Do humans and large language models agree on the quality of synthesis plans?

Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

Published on 12 August 2026

Cite this: DOI: 10.1039/d6dd00266h

Do humans and large language models agree on the quality of synthesis plans?

Varvara Voinarovska, *[ab] Rocío Mercado, [bc] Mikhail Kabeshov [a] and Samuel Genheden [a]

Received 7th May 2026
Accepted 11th August 2026

DOI: 10.1039/d6dd00266h

rsc.li/digitaldiscovery

Large language models (LLMs) have seen widespread adoption in all spheres of science including chemistry and cheminformatics. Nevertheless, our knowledge of how they operate is limited, giving rise to exploration of their capabilities in different areas of science and different operation modes. Here, we investigated whether LLMs could mimic human experts on the challenging task of assessing retrosynthetic path feasibility in routes generated by a popular computer-aided synthesis planning tool (AiZynthFinder). We evaluated the agreement between LLMs and expert chemists on holistic evaluations of the proposed routes as well as the individual chemical reactions in them. We used four frontier LLMs, three proprietary models and one open-source model (Claude Opus 4.8, GPT-5.5, Llama 3.1 70B, and Gemini 3.1 Pro) and employed 17 expert chemists to grade 50 retrosynthetic paths. We found that, when provided with clearly defined reaction-evaluation categories, human experts tended to converge in their assessments. Among the evaluated models, Gemini 3.1 Pro achieved the highest agreement with the human majority vote. GPT-5.5 and Claude Opus 4.8 were comparatively more pessimistic, whereas Llama 3.1 70B showed a pronounced optimistic bias.

1 Introduction

Retrosynthesis planning identifies viable synthetic routes to target molecules, but it is challenging due to the combinatorial explosion of possible chemical disconnections as molecular complexity increases.[1,2] Many options which appear formally valid often fail due to practical considerations, such as selectivity, protecting group strategies, solvent effects, or scale-up constraints, rendering otherwise plausible synthetic routes infeasible.[3] Furthermore, reaction databases derived from patents and literature are systematically biased towards well-studied and successful transformations, noisy, and frequently omit critical experimental details such as full conditions or yields.[4–6] This biased training data can lead models to overestimate the feasibility of routes in underrepresented chemical space and favor common transformations over rare but strategically valuable ones.[7] As a result, expert judgment remains important for distinguishing between formally valid routes and those that are also practical, safe, and economical, motivating the need for careful, expert-informed evaluation of computational retrosynthesis tools.

Assessing route quality requires more than counting steps. Useful criteria include convergence, confidence in each step, starting material availability, expected selectivity, number of protecting group operations, reagent choice, waste considerations, and expected yield.[1] Practical considerations such as cost, safety, and ease of workup also matter.[8] Making these criteria explicit helps align both human and model-based grading.

Retrosynthesis planning[9] has evolved from rule-based systems to template-based models that apply learned or curated reaction rules, and to template-free models[10] that predict changes directly on molecular graphs or SMILES (Simplified Molecular Input Line Entry Specification).[11] Search algorithms such as Monte Carlo tree search and A* guide route expansion toward purchasable starting materials.[12] AiZynthFinder,[13] ASKCOS,[14] Syntheseus,[15] and SynPlanner[16] are representative open-source packages that combine reaction prediction with guided search. Here, we focus on AiZynthFinder as a baseline for retrosynthesis planning.

In the computer-aided synthesis planning context, large language models (LLMs) are increasingly explored for molecular design, reaction prediction, and synthesis planning,[17–27] where they can complement established computational methods and expert workflows by quickly searching large information spaces and matching relevant precedents to a task.

The most recent research uses LLM-based agentic systems[28] which utilizes LLM as an entity with agency, aka the possibility to act upon decisions. Early chemistry-focused LLM agents,

[a]Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden. E-mail: varvara.voinarovska@astrazeneca.com

[b]AI Laboratory for Molecular Engineering, Department of Computer Science and Engineering, Chalmers University of Technology, University of Gothenburg, Sweden. E-mail: varvarav@chalmers.se

[c]Science for Life Laboratory (SciLifeLab), Gothenburg, Sweden

such as ChemCrow and related agent frameworks for automated scientific discovery and molecular design, demonstrated tool-augmented planning and execution across synthesis, analysis, and hypothesis generation.[18,29] ChemBench[30] evaluated LLMs against expert-level questions across multiple categories, and Coscientist investigated the capabilities of LLMs to be scientific assistants.[31] At the same time, we still do not fully understand how LLMs work or whether their abilities generalize beyond advanced pattern matching,[32] so researchers continue to test them to identify where they help, where they fail, and when their outputs can be trusted.[33] Applications to retrosynthetic analysis and route planning highlight both the promise of LLMs for knowledge extraction and hypothesis generation[24] and the need for rigorous benchmarking and interpretability.[18,30] Beyond ChemBench, several domain-specific and reasoning-centric benchmarks have emerged for chemistry, including error-detection and correction, chain-of-thought chemical reasoning, and function-group and reaction understanding.[34–37]

More broadly, using LLMs to evaluate the outputs of other models or systems, rather than only to generate content, has been formalized as the LLM-as-a-judge paradigm. Zheng *et al.*[38] introduced MT-Bench and Chatbot Arena to systematically study this approach, demonstrating that LLM judges can achieve over 80% agreement with human evaluators on open-ended tasks while also revealing systematic biases such as position bias, verbosity bias, and self-enhancement bias. Subsequent work has shown that LLM judges tend to align more closely with non-expert than expert annotators across diverse NLP tasks,[22] raising questions about whether high surface-level agreement masks deficiencies in domain-specific reasoning. Within chemistry, this paradigm is already being adopted: AlchemyBench employs an LLM-as-a-judge framework to evaluate materials synthesis recipes against expert-defined criteria and reports strong but imperfect correlation with human specialists,[39] while RetroReasoner uses LLM-based reward signals to verify retrosynthetic predictions.[40] Our work is a domain-specific instantiation of this paradigm applied to retrosynthetic route assessment. By comparing LLM judgments directly against expert consensus on structured feasibility criteria, we can disentangle what aspects of the observed agreement and disagreement reflect general LLM evaluation behavior, such as optimism bias, from what is specific to the practical reasoning demands of synthetic chemistry.

Human-in-the-loop (HITL) approaches offer another way to strengthen computer-aided synthesis planning (CASP) by incorporating expert knowledge where it is most impactful: proposing or discarding key disconnections, ranking alternative precursors, and flagging unreliable reaction classes, or filtering LLM-generated retrosynthetic transformations.[41,42] This feedback can be used to re-rank routes, update scoring functions, or train preference models that guide search. Compared to HITL in *de novo* molecular design, feedback in retrosynthesis planning focuses on route feasibility and operational practicality rather than molecule generation alone. HITL has been applied to guide molecule generation using reinforcement learning policies in REINVENT, and recent work has used expert knowledge to rank retrosynthetic paths.[43–45]

A central question for CASP practice is whether LLMs can approximate expert preferences when ranking candidate routes, a theme paralleled in other domains, such as conference reviewing.[46–48] If LLM ratings correlate with expert judgments, they could be used for rapid filtering before deeper review, reducing time spent on routine triage. If they diverge, the analysis reveals where LLMs miss practical constraints or overvalue formal feasibility. Measuring agreement, bias, and variance against expert consensus clarifies when LLMs can be trusted as proxies and when human review is essential.

In this paper, we investigate whether general-purpose proprietary and open-source LLMs can detect and reason about retrosynthetic plans in meaningful agreement with human-expert consensus in a CASP setting. We ask whether LLMs can be used to filter reactions and retrosynthetic paths in alignment with expert concerns and whether they can identify problematic steps in route generation. To do this, we compare human expert opinions and LLM-based grading on the same retrosynthetic routes generated by AiZynthFinder. We use standardized categories that cover feasibility, step reliability, and practicality, and we evaluate agreement and systematic differences between human experts and LLMs. This design isolates the grading task from route generation and identifies conditions where LLM assessments align with expert practice. We demonstrate our approach in Fig. 1.

2 Methods

2.1 Dataset

We selected 50 drug-like compounds from the *Journal of Medicinal Chemistry* (JMC) with the help of the experts to identify drug-like molecules challenging for retrosynthesis analysis. We solved routes for each of the 50 selected molecules through AiZynthFinder.[13] We used AiZynthFinder v4.3.1 and default settings to generate 10 routes per molecule, from which we selected the top-1 route to present to chemists and LLMs for evaluation. We presented routes leading to commercially available starting material (solved routes) as well as routes that only partially lead to commercial starting materials (unsolved routes). We presented the selected routes to 17 chemists with different backgrounds and education levels; the distribution of the level of education can be found in the SI. Most of the experts were from AstraZeneca, with some external experts. Each expert graded 20 routes which were randomly sub-selected from the 50 available. On average, 4 experts analyzed each route.

2.2 Human expert evaluation

To collect human expert feedback, we developed a website that allowed chemists to input their opinions on presented retrosynthetic paths. The website interface shows retrosynthetic paths, with the option to select categories for each reaction and route, as well as to enter raw text commentary on each reaction and route. Experts were allowed to select more than one option for reactions and only one for routes.

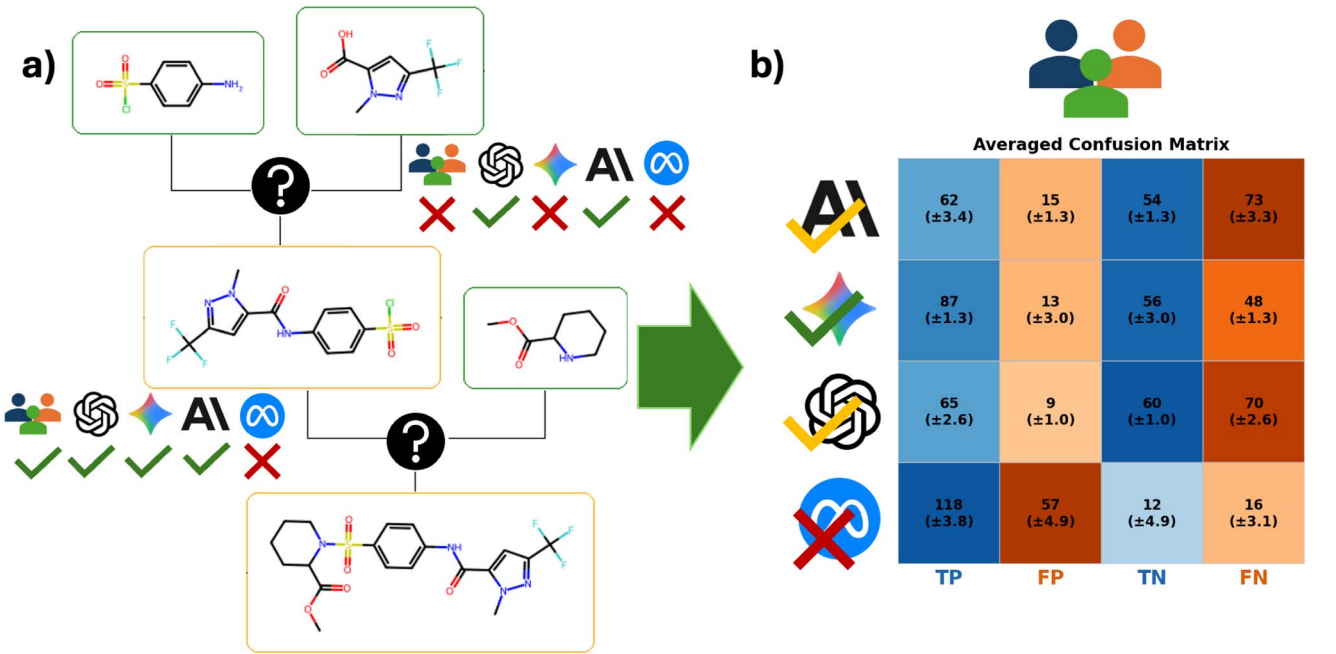


Fig. 1 Panel (a) illustrates chemists and per-LLM evaluation of the feasibility of the reactions in 50 retrosynthetic trees of the 50 molecules from JMC. Panel (b) illustrates the analysis of the majority voting of chemists and LLMs with Gemini 3.1 Pro performing better than Claude Opus 4.8, Llama 3.1 70B and GPT-5.5.

We used eight different categories for reaction-level evaluation (Table 1) and five categories for route-level assessment (see SI). The reaction-level categories are more precise in definition and leave for less room for individual interpretation, while the route-wise definitions are more general.

2.3 LLM evaluation

We evaluated four LLMs here: Claude Opus 4.8, GPT-5.5, Llama 3.1 70B and Gemini 3.1 Pro. We used these models with the default temperature, and the default reasoning budget, as reported in the GitHub repository metadata on models. We queried the LLMs *via* LangChain[49] using the same prompt; the prompt instructed each LLM to act as an expert organic chemist with extensive experience in medicinal chemistry, total synthesis, and reaction methodology, and tasked each model with evaluating a single retrosynthetic tree (AiZynthFinder JSON-formatted) provided at input. The models were instructed to follow a strict JSON format in the output. They were also asked in the prompt to give an explanation for selecting a given choice and provide confidence from 0 to 100. Out-of-vocabulary categories were mapped according to a predefined mapping dictionary provided in the GitHub repository; responses for which no unambiguous mapping was possible were excluded. Malformed outputs were dropped from the raw data during preprocessing.

2.4 Statistical analysis & agreement metrics

For the statistical analysis, we used majority voting of the 17 human experts' responses and compared results to majority voting across all four LLMs' outputs. The resolution of the majority voting was established the following way: (1) if the majority choose one category, then this category is the category of choice; (2) if there are different categories and no majority, then we look at the majority sentiment (based on count) and choose this sentiment as the majority sentiment, then choose the category with the maximal votes within this sentiment (if there is an equal amount, then we choose the worse category); (3) if there is a tie at the sentiment-level, then we assume a pessimistic tie, assuming that the general sentiment is negative. One expert selecting 2 categories contributes 2 distinct votes (one per category), exactly as if two experts had each selected one category, which allows them to flag multiple negative issues, though the usage of multi-selection was sparse. We used Matthews Correlation Coefficient (MCC)[50] as the primary measure of agreement between LLM and human expert sentiment classifications, selected for its robustness to class imbalance. Uncertainty was quantified by computing MCC independently for each of the 4 independent repeats and reporting mean ± SD with 95% bootstrap confidence intervals ($n = 10,000$ resamples). Pairwise differences between models were assessed using an exact permutation test enumerating all $\binom{8}{4} = 70$ arrangements of repeat labels under the null hypothesis (minimum achievable one-sided $p = 1/70 \approx 0.014$). Inter-rater agreement between LLM pairs and between individual evaluators was quantified using Cohen's Kappa.[51] Differences in confidence score distributions between correct and erroneous LLM predictions were tested with the two-sided Mann–Whitney U test.[52]

Table 1 The eight reaction categories used in the questionnaire and as definitions for the LLM prompt. The points are a heuristic value assigned to each category, to be used later for scoring

Category name	Category description	Points	Sentiment
Reaction feasible, all good	The proposed transformation is well-established and likely to proceed as expected and written	5	Positive
Reaction feasible, unexpected disconnection	The reaction would work but represents non-typical bond disconnection which is inventive and effective, but potentially less precedent	5	Positive
Non-optimal reagent	The exact reactant may not work in this reaction, but close analogues exist for this transformation with the same bond disconnection	4	Negative
Unnecessary step	The transformation adds complexity to the synthesis or may not work at all, but is unnecessary as its product is commercially available	4	Negative
Protecting group strategy is wrong/non-optimal	The protecting group selection is flawed, unnecessary, or inefficient	3	Negative
Selectivity (regio-, stereo-, chemo-) issues	The reaction would face significant regio-, stereo-, or chemoselectivity challenges	2	Negative
Functional group compatibility problems	Incompatible functional groups are present in the reactant molecule that would interfere with the desired transformation	2	Negative
Unlikely disconnection	The proposed transformation is chemically impossible	1	Negative

3 Results and discussion

3.1 *Vox Populi, Vox Dei*

In this section, we dive into the expert responses collected with the interface and analyze how human experts agree with each other when grading a reaction or route. We also illustrate complex cases when the experts disagree due to a different level of domain expertise for a given reaction type or peculiarities in personal interpretation of a given category. In each case, we measured the level of agreement as the percentage of experts who were in agreement with the majority category based on reaction-wise majority voting.

We analyzed how experts agree with each other when grading both at the reaction- and route-level. The experts showed a high degree of agreement on reactions (Fig. 2b) and a lower degree of agreement on routes due to more vague route score definitions relative to the reaction score definitions (Table S2). In previous work, we investigated whether assigning the route goodness to the worst grade presented by any of the expert chemists could be a meaningful metric to evaluate the quality of the route.[45] We measured the agreement by the percentage of experts agreeing with the majority category selected when grading a given reaction.

We can also see in Fig. 2a that the majority class for the selected reaction-level scores was "Reaction feasible"; 55.6% of human expert votes fall into this category, meaning that AiZynthFinder solution trees often included easy-to-execute reactions which seemed plausible to many experts. The category "Selectivity issues" had 11.5% which shows that it is not rare for AiZynthFinder to generate reactions that have

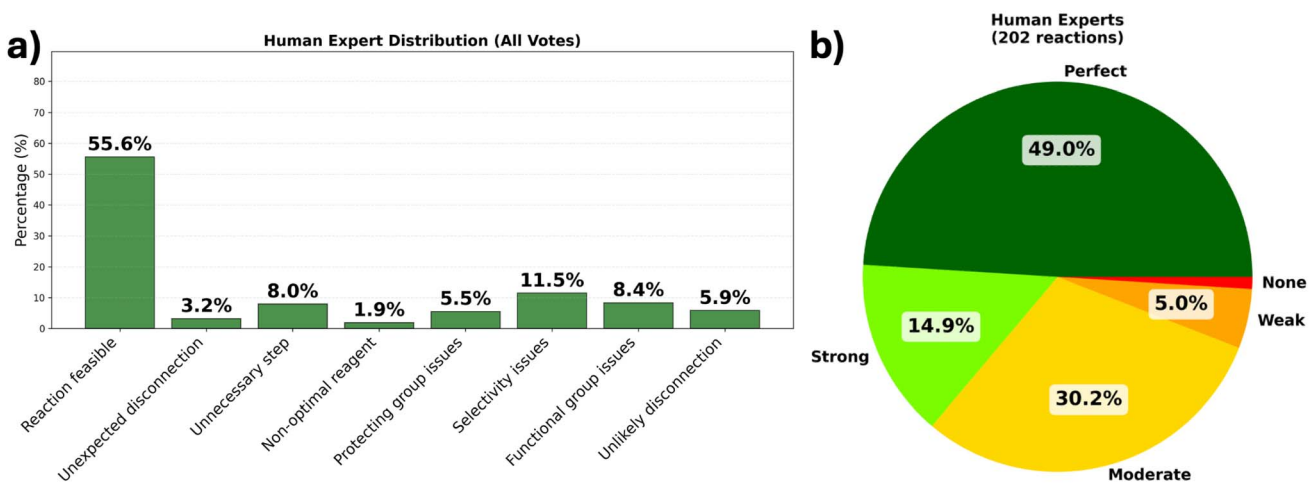
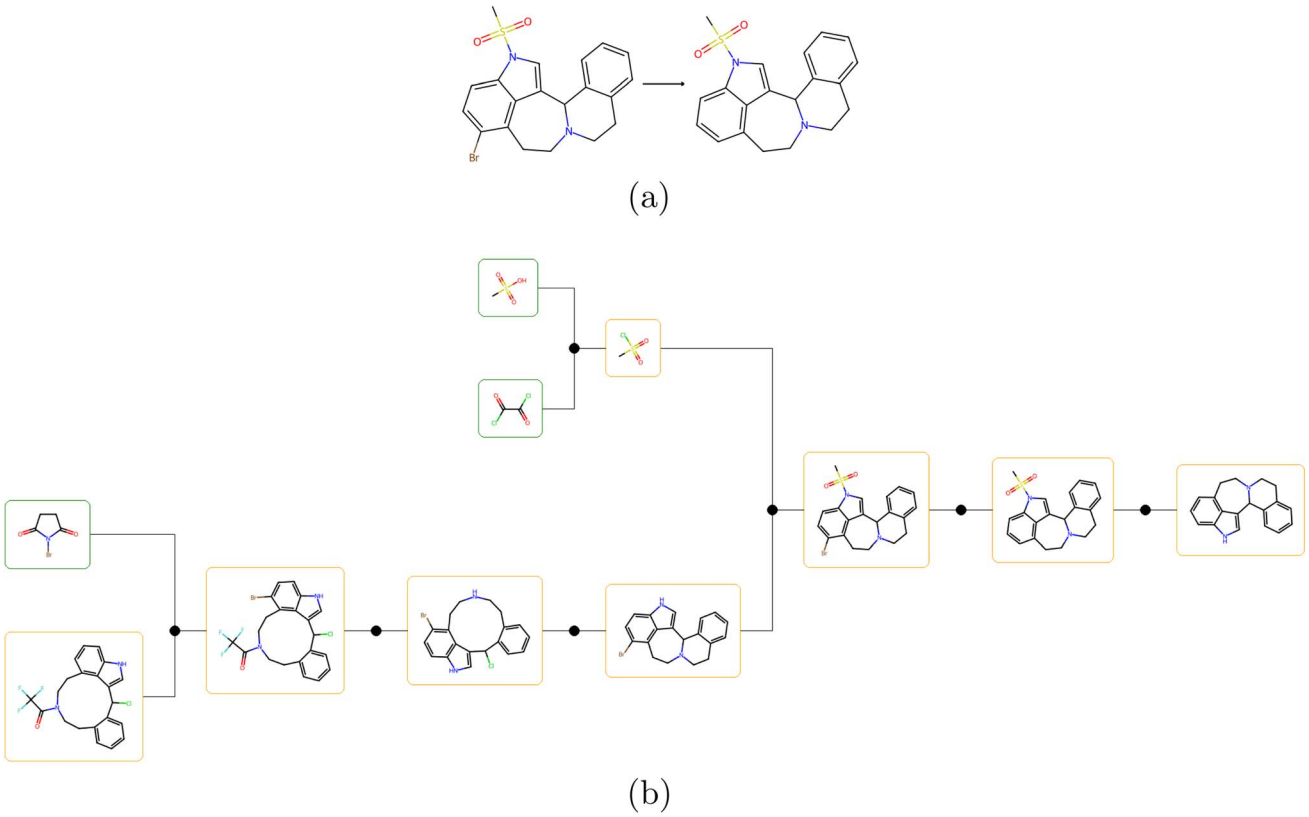


Fig. 2 (a) Distribution of selected reaction categories among human experts. Majority category selected by experts is "Reaction feasible" with 55.6%. The bar chart is sorted by the level of positivity of a given category. We sorted the categories in the bar plot by the points assigned to the category, from best to worst. (b) Human experts' sentiment agreement for reactions. 202/204 reactions had more than one expert evaluating them. Agreement in % of experts who had an agreement with the majority category selected for a given reaction. Perfect – 100%, strong – 75% or more, moderate – 50% or more, weak – 30% or more and none less than 30%.

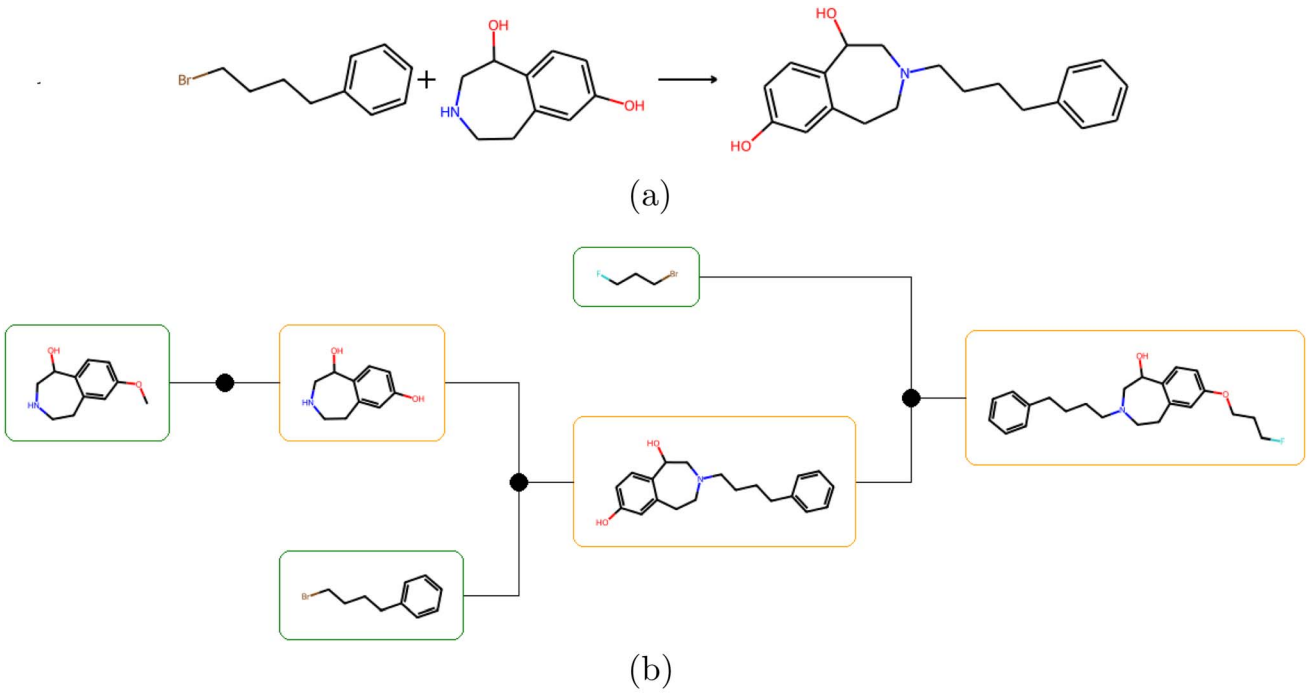


Unlikely disconnection	Then why was bromide installed in the first place!?
Unnecessary step	Then why was bromide installed in the first place!?
Protecting group strategy is wrong / non-optimal	Then why was bromide installed in the first place!?
Protecting group strategy is wrong / non-optimal	None
Unnecessary step	I don't see the need for bromination and debromination in this route - it doesn't look to be a blocking group?
Reaction feasible, all good	Yes, also works without the protecting group
Reaction feasible, all good	None

Fig. 3 Example of experts selecting different categories for representing the same opinion. (a) An example reaction on which experts had poor agreement with the available categories according to their raw text commentary. Out of context, this reaction could appear as a regular debromination reaction; however, within the context of the full route, illustrated in (b), this reaction step is not necessary. (b) The full route corresponding to the presented example. The figure was generated with AiZynthFinder built-in functions. The reaction discussed in (a) occurs before the last step of the synthesis going in the forward direction. Note that the route was not fully solved (one of the leaf nodes is not an available building block).

competing sites in experts' opinion, as well as 8.4% for "Functional group issues" which indicates that sometimes there were incompatible functional groups for a given reaction. 8.0% for "Unnecessary step" upon a deeper inspection revealed that for the majority of the cases AiZynthFinder generated compounds that were supposed to be in stock but were tagged as not in stock in the stock solutions file.

3.1.1 When do experts disagree, and why? To investigate the disagreement we analyzed the raw expert comments on disagreed reactions. We observed that, sometimes, a low level of agreement stemmed from distinct interpretations of a given reaction within a route. Some patterns of disagreement were often seen for the ratings "Unnecessary step" and "Reaction feasible, all good"; here, chemists were occasionally labeling a truly unnecessary step (*e.g.*, generating a reagent that is available in stock) as feasible despite that it is unnecessary. This is best illustrated *via* an example in Fig. 5. Another interesting example of expert disagreement includes a bromination reaction, illustrated in Fig. 3; here, we found out that experts were actually agreeing on the general idea but not selecting the category that fits best due to difference between their personal interpretations of the categories.



Reaction feasible, all good	None
Reaction feasible, all good	None
Reaction feasible, all good	None
Reaction feasible, all good	None
Selectivity (regio-, stereo-, chemo-) issues	None
Selectivity (regio-, stereo-, chemo-) issues	Phenol- vs. N-alkylation. More selective via reductive amination
Selectivity (regio-, stereo-, chemo-) issues	Step order swap: remove metoxy after Sn2 to solve selectivity issue
Protecting group strategy is wrong / non-optimal	Step order swap: remove metoxy after Sn2 to solve selectivity issue

Fig. 4 An example route where experts demonstrated strong disagreement in certain reaction-level categories, possibly due to some of the experts having domain knowledge on this specific reaction type that other experts may not have had. (a) An example reaction on which experts had poor agreement. This reaction is a regular bromo *N*-alkylation reaction; however, within the context of the full route, illustrated in (b), some chemists wrote that this reaction should be swapped with the previous step. (b) The full route corresponding to the presented example. The figure was generated with AiZynthFinder built-in functions.

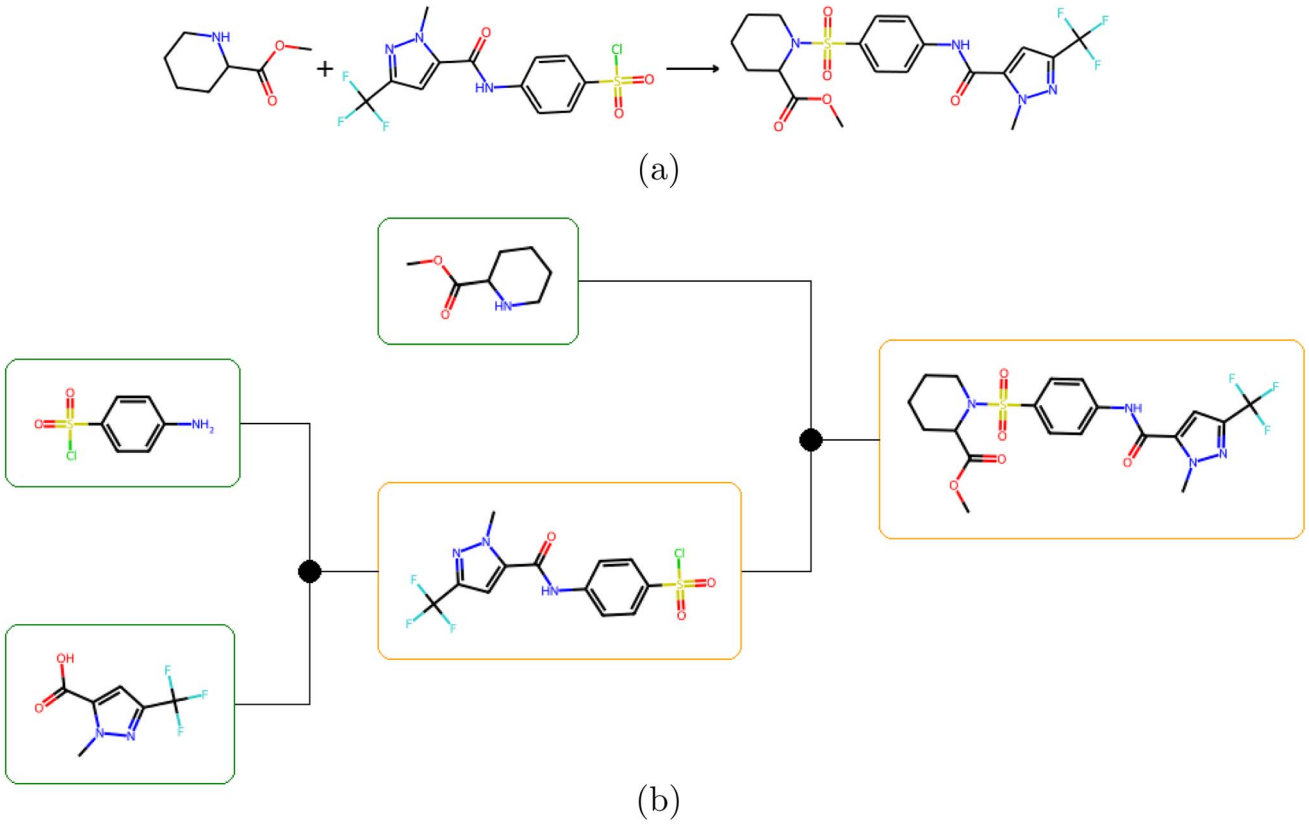
There were also instances of genuine disagreement. In some reactions, certain experts judged that there were no obvious problems, while others flagged potential issues based on deeper experience in specific chemistry domains, as shown in Fig. 4. At times, experts selected a variety of different options, yet closer inspection revealed they were converging on the same underlying concern. For example, in Fig. 5, we show how experts questioned possible self-polymerization of the starting compound, expressing that concern through different categories. Nevertheless, the cases presented above reflect some of the most extreme disagreements; overall, responses showed a strong tendency toward convergence in expert opinions.

3.2 LLMs as a consortium of experts

The investigation of LLMs' reasoning and recognition capabilities was conducted by testing whether re-querying LLMs with the same prompt and under the same set of parameters can be used to simulate a consortium of experts, with each query response treated as a new artificial expert, in a similar fashion to different human experts grading the same reaction. Because the nature of LLMs is stochastic, the same parameters and temperatures can yield different outputs. With this in mind, we also investigated stability in terms of sentiment agreement for each LLM, measuring how often the answer for the same prompt and the same reaction deviated with each new query.

To assess run-to-run variability, we conducted four independent repeat experiments, each comprising four evaluations of every retrosynthetic tree. Thus, each model evaluated every tree 16 times in total.

We assessed how stable the re-querying outputs were in Fig. 6. The majority of reactions received perfect or strong agreement across all LLMs, with GPT-5.5 showing the highest perfect agreement (59.3 $\pm$ 0.4%) and Llama 3.1 70B the most fragmentation. Human experts show a notably higher proportion of Moderate agreement (30.2%) compared to LLMs (14–



Reaction feasible, unexpected disconnection	The starting material might spontaneously polymerize, but I think the right conditions could enable selective amide coupling
Problems with reaction type and functional group compatibility	The starting material might spontaneously polymerize, but I think the right conditions could enable selective amide coupling
Selectivity (regio-, stereo-, chemo-) issues	Could have some self-reaction with sulfonyl chloride and amine?
Problems with reaction type and functional group compatibility	Could have some self-reaction with sulfonyl chloride and amine?
Reaction feasible, all good	None
Non-optimal reagent	None
Unlikely disconnection	None

Fig. 5 Example of experts showing a strong disagreement in categories selected and having non-obvious agreement. (a) An example reaction on which experts had poor agreement. This is a sulfonamide Schotten–Baumann reaction. (b) The full route corresponding to the presented example. The figure was generated with AiZynthFinder built-in functions.

22%), consistent with greater diversity of expert opinion on tricky reactions. Also we observe that no single LLM demonstrates an overwhelming level of internal consistency, which was more prominent in earlier models.

We observed that this moderate internal consistency comes with a moderate distribution of categories selected across all reruns and repeats, as one could see in Fig. 7. Llama is distinctively skewed toward over-selecting the "Reaction feasible" category, whereas all the others gave much more diverse rankings with Claude and GPT-5.5 selecting more "Selectivity issues" category compared to human experts with both over 20% when for human experts this category reaches 11.6%.

The category-level confusion matrices reveal the nature of disagreements across models in Fig. 9. Gemini 3.1 Pro, GPT-5.5, and Claude Opus 4.8 distributed their errors across several specific problem categories, including selectivity issues, functional-group incompatibilities, and unnecessary steps. This pattern is consistent with the models responding to several distinct chemical concerns rather than assigning all problematic reactions to a single category.

Llama 3.1 70B stands in contrast, showing a pronounced positive bias, with the majority of predictions collapsing into "Reaction feasible, all good" regardless of the human ground truth category. This pattern suggests that Llama's output variability does not arise from chemical reasoning but from instability in format and category selection. This model was also most prone to malformed outputs.

We examined whether the models' reported confidence scores, ranging from 0 to 100, discriminated between correct and incorrect predictions (Fig. S16). No significant difference

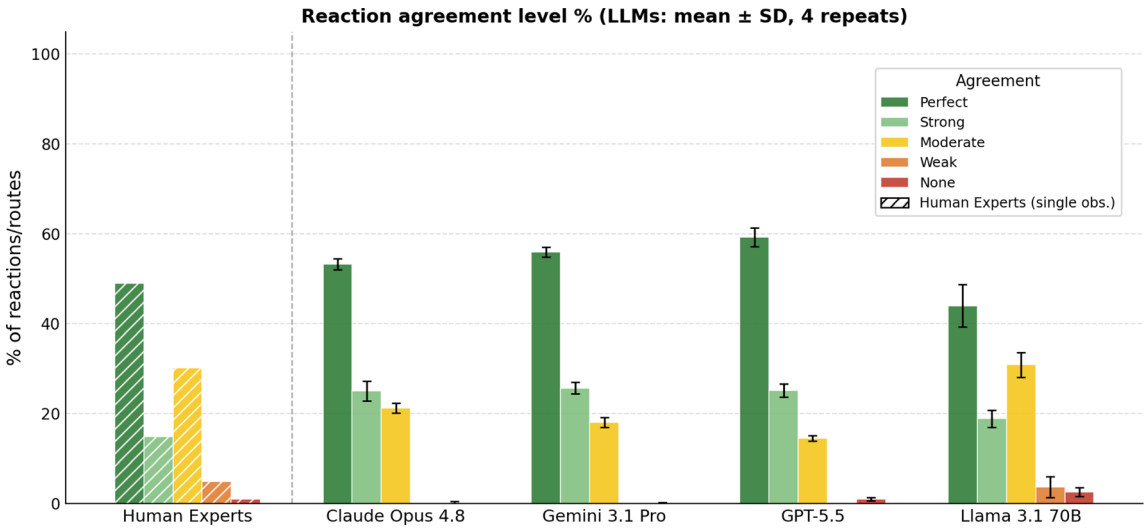


Fig. 6 Internal consistency of LLM evaluators and human experts when assigning reaction-level feedback categories. For each LLM, agreement level per reaction was determined from 4 reruns within each of 4 independent repeat runs; bars show mean ± SD across repeats. Human expert bars (hatched) represent a single observation from the full expert pool and carry no repeat-level uncertainty estimate.

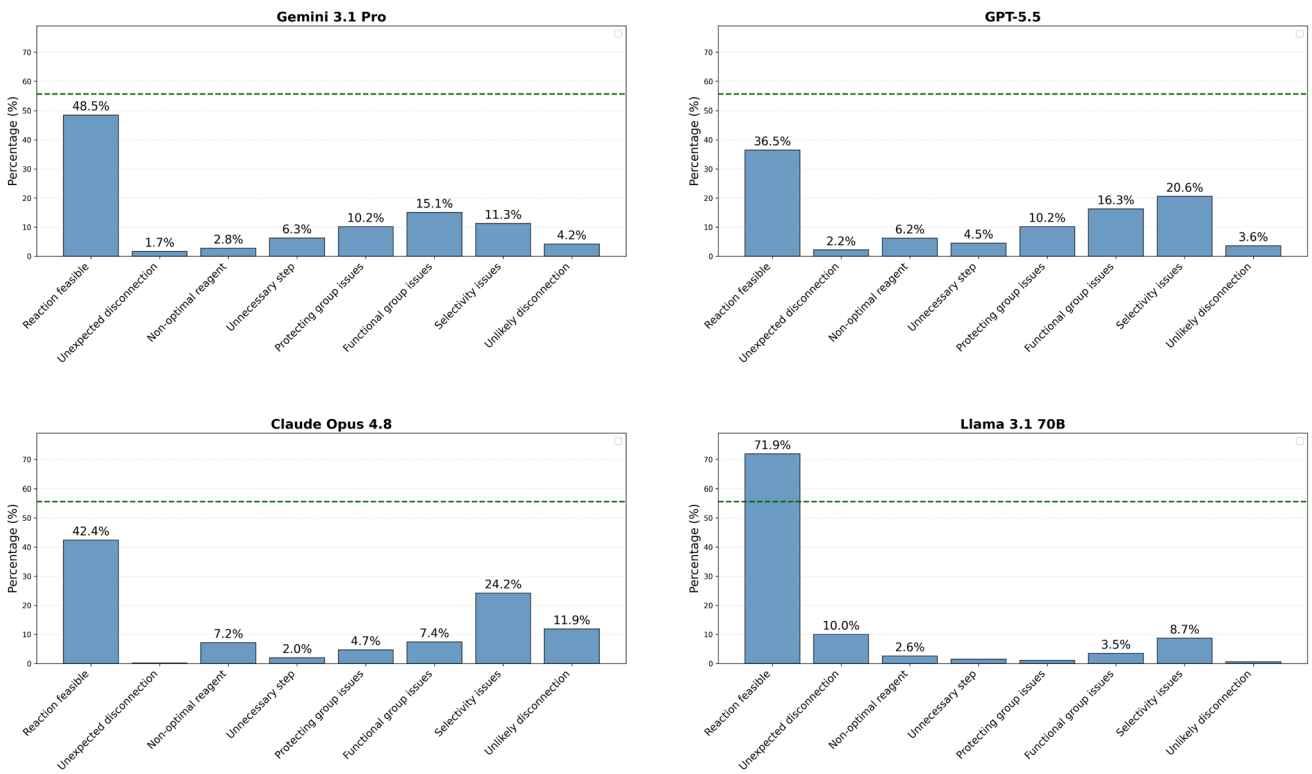


Fig. 7 Histogram of the selected categories for all re-runs and repeats of the LLMs, aggregated across re-queries and all reactions. We observe that Llama is over-selecting the "Reaction feasible" category compared to human experts (55.6%, green dashed line). The bar chart is sorted by the level of positivity of a given category (more positive categories on the left).

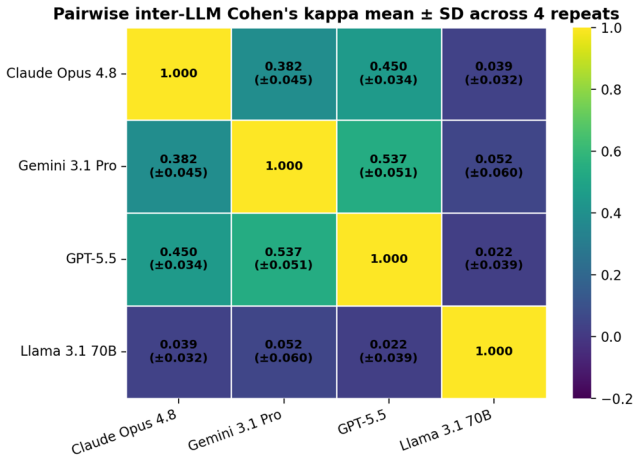


Fig. 8 Pairwise inter-LLM agreement measured using Cohen's κ, averaged over four independent repeats.

was observed for Gemini 3.1 Pro, GPT-5.5, or Llama 3.1 70B. Claude showed an approximately seven-point difference, although the two distributions overlapped substantially, limiting its practical utility. Confidence scores were also not associated with the level of human-expert agreement for a given reaction (Fig. S17). Thus, self-reported confidence was not a reliable indicator of correctness in this setting.

We then collapsed the confusion matrices to positive *vs.* negative sentiment (Fig. 10), selecting MCC as the metric given the class imbalance and our interest in an LLM's ability to correctly highlight problematic steps. In unadjusted one-sided pairwise permutation tests all six model pairs yielded $p \leq 0.029$; however, given that only 4 independent repeats were available and six comparisons were performed simultaneously, these results should be interpreted as exploratory. The bootstrap confidence intervals and the consistent direction of differences across all repeats nonetheless support the ranking

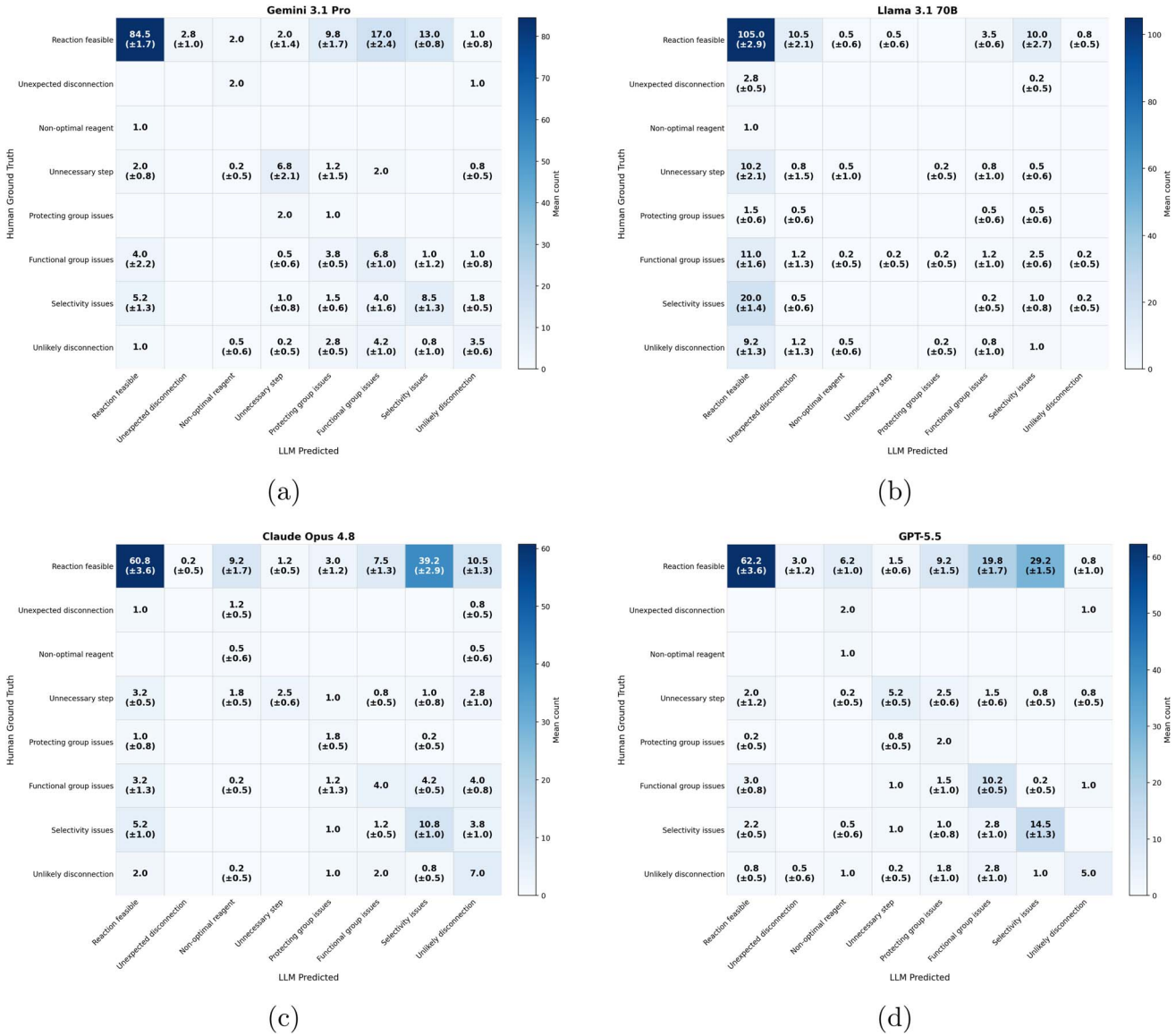


Fig. 9 Averaged over 4 repeats confusion matrices with the counts of the categories for the outputs of the four LLMs explored herein compared to human experts, showing how Llama (b) tends to rarely categorize reactions negatively, whereas Claude Opus (c) and GPT-5.5 (d) tend to have a slightly negative bias towards "Selectivity issues" category and Gemini 3.1 (a) tends to show a better agreement with expert chemist rankings.

of model performance. Gemini 3.1 Pro achieves the highest agreement with human expert consensus (MCC 0.43 ± 0.044), significantly outperforming GPT-5.5 (MCC 0.351 ± 0.028, $p = 0.029$), while Claude Opus 4.8 (MCC 0.241 ± 0.024) and Llama 3.1 70B (MCC 0.072 ± 0.102) show progressively lower correlation. The large standard deviation for Llama 3.1 70B reflects its unstable behavior across repeats.

To further assess whether there are overlaps in judgments of the models, we performed analysis of the averaged over 4 repeats Cohen's Kappa. The pairwise inter-LLM kappa matrix (Fig. 8) reveals that Claude Opus 4.8, Gemini 3.1 Pro, and GPT-5.5 form a coherent cluster of moderate agreement ($\kappa = 0.38$–0.54, SD ≤ 0.05 across repeats), indicating that these models make substantially similar judgments on the same reactions across independent runs. Llama 3.1 70B shows near-zero kappa with all three, confirming that its predictions are effectively independent of the consensus formed by the other models. The consistency of the SD values across repeats (all $\leq$0.06) shows that these inter-model agreement patterns are stable properties of the models rather than run-to-run artifacts. The fact that three out of four models form a coherent cluster, despite being independently trained, provides slight evidence that their shared judgments reflect internalized chemical signal rather than coincidental alignment.

We also examined the model commentaries for the three cases discussed in Section 3.1.1. For the first example (Fig. 3), GPT-5.5, Claude Opus 4.8, and Gemini 3.1 Pro generally recognized the transformation and considered it in the context of the full route. Across runs, Claude assigned several categories, including "Reaction feasible, all good", "Unlikely disconnection", and "Selectivity issues", whereas Gemini 3.1 Pro and GPT-5.5 most often selected "Unnecessary step" or "Reaction feasible, all good". Several responses explicitly questioned why bromine had been introduced only to be removed later. Llama 3.1 70B rarely identified this route-level issue.

For the second example (Fig. 4), Claude Opus 4.8, GPT-5.5, and Gemini 3.1 Pro identified the potential competition between *O*- and *N*-alkylation in most responses. They frequently proposed milder conditions or reductive amination as a more selective alternative, but rarely suggested reversing the order of the steps. Llama 3.1 70B rarely identified the selectivity concern.

For the third example (Fig. 5), Claude Opus 4.8, GPT-5.5, and Gemini 3.1 Pro recognized the transformation as a sulfonamide-forming Schotten–Baumann reaction and raised concerns about possible self-reaction or polymerization of the starting material. Claude sometimes suggested reordering the steps, whereas Gemini 3.1 Pro and GPT-5.5 suggested alternative starting materials. Llama 3.1 70B generally did not flag this issue.

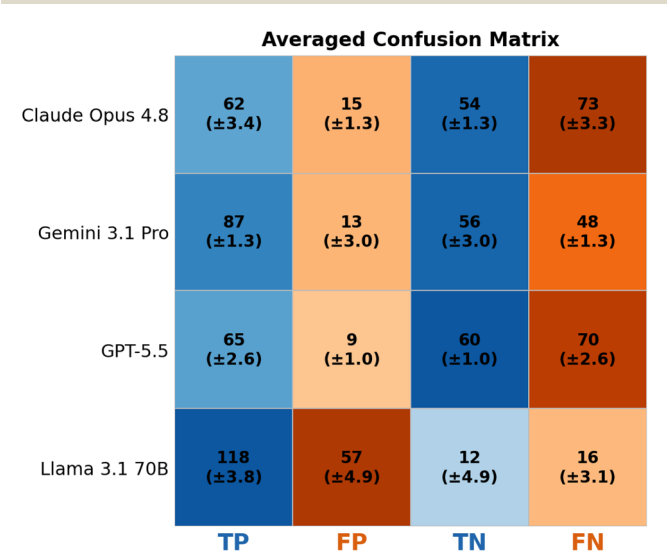


Fig. 10 Matrix illustrating averaged reaction counts ± SD for the majority sentiment vote of LLMs ($n = 4$ re-runs per LLM and $n = 4$ for repeats) compared to the majority of human expert sentiment votes, and organized by true positives (TN), false positives (FN), false negatives (FP), and true negatives (TP). Mean MCC ± SD: GPT-5.5 = 0.3508 ± 0.0281, Claude Opus 4.8 = 0.2407 ± 0.0244, Llama 3.1 70B = 0.0719 ± 0.1018, Gemini 3.1 Pro = 0.43 ± 0.0438.

4 Conclusion and future outlook

This work has two main conclusions and contributions. First, we have demonstrated that it is possible to use human expertise to grade the feasibility of reactions in retrosynthetic paths as a baseline consensus. Second, we have shown that some general-purpose LLMs can partially approximate human-expert consensus when evaluating the reaction-level feasibility of computer-generated retrosynthetic plans. Among the evaluated models, Gemini 3.1 Pro achieved the highest sentiment-level agreement with human experts (mean MCC = 0.43) and produced the most balanced category distribution. GPT-5.5 and Claude Opus 4.8 showed stronger negative biases, whereas Llama 3.1 70B showed a pronounced positive bias and substantially lower agreement. Compared with the models assessed in our preprint, the newer proprietary models achieved higher and more robust agreement on this benchmark.

We believe this work will pave the way to more research on synergies between HITL and LLM workflows. Future directions include a more diverse questionnaire for human experts with a greater variety of molecules and including more ambiguous solutions by retrosynthesis-planning tools. Also designing more robust categorization of the possible reactivity concerns that would leave no room for personal interpretation is necessary for future works in this direction.

Automatic prompt engineering and optimization were outside the scope of this study. Other possible future work directions could investigate a reinforcement learning-like approach for route evaluation, where LLMs evaluate the routes and consult human experts as necessary to tune the prompt in a way that leads to better agreement with human experts' opinions. Another possible future work could be on assessing whether a visual component of the route visualizations could improve LLM judgments of the feasibility of the reactions in this route and route as a whole.

Finally, automated grading can save expert time by filtering large candidate sets, but it should not replace human oversight for decisions with critical safety or scale-up implications. Clear

reporting of model limits and uncertainty is necessary, and any deployment should include safeguards that re-route high-risk cases for expert review.

Author contributions

Varvara Voinarovska performed the main research activities, collected and curated the dataset, analyzed the results, prepared the visualizations, and wrote the original draft of the manuscript. Samuel Genheden and Rocío Mercado contributed through conceptual input, critical discussion, supervision, and manuscript review and editing. Mikhail Kabeshov contributed to coordination of the people and activities required for dataset creation, and contributed to supervision and manuscript review and editing.

Conflicts of interest

The authors declare no competing interests.

Data availability

All source code, prompts, and the dataset of molecules from JMC, as well as full AiZynthFinder trees can be found at https://github.com/v-in-cube/HITLLLMs. Supporting datasets can be found at Zenodo at https://doi.org/10.5281/zenodo.21839555.

Supplementary information (SI): system prompt, routes categories, plots for general route feedback analysis and extra convergence analysis. See DOI: https://doi.org/10.1039/d6dd00266h.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation. We thank all chemists who participated in responding to the questionnaire and contributed valuable comments on the reaction categories.

References

1 E. J. Corey, The Logic of Chemical Synthesis: Multistep Synthesis of Complex Carbogenic Molecules (Nobel Lecture), *Angew Chem. Int. Ed. Engl.*, 1991, **30**, 455–465, DOI: 10.1002/anie.199104553.

2 C. M. Gothard, S. Soh, N. A. Gothard, B. Kowalczyk, Y. Wei, B. Baytekin and B. A. Grzybowski, Rewiring Chemistry: Algorithmic Discovery and Experimental Validation of One-Pot Reactions in the Network of Organic Chemistry, *Angew. Chem., Int. Ed.*, 2012, **51**, 7922–7927, DOI: 10.1002/anie.201202155.

3 C. W. Coley, D. A. Thomas III, J. A. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, *et al.*, A robotic platform for flow synthesis of organic compounds informed by AI planning, *Science*, 2019, **365**, eaax1566.

4 R. Mercado, S. M. Kearnes and C. W. Coley, Data Sharing in Chemistry: Lessons Learned and a Case for Mandating Structured Reaction Data, *J. Chem. Inf. Model.*, 2023, **63**, 4253–4265, DOI: 10.1021/acs.jcim.3c00607.

5 M. P. Maloney, C. W. Coley, S. Genheden, N. Carson, P. Helquist, P.-O. Norrby and O. Wiest, Negative Data in Data Sets for Machine Learning Training, *J. Org. Chem.*, 2023, **88**, 5239–5241, DOI: 10.1021/acs.joc.3c00844.

6 V. Voinarovska, M. Kabeshov, D. Dudenko, S. Genheden and I. V. Tetko, When Yield Prediction Does Not Yield Prediction: An Overview of the Current Challenges, *J. Chem. Inf. Model.*, 2023, **64**, 42–56, DOI: 10.1021/acs.jcim.3c01524.

7 A. Thakkar, T. Kogej, J.-L. Reymond, O. Engkvist and E. J. Bjerrum, Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain, *Chem. Sci.*, 2020, **11**, 154–168.

8 T. J. Struble, J. C. Alvarez, S. P. Brown, M. Chytil, J. Cisar, R. L. DesJarlais, O. Engkvist, S. A. Frank, D. R. Greve, D. J. Griffin, *et al.*, Current and future roles of artificial intelligence in medicinal chemistry synthesis, *J. Med. Chem.*, 2020, **63**, 8667–8682.

9 E. J. Corey and W. T. Wipke, Computer-Assisted Design of Complex Organic Syntheses: Pathways for molecular synthesis can be devised with a computer and equipment for graphical communication, *Science*, 1969, **166**, 178–192, DOI: 10.1126/science.166.3902.178.

10 P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas and A. A. Lee, Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction, *ACS Cent. Sci.*, 2019, **5**, 1572–1583.

11 D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36, DOI: 10.1021/ci00057a005.

12 M. H. Segler, M. Preuss and M. P. Waller, Planning chemical syntheses with deep neural networks and symbolic AI, *Nature*, 2018, **555**, 604–610.

13 S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist and E. Bjerrum, AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning, *J. Cheminf.*, 2020, **12**, 70, DOI: 10.1186/s13321-020-00472-1.

14 Z. Tu, S. J. Choure, M. H. Fong, J. Roh, I. Levin, K. Yu, J. F. Joung, N. Morgan, S.-C. Li, X. Sun, *et al.*, ASKCOS: open-source, data-driven synthesis planning, *Acc. Chem. Res.*, 2025, **58**, 1764–1775.

15 K. Maziarz, A. Tripp, G. Liu, M. Stanley, S. Xie, P. Gaiński, P. Seidl and M. H. Segler, Re-evaluating retrosynthesis algorithms with syntheseus, *Faraday Discuss.*, 2025, **256**, 568–586.

16 T. Akhmetshin, D. Zankov, P. Gantzer, D. Babadeev, A. Pinigina, T. Madzhidov and A. Varnek, SynPlanner: an end-to-end tool for synthesis planning, *J. Chem. Inf. Model.*, 2024, **65**, 15–21.

17 A. D. White, G. M. Hocky, H. A. Gandhi, M. Ansari, S. Cox, G. P. Wellawatte, S. Sasmal, Z. Yang, K. Liu, Y. Singh and W. J. Peña Ccoa, Assessment of chemistry knowledge in

large language models that generate code, *Digital Discovery*, 2023, **2**, 368–376, DOI: 10.1039/d2dd00087c.

18 A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White and P. Schwaller, Augmenting large language models with chemistry tools, *Nat. Mach. Intell.*, 2023, **6**, 525–535, DOI: 10.1038/s42256-024-00832-8.

19 T. Guo, K. Guo, B. Nan, Z. Liang, Z. Guo, N. V. Chawla, O. Wiest and X. Zhang, *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Curran Associates Inc., New Orleans, LA, USA, 2023.

20 M. C. Ramos, C. J. Collison and A. D. White, A review of large language models and autonomous agents in chemistry, *Chem. Sci.*, 2024, **16**, 2514–2572, DOI: 10.48550/arxiv.2407.01603.

21 K. Jablonka, P. Schwaller, A. Ortega-Guerrero and B. Smit, Leveraging large language models for predictive chemistry, *Nat. Mach. Intell.*, 2024, **6**, 161–169, DOI: 10.1038/s42256-023-00788-1.

22 A. Bavaresco, R. Bernardi, L. Bertolazzi, D. Elliott, R. Fernández, A. Gatt, E. Ghaleb, M. Giulianelli, M. Hanna, A. Koller, A. F. T. Martins, P. Mondorf, V. Neplenbroek, S. Pezzelle, B. Plank, D. Schlangen, A. Suglia, A. K. Surikuchi, E. Takmaz and A. Testoni, LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks, *arXiv*, 2024, preprint, arxiv:2406.18403, DOI: 10.48550/arXiv.2406.18403.

23 H. Li, S. Sarkar, W. Lu, P. O. Loftus, T. Qiu, Y. Shee, A. E. Cuomo, J.-P. Webster, H. R. Kelly, V. Manee, S. Sreekumar, F. G. Buono, R. H. Crabtree, T. R. Newhouse and V. S. Batista, Collective intelligence for AI-assisted chemical synthesis, *Nature*, 2026, **651**, 107–115, DOI: 10.1038/s41586-026-10131-4.

24 K. K. L. Tharwani, R. Kumar, Sumita, N. Ahmed and Y. Tang, Large Language Models Transform Organic Synthesis From Reaction Prediction to Automation, *arXiv*, 2025, preprint, arXiv:2508.05427, DOI: 10.48550/arXiv.2508.05427.

25 A. M. Bran, T. A. Neukomm, D. Armstrong, Z. Jončev and P. Schwaller, Chemical reasoning in LLMs unlocks strategy-aware synthesis planning and reaction mechanism elucidation, *Matter*, 2026, **9**, 102812, DOI: 10.1016/j.matt.2026.102812.

26 Y. Zhang, Y. Han, S. Chen, R. Yu, X. Zhao, X. Liu, K. Zeng, M. Yu, J. Tian, F. Zhu, X. Yang, Y. Jin and Y. Xu, Large language models to accelerate organic chemistry synthesis, *Nat. Mach. Intell.*, 2025, **7**, 1010–1022, DOI: 10.1038/s42256-025-01066-y.

27 Z. Zhao, Z. Huang, J. Li, S. Lin, J. Zhou, F. Cao, K. Zhou, R. Ge, T. Long, Y. Zhu, Y. Liu, J. Zheng, J. Wei, R. Zhu, P. Zou, W. Li, Z. Cheng, T. Ding, Y. Wang, Y. Yan, *et al.*, SUPERChem: A Multimodal Reasoning Benchmark in Chemistry, *arXiv*, 2025, preprint, arXiv:2512.01274, DOI: 10.48550/arXiv.2512.01274.

28 Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, R. Zheng, X. Fan, X. Wang, L. Xiong, Y. Zhou, W. Wang, C. Jiang, Y. Zou, X. Liu and Z. Yin, The rise and potential of large language model based agents: a survey, *Sci. China Inf. Sci.*, 2025, **68**, 121101, DOI: 10.1007/s11432-024-4222-0.

29 H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, A. Anandkumar, K. Bergen, C. P. Gomes, S. Ho, P. Kohli, J. Lasenby, J. Leskovec, T.-Y. Liu, A. Manrai, D. Marks, *et al.*, Scientific discovery in the age of artificial intelligence, *Nature*, 2023, **620**, 47–60, DOI: 10.1038/s41586-023-06221-2.

30 A. Mirza, N. Alampara, S. Kunchapu, M. Ríos-García, B. Emoekabu, A. Krishnan, T. Gupta, M. Schilling-Wilhelmi, M. Okereke, A. Aneesh, M. Asgari, J. Eberhardt, A. M. Elahi, H. M. Elbeheiry, M. V. Gil, C. Glaubitz, M. Greiner, C. T. Holick, T. Hoffmann, A. Ibrahim, *et al.*, A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists, *Nat. Chem.*, 2025, **17**, 1027–1034, DOI: 10.1038/s41557-025-01815-x.

31 D. A. Boiko, R. MacKnight, B. Kline and G. Gomes, Autonomous chemical research with large language models, *Nature*, 2023, **624**, 570–578, DOI: 10.1038/s41586-023-06792-0.

32 A. Birhane, A. Kasirzadeh, D. Leslie and S. Wachter, Science in the age of large language models, *Nat. Rev. Phys.*, 2023, **5**, 277–280, DOI: 10.1038/s42254-023-00581-4.

33 N. Alampara, M. Schilling-Wilhelmi, M. Ríos-García, I. Mandal, P. Khetarpal, H. S. Grover, N. M. A. Krishnan and K. M. Jablonka, Probing the limitations of multimodal language models for chemistry and materials research, *Nat. Comput. Sci.*, 2025, **5**, 952–961, DOI: 10.1038/s43588-025-00836-3.

34 N. T. Runcie, C. M. Deane and F. Imrie, Assessing the Chemical Intelligence of Large Language Models, *arXiv*, 2025, preprint, arXiv:2505.07735, DOI: 10.48550/arXiv.2505.07735.

35 H. Li, X. Fang, Y. Li, C. Huang, J. Wang, X. Wang, H. Bai, B. Hao, S. Lin, H. Liang, L. Zhang and G. Ke, RxnBench: A Multimodal Benchmark for Evaluating Large Language Models on Chemical Reaction Understanding from the Scientific Literature, *J. Chem. Inf. Model.*, 2026, **66**, 5747–5756, DOI: 10.1021/acs.jcim.6c00286.

36 C. Bartmann, J. Schimunek, M. Ielanskyi, P. Seidl, G. Klambauer and S. Luukkonen, MolecularIQ: Characterizing Chemical Reasoning Capabilities Through Symbolic Verification on Molecular Graphs, *arXiv*, 2026, preprint, arXiv:2601.15279, DOI: 10.48550/arXiv.2601.15279.

37 Y. Wu, J. Ye, S. Zhang, L. Dai, Y. Bisk and O. Isayev, MolErr2Fix: Benchmarking LLM Trustworthiness in Chemistry via Modular Error Detection, Localization, Explanation, and Revision, *arXiv*, 2025, preprint, arXiv:2509.00063, DOI: 10.48550/arXiv.2509.00063.

38 L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez and I. Stoica, *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Curran Associates Inc., New Orleans, LA, USA, 2023.

39 H. Kim, T. Jeon, S. Choi, J. H. Hong, D. W. Jeon, G.-Y. Baek, G.-W. Kwak, D.-H. Lee, J. Bae, C. Lee, Y.-S. Kim, S.-J. Choi,

J.-S. Park, S. B. Cho and H. Cho, *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, Association for Computing Machinery, Seoul, Republic of Korea, 2025, pp. 1302–1312, DOI: 10.1145/3746252.3761359.

40 H. Ko, C. Lee, Y. R. Kim, R. Hormazabal, S. Han, S. Lim and S. Kim, RetroReasoner: A Reasoning LLM for Strategic Retrosynthesis Prediction, *arXiv*, 2026, preprint, arXiv:2603.12666, DOI: 10.48550/arXiv.2603.12666.

41 O.-H. Choung, R. Vianello, M. Segler, N. Stiefl and J. Jiménez-Luna, Learning chemical intuition from humans in the loop, *ChemRxiv*, 2023, preprint, DOI: 10.26434/chemrxiv-2023-knwnv-v2.

42 S. V. Sathyanarayana, S. D. Hiremath, S. Rahil Kirankumar, R. Panda, R. Jana, R. Singh, R. Irfan, A. Murali and B. Ramsundar, DeepRetro discovers retrosynthetic pathways through iterative large language model reasoning, *Sci. Rep.*, 2026, **16**(1), 8448, DOI: 10.1038/s41598-026-38821-z.

43 Y. Nahal, J. Menke, J. Martinelli, M. Heinonen, M. Kabeshov, J. P. Janet, E. Nittinger, O. Engkvist and S. Kaski, Human-in-the-loop active learning for goal-oriented molecule generation, *J. Cheminf.*, 2024, **16**, 138, DOI: 10.1186/s13321-024-00924-y.

44 H. H. Loeffler, J. He, A. Tibo, J. P. Janet, A. Voronov, L. H. Mervin and O. Engkvist, Reinvent 4: Modern AI–driven generative molecule design, *J. Cheminf.*, 2024, **16**, 20, DOI: 10.1186/s13321-024-00812-5.

45 G. Yujia, M. Kabeshov, T. H. D. Le, S. Genheden, G. Bergonzini, O. Engkvist and S. Kaski, An Expert-Augmented Deep Learning Approach for Synthesis Route Evaluation, *ChemRxiv*, 2025, preprint, DOI: 10.26434/chemrxiv-2024-tp7rh-v2.

46 J. Zhou and N. Thakkar, Leveraging LLM Feedback to Enhance Review Quality, *ICLR Blog*, 2025, accessed: 2025-02-18.

47 Y. Luo, Y. Li, O. Ogunyemi, E. Koski and B. E. Himes, Leveraging large language models for academic conference organization, *NPJ Digit. Med.*, 2025, **8**, 101.

48 L. Zhu, Y. Lai, J. Xie, W. Mou, L. Huang, C. Qi, T. Yang, A. Jiang, W. Gan, D. Zeng, *et al.*, Evaluating the potential risks of employing large language models in peer review, *Clin. Transl. Discovery*, 2025, **5**, e70067.

49 H. Chase, *LangChain*, 2022, https://github.com/langchain-ai/langchain.

50 B. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochim. Biophys. Acta, Protein Struct.*, 1975, **405**, 442–451, DOI: 10.1016/0005-2795(75)90109-9.

51 J. Cohen, A Coefficient of Agreement for Nominal Scales, *Educ. Psychol. Meas.*, 1960, **20**, 37–46, DOI: 10.1177/001316446002000104.

52 H. B. Mann and D. R. Whitney, On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, *Ann. Math. Stat.*, 1947, **18**, 50–60, DOI: 10.1214/aoms/1177730491.

