields. Rate-capability accuracy combined six endpoints, and long-cycle accuracy combined three. Aggregate scores pooled correct and evaluated counts across rounds and runs before division and they were not arithmetic means of run-level percentages. Condition-resolved classification maps reported the number of correct predictions among three runs, and numerical maps retained endpoint-level judgements and distinguished unevaluated fields. Prediction–measurement plots included evaluated pairs only, and their equality line and nominal ±30% band were descriptive references (Fig. 2 and Extended Data Fig. 1).

Matched-condition Skill comparisons

Three dedicated comparison runs evaluated saved Skill configurations on the same eight-condition target rounds (Fig. 3). At V_2–V_7, the newer configuration used K_{x-1} and the older configuration used K_{x-2}. Scores were recalculated from frozen predictions and the corresponding measurements, with saved input, prediction and outcome hashes checked before aggregation. Saved request identities were additionally audited. At V_1, the initial V_0 Skill group was compared with a separate No Skill group on the V_1 conditions; this was a group contrast rather than a paired adjacent-version test.

Upstream predictions were not held fixed in every comparison. Across three runs, six target rounds and three downstream tasks, upstream inputs differed in 21 of 54 older-versus-newer pairs, including five phase comparisons and eight comparisons for each performance task. The analysis therefore compared team configurations, including possible propagation of upstream changes, rather than isolating the direct causal effect of one specialist's Skill.

Each configuration's task accuracy used measured-and-answered fields, so older and newer configurations could have different answered subsets and denominators. Within each run, team accuracy was the equally weighted mean of the four task accuracies. Displayed version changes were arithmetic means of the three run-level changes, rather than differences of pooled field counts.

Let $A(B, K)$ denote task or team accuracy for configuration K on round batch B. For $x \geq 2$, the same-batch update term was $G_x = A(B_x, K_{x-1}) - A(B_x, K_{x-2})$. The batch-shift term using the older configuration was $D_x = A(B_x, K_{x-2}) - A(B_{x-1}, K_{x-2})$. The net change was $N_x = G_x + D_x$. Differences were calculated in percentage points before display rounding. This decomposition is an accounting identity, and its terms can reflect changes in upstream inputs, answer coverage, endpoint composition and generated outputs as well as in the saved Skills.

Ablation protocols

The full Gemini 3.7 Flash workflow was compared with frozen-Skill, no-new-feedback and No Skill workflows across V_0–V_7, each with three runs (Fig. 4 and Supplementary Figs. S25–S27). In the frozen-

Skill workflow, the V_0 revision produced both textual Skills and executable rules specifying input domains, matching conditions, outputs or formulas, confidence and evidence. For V_1–V_7, a rule-application tool generated predictions from these frozen rules, and specialists submitted the results without reinterpretation or coefficient revision. Uncovered conditions or multiple matching rules stopped progression. The published Skills remained those derived from V0, so this control froze both the knowledge and its executable application.

The no-new-feedback workflow received experimental feedback exactly once, with the V_0 outcomes and judgements released after V_0 prediction had been frozen and used for the first Skill revision, as in the full workflow. In V_1–V_7, specialists revised their Skills using only their own current prediction records and previous Skill, without rereading the V_0 feedback or receiving any later measurements, judgements or experimental-history tables. Saved predictions were evaluated externally against the reference dataset, and those scores were not returned to the agents.

The No Skill workflow omitted Skills throughout, used fresh chief and specialist prediction sessions, and prohibited model access to historical outcomes, judgements and prediction archives. Its local evaluation and archiving returned workflow status without exposing true labels or scores to the agents. All controls retained the four specialist roles and upstream dependencies, and no single-agent control was part of these comparisons.

V_0 was included in aggregate scores. The workflows were allowed to differ in model-call count, token use and numerical coverage, so comparisons were not matched for computational budget or answered field set.

A supplementary comparison evaluated Gemini 3.7 Flash request configurations labelled default and max (Supplementary Fig. S28). The max configuration set reasoningEffort to max for all five roles, whereas default did not explicitly set this option, with both using the same model identifier and configured limits. This comparison specifies the requested option and does not establish the effort actually applied within the serving model.

Single-model teams and Hybrid assembly

Seven single-model teams used GPT-5.6 Sol, DeepSeek V4 Pro, Gemini 3.7 Flash, Claude Fable 5, GLM-5.3, Grok 4.6 or GPT-6 Astra for all five agent roles (Fig. 5 and Supplementary Figs. S29–S35). Their configured model identifiers were gpt-5.6-sol, deepseek-v4-pro, gemini-3.7-flash, claude-fable-5, glm-5.3, grok-4.6 and gpt-6-astra, respectively. A single-model team remained a multi-agent system. Each configuration used three runs with 32 Skills, the same 64 conditions and the same disclosure sequence.

For each specialist task, models were ranked by pooled prediction accuracy across V_0–V_7 and the three runs. Ties in accuracy were resolved by selecting the lowest-priced model. The Hybrid team assigned GPT-6 Astra to reaction-process prediction, Gemini 3.7 Flash to phase prediction, and DeepSeek V4 Pro to the chief and both electrochemical-performance roles. It was executed through the complete workflow with its own predictions and Skill histories, rather than constructed by selecting correct outputs from single-model teams. Role assignment and Hybrid evaluation used the same benchmark, with no independent model-selection dataset used.

Repeated runs and statistical summaries

Three computational runs were performed for each reported configuration. All runs reused the same sealed experimental dataset and did not represent independent synthesis or cell replicates. Multiple numerical endpoints from one material condition were output fields rather than independent physical samples. Within each configuration, runs are numbered 1–3 in all figures.

Accuracy figures used pooled correct/evaluated counts for bar centers and the sample standard deviation of the three run-level accuracy percentages for error bars. Resource figures used the arithmetic mean of the three run totals and their sample standard deviation. Standard deviations used an $n - 1$ denominator, with $n = 3$. Ranges in the Skill-version records gave the minimum and maximum run-level percentages. These

quantities were descriptive and were not confidence intervals. Matched-condition comparisons displayed individual runs and arithmetic mean changes without inferential significance testing.

Token accounting and estimated cost

Resource accounting used provider-reported usage stored for assistant responses associated with the selected project sessions. Recorded components were uncached input, cache read, cache write, visible output and separately reported reasoning or thinking tokens. Component totals included all selected chief and specialist assistant responses, including coordination, prediction, Skill revision and recorded recovery. Zero separately reported reasoning tokens did not establish that a model performed no internal reasoning.

For response *j*, estimated US-dollar cost was

$$C_j = \frac{I_j r_{I,j} + R_j r_{R,j} + W_j r_{W,j} + (O_j + H_j) r_{O,j}}{10^6},$$

where *I*, *R*, *W*, *O* and *H* were the five recorded token components, and *r* denoted the corresponding rate per million tokens. Rates came from the model-specific tariff schedule used by the analysis, including context-length tiers (Supplementary Table S3). Prompt length for tier selection was input plus cache-read and cache-write tokens. Chinese-yuan tariffs were converted at 7.20 CNY per US dollar. When a cache-write tariff depended on retention duration and that duration was unavailable, the five-minute tariff was used. Hybrid responses were priced using the model assigned to each role. Response costs were summed within a run before calculating the three-run mean and sample standard deviation.

Costs were estimates of model-token charges under these assumptions, rather than reconciled invoices, and excluded physical experiments, tool charges, storage, taxes and regional adjustments. Token totals from different model tokenizers were not treated as equivalent amounts of physical computation.

Precursor preparation and Joule heating

Titanium butoxide (TBOT, ≥99.0%, Aladdin) and anhydrous ethanol (≥99.5%, Sinopharm Chemical Reagent) were used as received. TBOT (20 ml) was added to ethanol (80 ml) and stirred for 0.5 h. Ultrapure water (20 ml) was added under vigorous stirring, followed by a further 1 h of stirring. The suspension was aged and dried in a forced-air oven at 50 °C for 12 h to obtain the precursor powder.

Approximately 1.5 g of precursor was pressed at 10 MPa for 10 s into a pellet 2 cm in diameter. The recorded precursor masses ranged from 1.44 to 1.60 g. Each pellet was placed on carbon felt measuring 3.5 cm × 5 cm × 2 mm. The felt was clamped between upper and lower carbon plates connected to the power-supply leads by electrode clips. Heating was performed in air using a computer-controlled Chroma 62024P-80-60 power supply in constant-current mode, followed by natural cooling in air. Reaction behavior was classified by whether the reaction remained self-sustaining after electrical power was switched off. Process 1 involves Joule heating alone, whereas Process 2 additionally induces a self-propagating combustion reaction, triggered by the high-temperature ignition of the butoxy- and hydroxy-containing TiO_2 precursor.

Current and voltage were logged by the power-supply control software throughout each heating event. The temperature at the center of the carbon felt was recorded at 1-s intervals using an SA-JS16A infrared pyrometer, with emissivity set to 0.95 and a measurement range of 300–1600 °C. Peak felt temperature was the first occurrence of the recorded maximum within the first sustained heating event. Electrical power was calculated as $P_i = U_i I_i$, and electrical energy input was calculated by trapezoidal integration:

$$E = \Sigma_i \left[\frac{P_i + P_{i+1}}{2}\right] (t_{i+1} - t_i).$$

Integration used the recorded timestamps from the first to the final exported measurement, including initial zero-power records and the current-rise stage. Instantaneous power denoted the initial recorded power maximum during current rise after switch-on. Average power was the integrated energy divided by the

elapsed time between the first and final records. These quantities describe electrical input and felt temperature and the integration interval was not replaced by the nominal heating duration (Supplementary Fig. S1).

X-ray diffraction and phase characterization

Powder X-ray diffraction was performed using a PANalytical X'Pert3 diffractometer with Cu $K\alpha$ radiation at 40 kV and 20 mA. Patterns were acquired over $2\theta = 23–29°$ at 1° min^{-1}. This range contains the strongest reflections of anatase (101) and rutile (110). For several samples, refinement of full-range slow scans gave phase contents nearly identical to those obtained from the 23–29° range, so the shorter range was used for all samples to reduce measurement time. Phase contents and phase-specific crystallite sizes were obtained by fitting this range in GSAS-II (commit 3fc6dc, 14 May 2024; version tag 5786), using anatase and rutile starting structures from COD entry 9009086 and ICSD entry 88625, respectively. Reported fits had $R_w <$ 8%. No internal standard was used. Phase percentages therefore describe model-dependent fractions from the fitted range and do not establish absolute crystalline and amorphous mass fractions. Crystallite sizes refer to coherent diffraction domains rather than microscopy-derived particle sizes.

Products without resolved diffraction peaks within 23–29° were operationally screened as amorphous and were not subjected to phase refinement or electrochemical testing. Fourteen conditions fell in this screening group and their benchmark phase-map label was Amorphous/Anatase. Screening terminology and the prediction label set were kept separate. The experimental map contained four labels (Amorphous/Anatase, Anatase, Anatase/Rutile and Rutile), whereas the model output vocabulary additionally allowed Amorphous. Phase fractions and crystallite sizes were disclosed experimental descriptors and were not separate scored prediction targets (Supplementary Fig. S2).

Electrode preparation and electrochemical testing

TiO_2, acetylene black and poly(vinylidene fluoride) were mixed at an 8:1:1 mass ratio. N-Methyl-2-pyrrolidone was added, and the mixture was manually ground to form a homogeneous slurry. Aluminium foil was used as the current collector. Electrodes were vacuum-dried at 50 °C for 12 h and cut into discs 12 mm in diameter, with a nominal TiO_2 loading of approximately 1.3 mg cm^{-2}. Specific capacities and currents were normalized to TiO_2 mass alone.

CR2032 lithium half-cells were assembled in an argon-filled glovebox with water and oxygen levels below 1 ppm. Lithium foil served as the counter and reference electrode, and Celgard 2400 polypropylene was used as the separator. Each cell contained 0.4 ml of 1 M $LiPF_6$ in ethylene carbonate, diethyl carbonate and dimethyl carbonate (1:1:1 by volume; DoDoChem). Cells rested for 12 h before testing at room temperature. Galvanostatic tests were performed using a LAND CT2001A system over 1.0–3.0 V versus Li/Li^+. Fifty material conditions had electrochemical measurements. The remaining fourteen screened conditions described above had unavailable electrochemical endpoints because the resulting products contained unreacted precursor and were either electrochemically inactive or exhibited poor electrochemical performance.

The rate sequence was 1C, 2C, 5C, 10C and 20C followed by return to 1C, with ten cycles per stage and 1C defined as 170 mA g^{-1}. Each rate endpoint was the mean discharge capacity of its ten-cycle stage. Long-cycle testing comprised 500 cycles at 5C. The three long-cycle endpoints were the arithmetic mean discharge capacity over cycles 1–500, the arithmetic mean Coulombic efficiency over cycles 1–500, and capacity retention, calculated as $100\% \times Q_{500}/Q_2$, where Q_n is the discharge capacity at cycle n. Coulombic efficiency was expressed as a percentage. Capacities were reported in mAh g^{-1}, and efficiency and retention in percent (Supplementary Fig. S3).

References

1. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large

language models. *Nature* **624**, 570–578 (2023).

2. Guo, J. et al. AI-guided design of efficient perovskite solar cells operationally stable at 100 °C. *Science* **392**, 724–728 (2026).
3. Feng, B. et al. Agentic AI systems for scientific discovery. *Nat. Rev. Methods Primers* **6**, 81 (2026).
4. Tabor, D. P. et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **3**, 5–20 (2018).
5. de Pablo, J. J. et al. New frontiers for the materials genome initiative. *npj Comput. Mater.* **5**, 41 (2019).
6. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
7. Leeman, J. et al. Challenges in high-throughput inorganic materials prediction and autonomous synthesis. *PRX Energy* **3**, 011002 (2024).
8. Yao, Y. et al. Carbothermal shock synthesis of high-entropy-alloy nanoparticles. *Science* **359**, 1489–1494 (2018).
9. Greeley, J., Jaramillo, T. F., Bonde, J., Chorkendorff, I. & Nørskov, J. K. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nat. Mater.* **5**, 909–913 (2006).
10. Burger, B. et al. A mobile robotic chemist. *Nature* **583**, 237–241 (2020).
11. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864–B871 (1964).
12. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
13. Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
14. Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
15. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the Open Quantum Materials Database (OQMD). *JOM* **65**, 1501–1509 (2013).
16. Curtarolo, S. et al. AFLOW: an automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).
17. Urban, A., Seo, D.-H. & Ceder, G. Computational understanding of Li-ion batteries. *npj Comput. Mater.* **2**, 16002 (2016).
18. Cohen, A. J., Mori-Sánchez, P. & Yang, W. Challenges for density functional theory. *Chem. Rev.* **112**, 289–320 (2012).
19. Lejaeghere, K. et al. Reproducibility in density functional theory calculations of solids. *Science* **351**, aad3000 (2016).
20. Bursch, M., Mewes, J.-M., Hansen, A. & Grimme, S. Best-practice DFT protocols for basic molecular computational chemistry. *Angew. Chem. Int. Ed.* **61**, e202205735 (2022).
21. Jang, J., Gu, G. H., Noh, J., Kim, J. & Jung, Y. Structure-based synthesizability prediction of crystals using partially supervised learning. *J. Am. Chem. Soc.* **142**, 18836–18843 (2020).

22. Sun, W. & David, N. A critical reflection on attempts to machine-learn materials synthesis insights from text-mined literature recipes. *Faraday Discuss.* **256**, 614–638 (2025).
23. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
24. Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
25. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
26. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2019).
27. Merchant, A. et al. Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023).
28. Szymanski, N. J. et al. An autonomous laboratory for the accelerated synthesis of inorganic materials. *Nature* **624**, 86–91 (2023).
29. Zhang, Y. & Ling, C. A strategy to apply machine learning to small datasets in materials science. *npj Comput. Mater.* **4**, 25 (2018).
30. Artrith, N. et al. Best practices in machine learning for chemistry. *Nat. Chem.* **13**, 505–508 (2021).
31. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems 30* 5998–6008 (Curran Associates, 2017).
32. Brown, T. B. et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33* 1877–1901 (Curran Associates, 2020).
33. OpenAI. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).
34. Jablonka, K. M. et al. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digit. Discov.* **2**, 1233–1250 (2023).
35. Dagdelen, J. et al. Structured information extraction from scientific text with large language models. *Nat. Commun.* **15**, 1418 (2024).
36. Zheng, Z., Zhang, O., Borgs, C., Chayes, J. T. & Yaghi, O. M. ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. *J. Am. Chem. Soc.* **145**, 18048–18062 (2023).
37. Zheng, Z. et al. ChatGPT research group for optimizing the crystallinity of MOFs and COFs. *ACS Cent. Sci.* **9**, 2161–2170 (2023).
38. Romera-Paredes, B. et al. Mathematical discoveries from program search with large language models. *Nature* **625**, 468–475 (2024).
39. M. Bran, A. et al. Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* **6**, 525–535 (2024).
40. Antunes, L. M., Butler, K. T. & Grau-Crespo, R. Crystal structure generation with autoregressive large language modeling. *Nat. Commun.* **15**, 10570 (2024).
41. Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A. & Smit, B. Leveraging large language models for predictive chemistry. *Nat. Mach. Intell.* **6**, 161–169 (2024).
42. Niyongabo Rubungo, A., Arnold, C., Rand, B. P. & Dieng, A. B. LLM-Prop: predicting the properties of crystalline materials using large language models. *npj Comput. Mater.* **11**, 186

(2025).

43. Kim, S., Jung, Y. & Schrier, J. Large language models for inorganic synthesis predictions. *J. Am. Chem. Soc.* **146**, 19654–19659 (2024).
44. Hu, Y.-S., Kienle, L., Guo, Y.-G. & Maier, J. High lithium electroactivity of nanometer-sized rutile TiO_2. *Adv. Mater.* **18**, 1421–1426 (2006).
45. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K. & Yao, S. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36* 8634–8652 (Curran Associates, 2023).
46. Huang, J. et al. Large language models cannot self-correct reasoning yet. In *Proc. 12th International Conference on Learning Representations* (ICLR, 2024).
47. Liu, S. et al. Expert-guided LLM reasoning for battery discovery: from AI-driven hypothesis to synthesis and characterization. Preprint at https://arxiv.org/abs/2507.16110 (2025).
48. Wang, Y. et al. PeroMAS: a multi-agent system of perovskite material discovery. In *Proc. 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining* 12339–12350 (ACM, 2026).
49. Ji, K. et al. A closed-loop AI framework for hypothesis-driven and interpretable materials design. Preprint at https://arxiv.org/abs/2509.18604 (2025).
50. Sheng, F. et al. Synthesizing like a chemist: an iterative, feedback-driven loop for materials discovery. Preprint at https://arxiv.org/abs/2608.15928 (2026).

Data availability

The data supporting the findings of this study are available in the Article and its Supplementary Information. Run records, Skills and prompts are provided in Supplementary Data 1. Source data are provided with this paper.

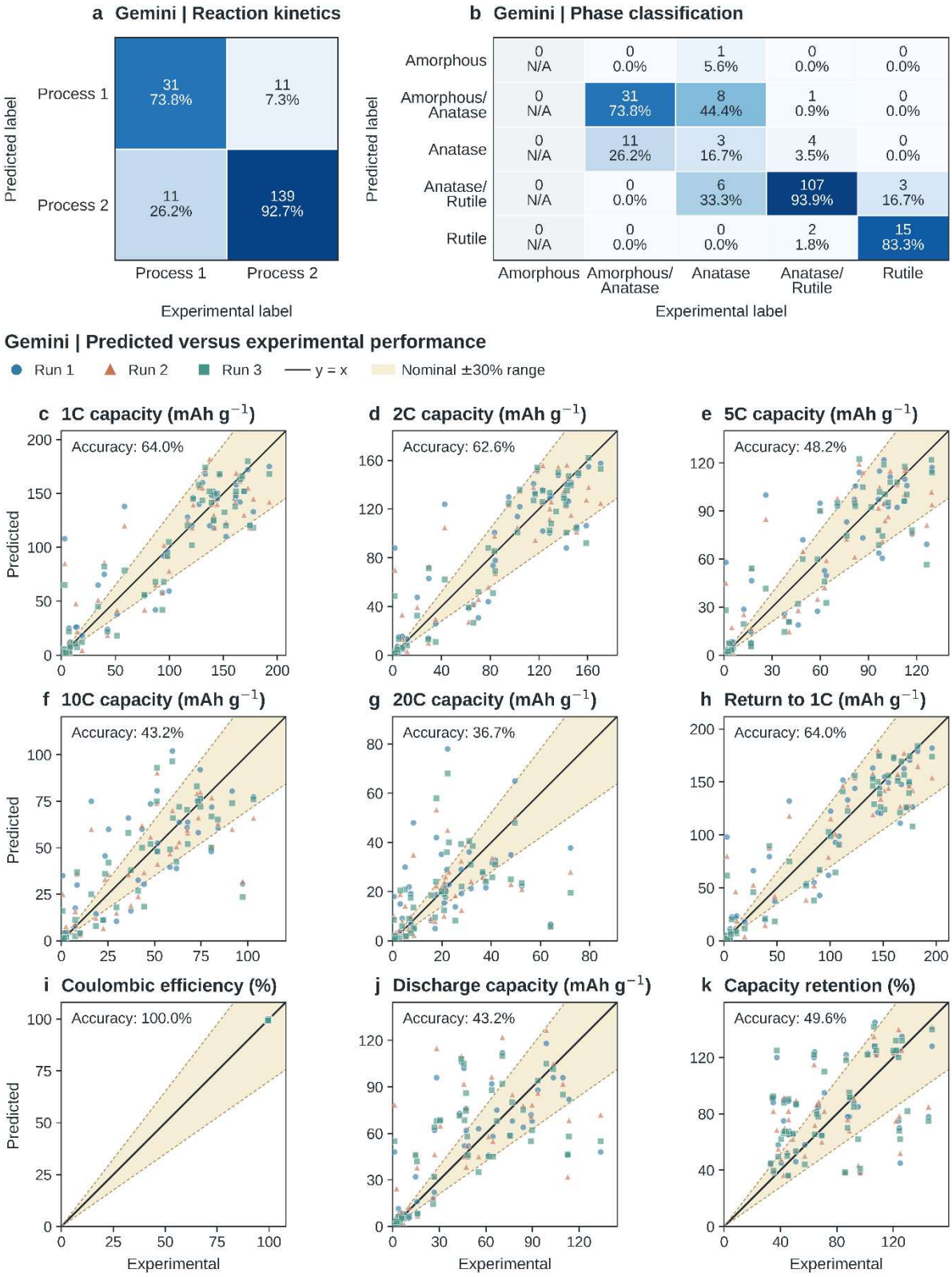
Code availability The code of the framework is provided as Supplementary Software under the MIT licence and will be made publicly available on GitHub and archived at Zenodo upon publication.

Acknowledgements and use of AI tools The LLMs evaluated in this study (GPT-5.6 Sol, DeepSeek V4 Pro, Gemini 3.7 Flash, Claude Fable 5, GLM-5.3, Grok 4.6 and GPT-6 Astra) are the object of the research. They generated the predictions and Skills analyzed here, under the configurations described above, and their prompts, predictions, judgements and Skills are reported in the Supplementary Information and Supplementary Data 1. Separately, AI assistants (ChatGPT and DeepSeek) were used to help edit the language of the manuscript. No generative AI tools were used to generate or modify experimental data or scientific conclusions. All experimental data were produced by the authors, and the authors reviewed and verified all AI-assisted text, figures and code and take full responsibility for the content of the manuscript.

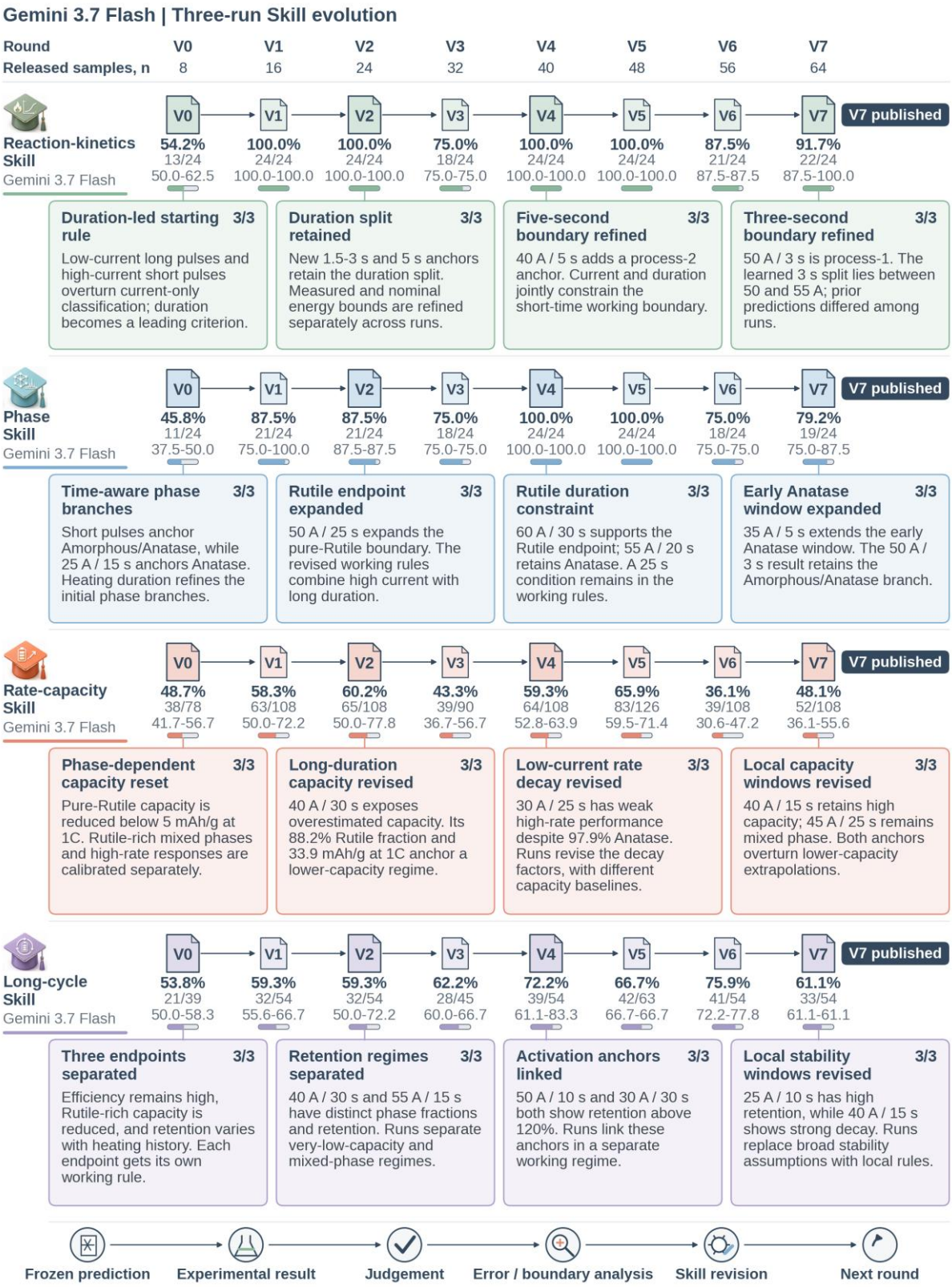
Competing interests The authors declare no conflicts of interest.

Additional information

Supplementary information The online version contains supplementary material available at xxx.



Extended Data Fig. 1 | Classification errors and numerical deviations of the Gemini 3.7 Flash predictions. a,b, Pooled confusion matrices for reaction-process (**a**) and phase (**b**) classification over three runs (192 judgements each). Columns are experimental labels and rows are predicted labels; each cell gives the count and the percentage of its column. No product was fully amorphous, so that column is marked N/A. **c–h,** Predicted versus measured capacity at 1C (**c**), 2C (**d**), 5C (**e**), 10C (**f**), 20C (**g**) and on returning to 1C (**h**). **i–k,** Predicted versus measured Coulombic efficiency (**i**), discharge capacity (**j**) and capacity retention (**k**) after 500 cycles at 5C. Symbols distinguish the three runs; the solid line marks equality, and the shaded band marks ±30% of the measured value. The percentage in each panel is the fraction of the 139 evaluated pairs within this band (rounded bounds; Methods).


Gemini 3.7 Flash | Three-run Skill evolution
Round
Released samples, n
V0 8
V1 16
V2 24
V3 32
V4 40
V5 48
V6 56
V7 64
Reaction-kinetics Skill
Gemini 3.7 Flash
V0 54.2% 13/24 50.0-62.5
V1 100.0% 24/24 100.0-100.0
V2 100.0% 24/24 100.0-100.0
V3 75.0% 18/24 75.0-75.0
V4 100.0% 24/24 100.0-100.0
V5 100.0% 24/24 100.0-100.0
V6 87.5% 21/24 87.5-87.5
V7 91.7% 22/24 87.5-100.0
V7 published
Duration-led starting rule 3/3
Low-current long pulses and high-current short pulses overturn current-only classification; duration becomes a leading criterion.
Duration split retained 3/3
New 1.5-3 s and 5 s anchors retain the duration split. Measured and nominal energy bounds are refined separately across runs.
Five-second boundary refined 3/3
40 A / 5 s adds a process-2 anchor. Current and duration jointly constrain the short-time working boundary.
Three-second boundary refined 3/3
50 A / 3 s is process-1. The learned 3 s split lies between 50 and 55 A; prior predictions differed among runs.
Phase Skill
Gemini 3.7 Flash
V0 45.8% 11/24 37.5-50.0
V1 87.5% 21/24 75.0-100.0
V2 87.5% 21/24 87.5-87.5
V3 75.0% 18/24 75.0-75.0
V4 100.0% 24/24 100.0-100.0
V5 100.0% 24/24 100.0-100.0
V6 75.0% 18/24 75.0-75.0
V7 79.2% 19/24 75.0-87.5
V7 published
Time-aware phase branches 3/3
Short pulses anchor Amorphous/Anatase, while 25 A / 15 s anchors Anatase. Heating duration refines the initial phase branches.
Rutile endpoint expanded 3/3
50 A / 25 s expands the pure-Rutile boundary. The revised working rules combine high current with long duration.
Rutile duration constraint 3/3
60 A / 30 s supports the Rutile endpoint; 55 A / 20 s retains Anatase. A 25 s condition remains in the working rules.
Early Anatase window expanded 3/3
35 A / 5 s extends the early Anatase window. The 50 A / 3 s result retains the Amorphous/Anatase branch.
Rate-capacity Skill
Gemini 3.7 Flash
V0 48.7% 38/78 41.7-56.7
V1 58.3% 63/108 50.0-72.2
V2 60.2% 65/108 50.0-77.8
V3 43.3% 39/90 36.7-56.7
V4 59.3% 64/108 52.8-63.9
V5 65.9% 83/126 59.5-71.4
V6 36.1% 39/108 30.6-47.2
V7 48.1% 52/108 36.1-55.6
V7 published
Phase-dependent capacity reset 3/3
Pure-Rutile capacity is reduced below 5 mAh/g at 1C. Rutile-rich mixed phases and high-rate responses are calibrated separately.
Long-duration capacity revised 3/3
40 A / 30 s exposes overestimated capacity. Its 88.2% Rutile fraction and 33.9 mAh/g at 1C anchor a lower-capacity regime.
Low-current rate decay revised 3/3
30 A / 25 s has weak high-rate performance despite 97.9% Anatase. Runs revise the decay factors, with different capacity baselines.
Local capacity windows revised 3/3
40 A / 15 s retains high capacity; 45 A / 25 s remains mixed phase. Both anchors overturn lower-capacity extrapolations.
Long-cycle Skill
Gemini 3.7 Flash
V0 53.8% 21/39 50.0-58.3
V1 59.3% 32/54 55.6-66.7
V2 59.3% 32/54 50.0-72.2
V3 62.2% 28/45 60.0-66.7
V4 72.2% 39/54 61.1-83.3
V5 66.7% 42/63 66.7-66.7
V6 75.9% 41/54 72.2-77.8
V7 61.1% 33/54 61.1-61.1
V7 published
Three endpoints separated 3/3
Efficiency remains high, Rutile-rich capacity is reduced, and retention varies with heating history. Each endpoint gets its own working rule.
Retention regimes separated 3/3
40 A / 30 s and 55 A / 15 s have distinct phase fractions and retention. Runs separate very-low-capacity and mixed-phase regimes.
Activation anchors linked 3/3
50 A / 10 s and 30 A / 30 s both show retention above 120%. Runs link these anchors in a separate working regime.
Local stability windows revised 3/3
25 A / 10 s has high retention, while 40 A / 15 s shows strong decay. Runs replace broad stability assumptions with local rules.
Frozen prediction
Experimental result
Judgement
Error / boundary analysis
Skill revision
Next round

Extended Data Fig. 2 | Skill version histories in three Gemini 3.7 Flash runs. Rows show the reaction-kinetics, phase, rate-capacity and long-cycle Skills; columns show rounds V_0–V_7 and the cumulative number of conditions whose results had been released. Each version V_x is written after the frozen predictions of round V_x have been scored and is used from round V_{x+1}. The percentage and correct/evaluated count below each version therefore describe the predictions made before that update, and the range gives the minimum and maximum over the three runs. For example, the entry 54.2%, 13/24 and 50.0–62.5 below V0 of the reaction-kinetics Skill means that the V_0 predictions, made without any Skill, were correct in 13 of 24 judgements pooled over the three runs (54.2%), with run-level accuracies from 50.0% to 62.5%; the bar below each entry shows the pooled accuracy. V_0 predictions start without Skills, and the V_7 Skills have no V_8 test. Cards summarize representative revisions found in the Skill texts, such as current–duration boundaries, phase branches, capacity baselines and endpoint-specific long-cycle regimes; 3/3 indicates that the revision theme appears in all three runs, not that the rules are identical. The statements are empirical working relationships rather than established mechanisms. The bottom row shows the cycle of frozen prediction, experimental result, judgement, error and boundary analysis, Skill revision and the next round.

Supplementary Files

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

- SupplementaryInformation.docx