SYNREASON: ENHANCING SYNTHESIS REASONING VIA REINFORCEMENT LEARNING EXPERIMENTAL FEEDBACK

Elton Pan[1*], Thorben Prein[2,3,4,5*], Juno Nam[1], Xiaochen Du[1], Soojung Yang[1], Pengfei Cai[1] Jennifer L.M. Rupp[2,5], Rafael Gomez-Bombarelli[1], Elsa Olivetti[1]

[1]MIT, [2]TUM, [3]Munich Data Science Institute, [4]TUM.ai, [5]TUMint. Energy Research GmbH

[*] Equal contribution

ABSTRACT

Materials discovery efforts have produced millions of candidate inorganic compounds, yet identifying practical synthesis routes remains a major bottleneck. While the reasoning capabilities of LLMs are promising, their reasoning traces can result in poor synthesis prediction accuracy due to a misalignment with experimental data. We introduce ***R**einforcement **L**earning **E**xperimental **F**eedback* (**RLEF**), a simple RL post-training approach that samples multiple candidate completions per prompt, scores them with an *experimental alignment reward*, followed by group-normalized advantages and regularized updates with a KL penalty to a reference model. We demonstrate our method on materials precursor recommendation task: given a target composition, generate a ranked list of precursor sets that match experimental syntheses. Using RLEF, we develop SYNREASON, a synthesis planning LLM model that is capable of chemical reasoning. Across different model sizes, RLEF redefines the accuracy–speed Pareto front. A 4B model post-trained with RLEF outperforms even base models that are $8\times$ larger across all Top-K metrics. Interestingly, contrary to other RL post-training methods, RLEF shortens reasoning traces (instead of lengthening) while simultaneously improving performance. This *reasoning compression* suggests that RLEF induces more selective and task-relevant chemical reasoning in LLMs.

1 INTRODUCTION

The discovery of inorganic materials underpins modern technologies ranging from renewable energy to electronics. Large-scale computational exploration has produced millions of candidate compounds (*what* to synthesize) (Sriram et al., 2024; Merchant et al., 2023; Kim et al., 2021; Zhu et al., 2024; Merchant et al., 2023; Barroso-Luque et al., 2024; Saal et al., 2013; Zeni et al., 2023; Schmidt et al., 2024), but experimentally realizing these predictions remains a major bottleneck (*how* to synthesize) (Karpovich et al., 2023; Mahbub et al., 2020; Malik et al., 2021). Unlike organic synthesis—where retrosynthesis decomposes a target into multi-step transformations—inorganic synthesis is often a one-step process in which a *set* of precursors is reacted to form a target phase. The space of plausible precursor sets is large, and the lack of a unifying theory forces trial-and-error experimentation guided by incomplete intuition and costly physical modeling (Kononova et al., 2019; Bianchini et al., 2020).

We therefore focus on precursor recommendation for inorganic materials: given a target composition, produce a ranked list of precursor sets likely to result in a successful synthesis. Prior work has explored heuristic template completion (Kim et al., 2022), retrieval of similar known syntheses (He et al., 2023), and language-model-based synthesis prediction (Kim et al., 2024). Building on this progress, our work investigates whether modern reasoning-style LLMs can be improved via reinforcement learning post-training grounded in experimental recipes, yielding faster and more accurate synthesis planning. Our key contributions are as follows:

- **Reasoning models for materials synthesis planning**. We show that chemical reasoning is an essential part in achieving strong performance in materials synthesis planning.

Figure 1: **Reasoning models for materials synthesis planning** (a) Materials precursor recommendation task. Given a prompt that contains the target material, the LLM leverages its reasoning capabilities to output a reasoning trace followed by a diverse set of synthesis precursors. (b) Top-1 exact match accuracy vs. model size for Qwen3 (Yang et al., 2025) models. LLMs exhibit improved performance as they output reasoning traces, indicating the potential of chemical reasoning for synthesis planning task.

- **Reinforcement learning to enhance synthesis reasoning.** We introduce a RL post-training approach, reinforcement learning experimental feedback (RLEF) in Fig. 2a with an experimental alignment reward that aligns LLM completions with experimental synthesis.
- **Improved accuracy–efficiency tradeoffs.** We show that RLEF redefines a new accuracy–speed Pareto front for materials synthesis (Fig. 2b), while simultaneously performing a reasoning compression (Fig. 2d) resulting in a more task-selective reasoning.

2 Methods

2.1 Reinforcement Learning Experimental Feedback

We post-train a base language model π_{ref} into a synthesis reasoning model π_θ using group relative policy optimization (GRPO) (Shao et al., 2024) (Fig. 2a) with an *experimental alignment reward* (Fig. 2c).

Rollout generation. For each prompt x containing a target composition (Appendix B.4), the policy samples a group of G candidate completions $\{o_i\}_{i=1}^{G} \sim \pi_\theta(\cdot \mid x)$, each including a reasoning trace and a structured list of K candidate precursor sets. Each rollout o_i is parsed into a list of precursor sets.

Experimental alignment reward. We define a scalar reward $r_i = R(o_i) = 1/\text{rank}_{\text{exp}}$, where rank_{exp} is the rank of an experimentally verified precursor set in the list of generated precursor sets. Rollouts that place the ground-truth, experimental precursor set higher in the ranked list receive a higher reward. Fig. 2c illustrates this experimental-alignment signal. This strict, scalar reward simultaneously incentivizes three aspects of the LLM output: *(1) Chemistry:* The precursors must match exactly. Any deviation from ground-truth (e.g., additional hydrates $\cdot x\text{H}_2\text{O}$) is not tolerated. *(2) Cardinality:* Correct number of precursors in the ground truth set must be predicted (predicting an extra precursor is not tolerated). *(3) Ranking:* If (1) and (2) are fulfilled, the ground truth set should be ranked higher than other recommendations. This is due to the one-to-many relationship between target and synthesis (i.e., a target can be synthesized via multiple recipes) (Pan et al., 2025; 2024).

Group-normalized advantage. To reduce variance, we compute a normalized advantage within each group as $A_i = \frac{r_i - \mu_r}{\sigma_r + \epsilon}$, where $\mu_r = \frac{1}{G}\sum_{j=1}^{G} r_j$ and $\sigma_r = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(r_j - \mu_r)^2}$.

Regularized GRPO objective. To prevent reward hacking and preserve the base model distribution, we regularize toward a frozen reference policy π_{ref} (Fig. 2a). Overall, the objective is to maximize

$$\max_{\theta} \; \mathbb{E}_{x,\, o_{1:G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G} \min\Big(r_i A_i,\; \text{clip}(r_i, 1-\epsilon, 1+\epsilon)\, A_i \Big) \;-\; \beta\, \widehat{D}_{\text{KL}}\big(\pi_\theta \,\|\, \pi_{\text{ref}}\big) \right], \quad (1)$$

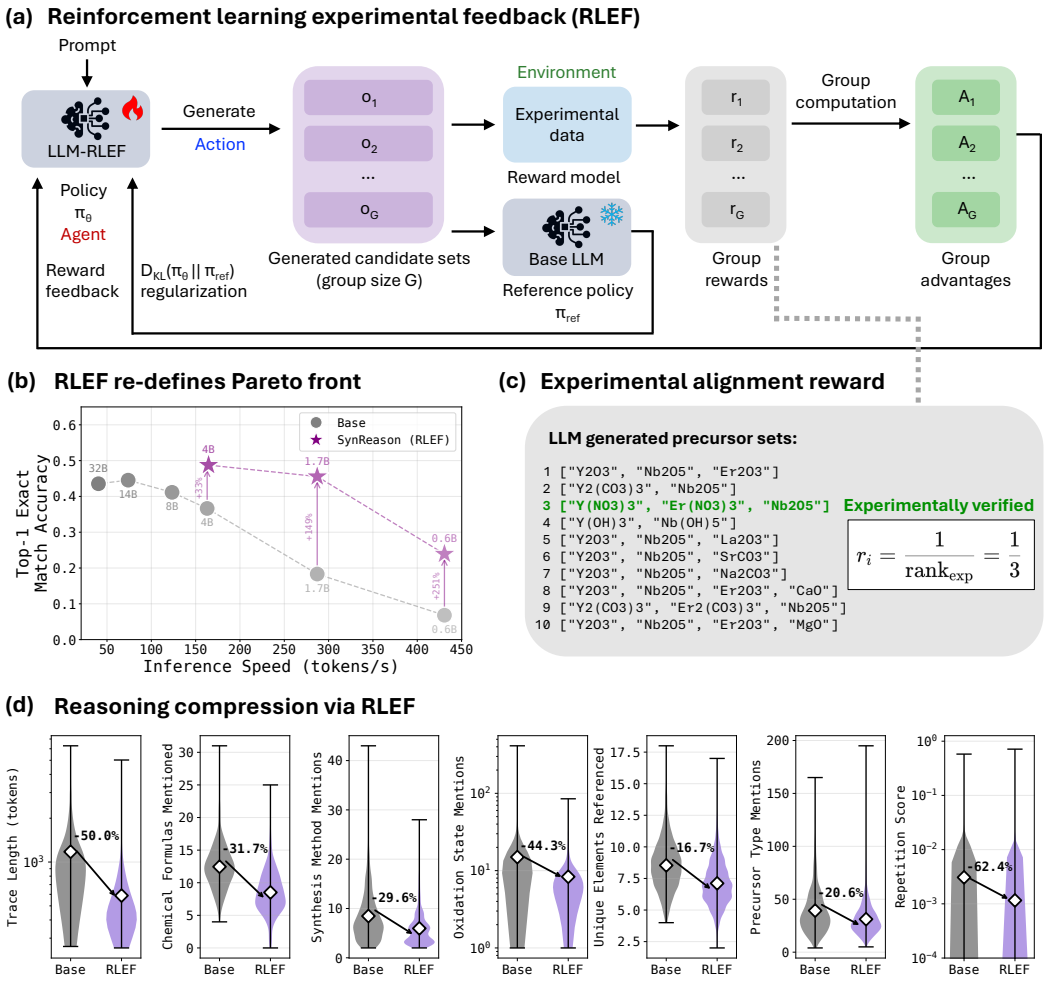

Figure 2: **Enhancing chemical reasoning with reinforcement learning experimental feedback (RLEF)** (a) RLEF post-training of LLMs. A LLM serves as an "agent" (π_θ) that takes "actions" to generate rollouts o_i of size G per prompt. Interaction with the environment (a reward model that measures agreement with literature recipes) yields reward r_i. Group normalized advantages (A_i) and KL divergence between finetuned and base models ($D_{\text{KL}}(\pi_\theta||\pi_{\text{ref}})$). (b) Impact of RLEF on accuracy-speed Pareto front of materials synthesis planning. (c) Experimental alignment reward for RLEF. For example, if a precursor set is ranked 3rd in the LLM output, the reward is the reciprocal rank ($\frac{1}{3}$). (d) RLEF compresses LLM reasoning. Surprisingly, after RLEF, LLMs output *shorter* reasoning traces and *fewer* chemical nomenclature mentions.

where $r_i = \frac{\pi_\theta(o_i|x)}{\pi_{\theta_{\text{old}}}(o_i|x)}$ is the importance ratio, $\widehat{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) = \frac{1}{G}\sum_{i=1}^{G}\left(\frac{\pi_{\text{ref}}(o_i|x)}{\pi_\theta(o_i|x)} - \log\frac{\pi_{\text{ref}}(o_i|x)}{\pi_\theta(o_i|x)} - 1\right)$, and ϵ is the clipping threshold. Intuitively, RLEF increases the probability of completions/recipes that result in *experimental* success while staying close to the base model.

3 Results and Discussion

3.1 Task and Experimental Setup

Precursor recommendation task. In forward inorganic synthesis, precursors $P_1, \ldots, P_m$ are combined and heated to form a target material T. Retrosynthesis inverts this process: given a target compound (e.g., $Li_7La_3Zr_2O_{12}$), it seeks a feasible precursor set (e.g., $\{LiOH, La_2O_3, ZrO_2\}$) that

can produce T. The problem is under-determined because many precursor sets may yield the same target under suitable conditions. Building upon previous work (He et al., 2023; Noh et al., 2024), the goal is to predict a ranked list of precursor sets, $(\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_K)$. Each set $\mathbf{S} = \{P_1, P_2, \ldots, P_m\}$ contains m precursor materials (the set size may vary), where each P_i is a single precursor (Fig. 1a). The ranking reflects the predicted likelihood that a set forms the target; literature-reported synthesis routes are treated as correct predictions.

Dataset. We use an experimental synthesis recipe dataset (Kononova et al., 2019), which contains 33,343 literature-extracted recipes (targets, precursors, and byproducts). To reduce noise from incomplete/ambiguous entries, we drop formulas with symbolic variables, or other placeholders, and enforce element-consistency: all target elements (except C/O/H/N) must appear in the precursor set. After filtering, we retain 9,255 unique recipes. Following prior work (Prein et al., 2025b;a), we evaluate on the novel materials system split, where no material system (defined by the set of elements in the target material) overlaps between the training and test sets.

Evaluation. We report Top-K exact match accuracy (Noh et al., 2024; He et al., 2023; Kim et al., 2022). For each target $\mathbf{x}_T$ with ground-truth precursor set $\mathbf{S}_{\text{true}}$, we predict ten candidate precursor sets in a structured list-of-lists format (Fig. 1a). This exact match is a strict metric as the predicted precursor sets have to match exactly (with the ground-truth) the correct number of precursors and their chemical identities.

3.2 Reasoning enhances synthesis planning

Fig. 1a shows the precursor recommendation task: given a target composition (e.g., $\mathrm{Y_{0.995}Er_{0.005}NbO_4}$) and a chemistry-oriented prompt (B.4), the model outputs 10 candidate precursor sets in a structured list-of-lists format. The example spans common precursor families-oxides (e.g., $\mathrm{Y_2O_3}, \mathrm{Nb_2O_5}, \mathrm{Er_2O_3}$), carbonates (e.g., $\mathrm{Y_2(CO_3)_3}$), nitrates (e.g., $\mathrm{Y(NO_3)_3}$), and hydroxides (e.g., $\mathrm{Y(OH)_3}$)—consistent with how solid-state and solution routes are parameterized in the literature (Fig. 1a). Beyond a single "canonical" recipe, this diversity enables chemically plausible alternatives that trade off availability, volatility, decomposition pathways, and mixing/solubility.

Fig. 1b reports Top-1 exact match accuracy vs. model size for Thinking vs Non-thinking modes. Non-thinking peaks around 0.3 (Fig. 1b). Thinking improves performance for mid-to-large models (4B to 32B), peaking at 0.45. However, these gains come with a computational cost: Thinking responses become longer as model size increases, while Non-thinking remains comparatively short (Fig. A3), motivating post-training methods that retain deliberate reasoning while reducing verbosity and latency.

3.3 RLEF redefines Pareto front

Compared to the base models, RLEF post-trained models achieve higher accuracy at comparable speed (Fig. 2b). We observe largest gains for the smallest models e.g., RLEF resulted in a +251% increase in Top-1 exact match accuracy for the 0.6B model. A 4B model post-trained with RLEF achieves a Top-1 exact match accuracy of 0.49, which *outperforms all base models up to 8× larger in size*. This result is also consistent across Top-K (Fig. A4). Importantly, RLEF redefines the Pareto front between accuracy and inference speed (Fig. 2b). This is important for synthesis planning systems that must evaluate multiple candidate materials. Evidently, RLEF provides a means to access more performant and deployable synthesis planning models.

3.4 Reasoning compression

For base models, the raw response-length distributions (Fig. A3) make clear that the Thinking mode can become increasingly expensive at larger scales. However, chemistry-metric analysis (Fig. A5) suggests that improvements do not simply come from mentioning more chemistry: surface-level indicators such as oxidation-state mentions, synthesis-method mentions, and formula mentions do not monotonically track accuracy.

Less (reasoning) is more. RLEF offers a mechanism to achieve higher accuracy while avoiding longer reasoning traces. Surprisingly, contrary to previous works (Guo et al., 2025) that report longer reasoning after RL post-training, we found that RLEF systematically reduces reasoning trace,

resulting in *reasoning compression* Fig 2d. A common concern with reasoning-style prompting is verbosity: long traces can increase latency and introduce contradictory/repetitive statements that confound the LLM (Wei et al., 2022; Turpin et al., 2023; Arcuschin et al., 2025; Guo et al., 2025). Notably, a 62 $\%$ reduction (Fig 2d) in repetition score is indicative of a shift toward more selective and task-relevant reasoning, rather than more verbose reasoning, contributing to improvement gains post RLEF.

Qualitative analysis. We show 2 qualitative examples to illustrate the behavioral changes induced by RLEF in concrete synthesis planning scenarios. In one case, RLEF recovers a missing literature precursor set that the base model fails to propose (Fig. A6); in another, both models include the correct set but RL ranks it higher (Fig. A7). These examples align with the experimental alignment reward signal in Fig. 2c and help explain why RLEF improves Top-K exact match without requiring longer or more elaborate traces.

4 Conclusion

We present RLEF, a RL post-training approach for aligning LLM generation with experimental outcomes. Empirically, RLEF improves Top-1 exact match accuracy, expands the accuracy–speed Pareto front on materials synthesis planning task, and results in reasoning compression, where reasoning is compacted and more task-selective while simultaneously improving performance.

Limitations. Our evaluation (Top-K exact match) focuses on exact match, which may under-credit near-miss candidates that are chemically plausible but differ from the canonical recipe. From the task-specific perspective, we do not explicitly model synthesis conditions (temperature, atmosphere, processing steps) or reaction kinetics/thermodynamics, which are critical for real-world experimental success. More broadly, precursor recommendation can also be viewed as a constraint-satisfaction problem, where feasible predictions must include target elements and satisfy chemical feasibility. Explicitly incorporating such constraints into the RL objective may further improve performance. Future work could integrate richer reward signals (e.g., thermodynamics/kinetics), and extend beyond precursor recommendation to joint planning of conditions and precursors.

References

Iván Arcuschin, Jett Janiak, Robert Krzyzanowski, Senthooran Rajamanoharan, Neel Nanda, and Arthur Conmy. Chain-of-thought reasoning in the wild is not always faithful, 2025.

Luis Barroso-Luque, Muhammed Shuaibi, Xiang Fu, Brandon M Wood, Misko Dzamba, Meng Gao, Ammar Rizvi, C Lawrence Zitnick, and Zachary W Ulissi. Open materials 2024 (omat24) inorganic materials dataset and models. *arXiv preprint arXiv:2410.12771*, 2024.

Matteo Bianchini, Jingyang Wang, Raphaële J Clément, Bin Ouyang, Penghao Xiao, Daniil Kitchaev, Tan Shi, Yaqian Zhang, Yan Wang, Haegyeom Kim, et al. The interplay between thermodynamics and kinetics in the solid-state synthesis of layered oxides. *Nature materials*, 19(10):1088–1095, 2020.

Daya Guo, Dejian Yang, Haowei Zhang, et al. Deepseek-r1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Tanjin He, Haoyan Huo, Christopher J Bartel, Zheren Wang, Kevin Cruse, and Gerbrand Ceder. Precursor recommendation for inorganic synthesis by machine learning materials similarity from scientific literature. *Science advances*, 9(23):eadg8180, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2106.09685.

Christopher Karpovich, Elton Pan, Zach Jensen, and Elsa Olivetti. Interpretable machine learning enabled inorganic reaction classification and synthesis condition prediction. *Chemistry of Materials*, 35(3):1062–1079, 2023.

Kun Joong Kim, Moran Balaish, Masaki Wadaguchi, Lingping Kong, and Jennifer LM Rupp. Solid-state li–metal batteries: challenges and horizons of oxide and sulfide solid electrolytes and their interfaces. *Advanced Energy Materials*, 11(1):2002689, 2021.

Seongmin Kim, Juhwan Noh, Geun Ho Gu, Shuan Chen, and Yousung Jung. Element-wise formulation of inorganic retrosynthesis. In *AI for Accelerated Materials Design NeurIPS 2022 Workshop*, 2022.

Seongmin Kim, Yousung Jung, and Joshua Schrier. Large language models for inorganic synthesis predictions. *Journal of the American Chemical Society*, 146(29):19654–19659, 2024.

Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. Text-mined dataset of inorganic materials synthesis recipes. *Scientific data*, 6 (1):203, 2019.

Ji Lin, Jiaming Tang, Haoyang Yu, Xiuyu Yang, Yong Li, Wei Zhang, et al. Awq: Activation-aware weight quantization for llm compression and acceleration. *arXiv preprint arXiv:2306.00978*, 2023.

Rubayyat Mahbub, Kevin Huang, Zach Jensen, Zachary D Hood, Jennifer LM Rupp, and Elsa A Olivetti. Text mining for processing conditions of solid-state battery electrolytes. *Electrochemistry Communications*, 121:106860, 2020.

Shreshth A Malik, Rhys EA Goodall, and Alpha A Lee. Predicting the outcomes of material syntheses with deep learning. *Chemistry of Materials*, 33(2):616–624, 2021.

Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

Paulius Micikevicius, Dusan Stosic, Neil Burgess, Marius Cornea, Pradeep Dubey, Richard Grisenthwaite, Sangwon Ha, Alexander Heinecke, Patrick Judd, John Kamalu, Naveen Mellempudi, Stuart Oberman, Mohammad Shoeybi, Michael Siu, and Hao Wu. Fp8 formats for deep learning, 2022.

Heewoong Noh, Namkyeong Lee, Gyoung S Na, and Chanyoung Park. Retrieval-retro: Retrieval-based inorganic retrosynthesis with expert knowledge. *arXiv preprint arXiv:2410.21341*, 2024.

Elton Pan, Soonhyoung Kwon, Sulin Liu, Mingrou Xie, Yifei Duan, Thorben Prein, Killian Sheriff, Yuriy Roman, Manuel Moliner, Rafael Gómez-Bombarelli, et al. A chemically-guided generative diffusion model for materials synthesis planning. In *AI for Accelerated Materials Design-NeurIPS 2024*, 2024.

Elton Pan, Soonhyoung Kwon, Sulin Liu, Mingrou Xie, Alexander J Hoffman, Yifei Duan, Thorben Prein, Killian Sheriff, Yuriy Roman-Leshkov, Manuel Moliner, et al. Diffsyn: A generative diffusion approach to materials synthesis planning. *arXiv preprint arXiv:2509.17094*, 2025.

Thorben Prein, Elton Pan, Sami Haddouti, Marco Lorenz, Janik Jehkul, Tymoteusz Wilk, Cansu Moran, Menelaos Panagiotis Fotiadis, Artur P Toshev, Elsa Olivetti, et al. Retro-rank-in: A ranking-based approach for inorganic materials synthesis planning. *arXiv preprint arXiv:2502.04289*, 2025a.

Thorben Prein, Elton Pan, Janik Jehkul, Steffen Weinmann, Elsa Olivetti, and Jennifer LM Rupp. Language models enable data-augmented synthesis planning for inorganic materials. *ACS Applied Materials & Interfaces*, 17(51):69221–69233, 2025b.

James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher Wolverton. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *Jom*, 65:1501–1509, 2013.

Jonathan Schmidt, Tiago FT Cerqueira, Aldo H Romero, Antoine Loew, Fabian Jäger, Hai-Chen Wang, Silvana Botti, and Miguel AL Marques. Improving machine-learning models in materials science through large datasets. *Materials Today Physics*, 48:101560, 2024.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Anuroop Sriram, Sihoon Choi, Xiaohan Yu, Logan M Brabson, Abhishek Das, Zachary Ulissi, Matt Uyttendaele, Andrew J Medford, and David S Sholl. The open dac 2023 dataset and challenges for sorbent discovery in direct air capture, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting, 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2022.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Sasha Shysheya, Jonathan Crabbé, Lixin Sun, Jake Smith, et al. Mattergen: a generative model for inorganic materials design. *arXiv preprint arXiv:2312.03687*, 2023.

Yuntong Zhu, Ellis R Kennedy, Bengisu Yasar, Haemin Paik, Yaqian Zhang, Zachary D Hood, Mary Scott, and Jennifer LM Rupp. Uncovering the network modifier for highly disordered amorphous li-garnet glass-ceramics. *Advanced Materials*, 36(16):2302438, 2024.

A ADDITIONAL DISCUSSION ON QUANTIZATION

Quantization confers time and space savings One potential way is quantization of the LLMs. Fig. A1 and Fig. A2 indicate that further throughput and memory gains are achievable with limited degradation in Top-1 exact match accuracy. We consider different quantization techniques, including 8-bit floating point (FP8) and AWQ (4-bit) (Micikevicius et al., 2022; Lin et al., 2023). Across model sizes, quantization yields substantial efficiency improvements. From Fig. A1, FP8 and AWQ typically increase throughput from roughly ∼50–450 tokens/s (non-quantized) to ∼200–900 tokens/s (quantized), corresponding to about ∼2–4× speedup depending on model size. Simultaenously, Fig. A2 shows that quantization reduces GPU memory usage by about ∼2× (FP8) to ∼4× (AWQ) at comparable accuracy. These gains enable practical deployment for high-throughput screening under tight latency and memory budgets.

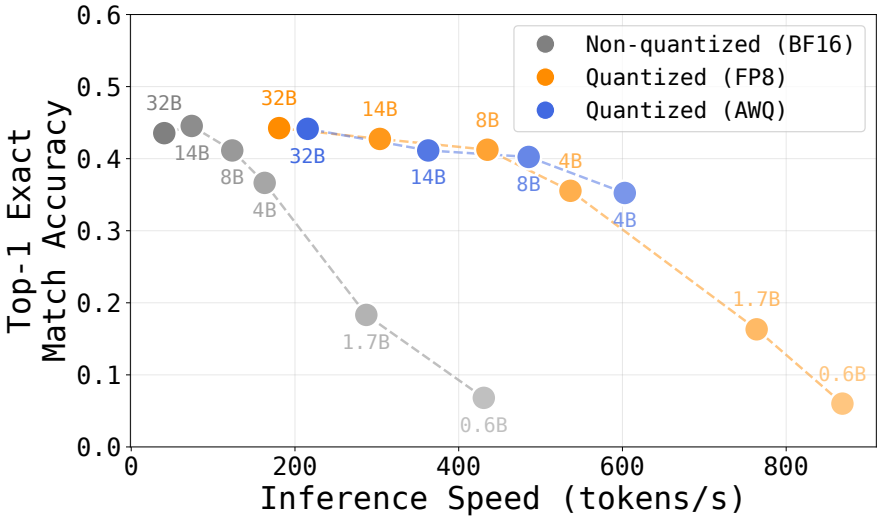

Figure A1: **Quantization tradeoff: speed vs. accuracy.** Top-1 exact-match accuracy versus inference speed under different quantization settings.

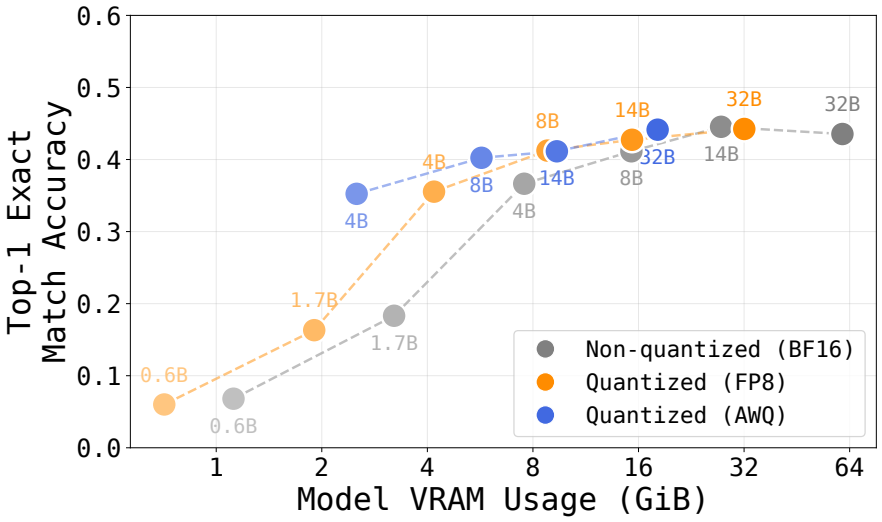

Figure A2: **Quantization tradeoff: VRAM vs. accuracy.** Top-1 exact-match accuracy versus GPU memory usage under different quantization settings.

B EXPERIMENTAL DETAILS

B.1 EXPERIMENTAL SETUP

We use Qwen3 family models with parameter scales from 0.6B to 32B. Unless otherwise specified, inference is performed with vLLM and RL post-training is performed with verl. We evaluate three numerical formats: BF16 (non-quantized), FP8, and AWQ. All inference and RLEF runs are performed on a single NVIDIA DGX Spark GB10 GPU.

B.2 LLM INFERENCE

Table 1 summarizes the inference hyperparameters (vLLM backend).

Table 1: Inference hyperparameters (vLLM).

Parameter	Value
Temperature	0.2
Batching	True
Batch size	128
Max batch tokens	262144
Thinking mode	True
GPU memory utilization	0.85

B.3 RLEF TRAINING

We use Low-Rank Adaptation (LoRA) for parameter-efficient finetuning, which freezes the base model weights and learns a small set of low-rank update matrices in attention/MLP layers (Hu et al., 2022). Table 2 summarizes the GRPO training hyperparameters (verl).

Table 2: GRPO training hyperparameters (verl).

Parameter	Value
Epochs	10
Train batch size	128
Mini-batch size	32
Micro-batch size	4
Learning rate	5e-5
KL coefficient	0.001
Max prompt length	1024
Max response length	2048
Temperature	0.7
Samples per prompt (G)	2
Tensor parallel size	1
GPU memory utilization	0.5
Log-prob micro-batch size	32
LoRA rank	32
LoRA alpha	32
GPUs	1
Nodes	1
Thinking mode	True

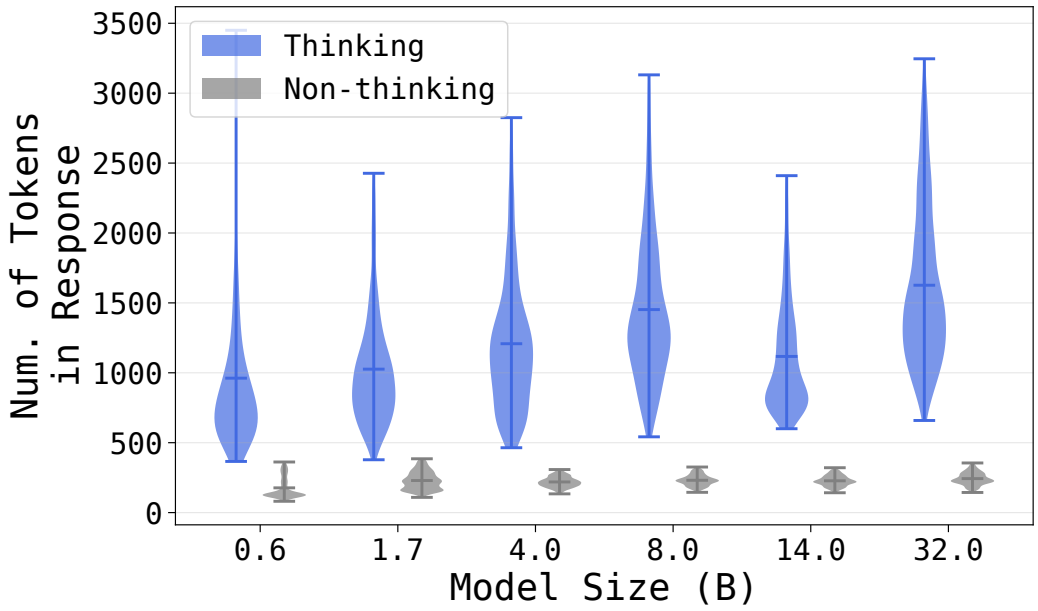

Figure A3: **Response length vs. model size.** Token-count distributions for Thinking vs. Non-thinking across model sizes.

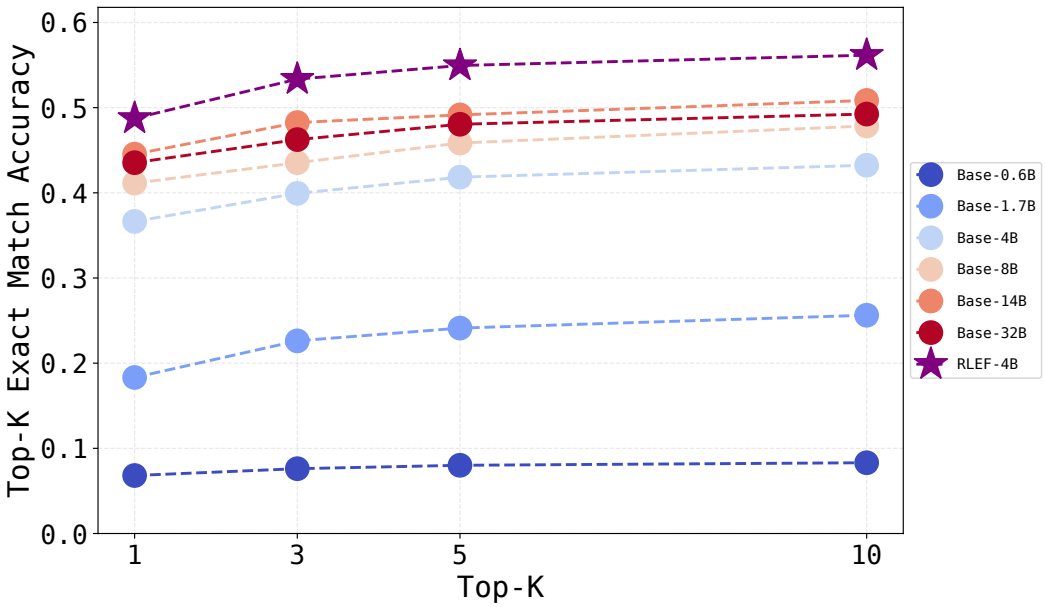

Figure A4: **Top-K exact match accuracy vs. K.** A 4B model trained with RLEF outperforms even base models 8× larger (e.g., 32B) across all Top-K.

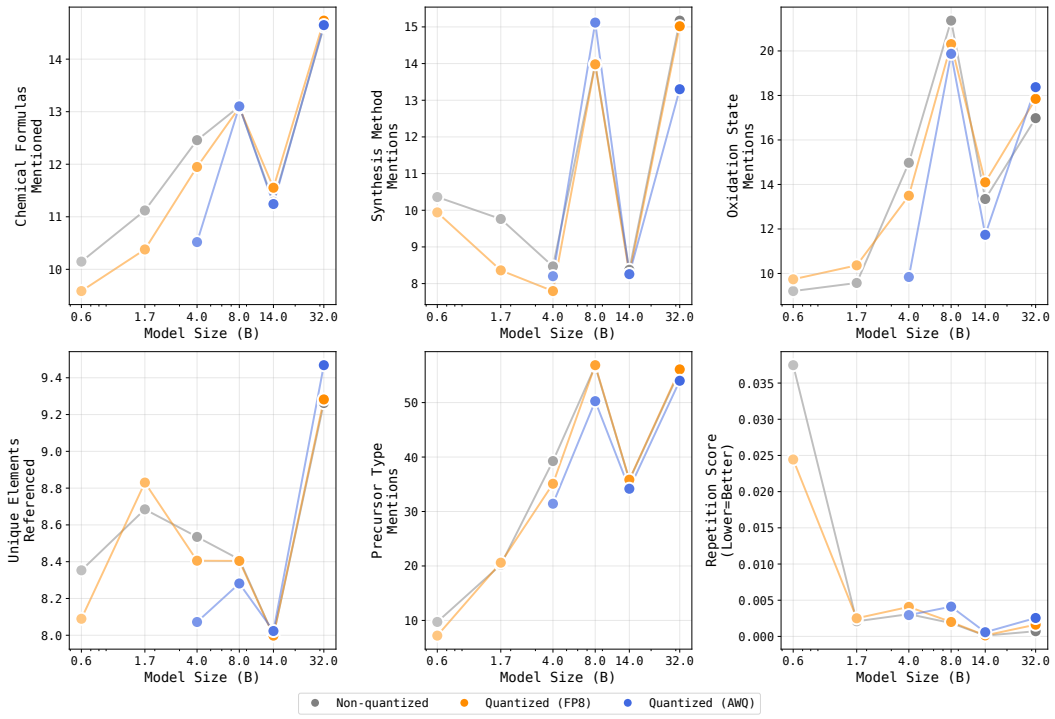

Figure A5: **Chemistry metrics vs. model size.** Mentions of formulas, synthesis methods, oxidation states, unique elements, precursor types, and a repetition score across model sizes and quantization settings.

Figure A6: **Qualitative example where RLEF fixes a failure case.** RL produces a ranked list that contains the literature precursor set, whereas the base model misses it.

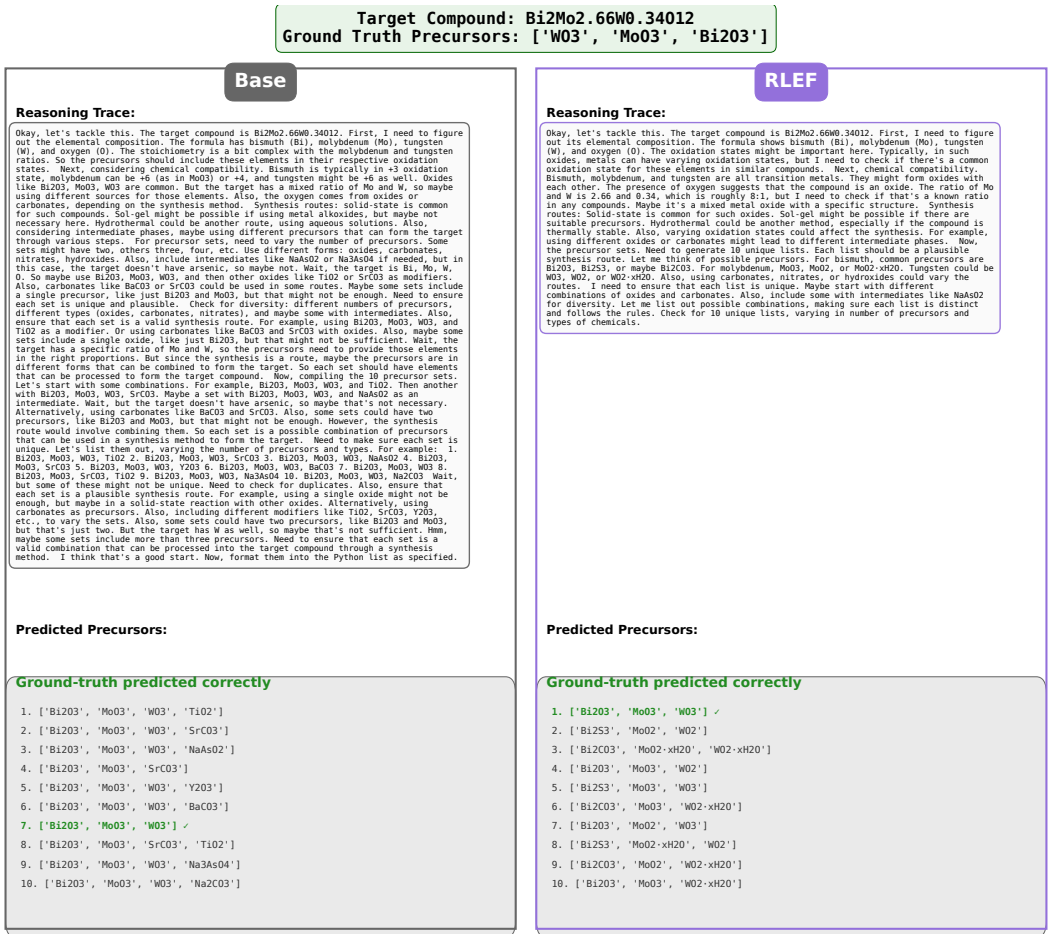

Figure A7: **Qualitative example where RLEF improves ranking.** Both models include the literature precursor set, but RL ranks it higher.

B.4 CHATML-STYLE PROMPT

We use the following prompt (in ChatML format) for both inference and RLEF.

```
[
  {
    "role": "system",
    "content": "You are a helpful chemistry assistant.
    Task: Given a target compound, output (A) a short reasoning and (B
    ) exactly 10 diverse precursor sets.
    Reasoning should consider: elemental composition, chemical
    compatibility, synthesis routes (solid-state, sol-gel, hydrothermal
    ), and oxidation states.
    You must follow formatting rules exactly."
  },
  {
    "role": "user",
    "content": "Target compound: <TARGET_FORMULA>

    Output has TWO parts in this exact order:
    1) Reasoning (STRICT): Under 10 sentences. Consider:
       - Elemental composition match with the target
       - Chemical compatibility and common synthetic pathways
       - Different synthesis routes (e.g., solid-state, sol-gel,
    hydrothermal)
       - Varying levels of oxidation states or intermediate phases
       No bullet points, no numbering, no repeated lists.
    2) Precursor sets: a Python list of lists inside a ```python code
    block.

    Format example (this is just the FORMAT, not actual content. Do
    not use precursors in this example unless necessary):

    ```python
    [
        ["BaCO3", "TiO2"],
        ["BaO", "TiO2"],
        ["BaCO3", "TiO2", "La2O3"],
        ["BaCO3"],
        ["BaO2", "TiO2"],
        ["BaCO3", "SrCO3", "TiO2"],
        ["BaO"],
        ["BaCO3", "Ti2O3"],
        ["BaCO3", "TiO2", "Y2O3"],
        ["BaO", "Ti2O3"]
    ]
    ```

    Precursor-set rules:
    - Exactly 10 inner lists (no more, no fewer).
    - Each inner list contains ONLY chemical formulas (no names).
    - Each inner list is a plausible synthesis route and must be
    UNIQUE.
    - Diversity: vary #precursors (25), use alternative chemistries (
    oxides/carbonates/nitrates/hydroxides/arsenates), and optionally
    include intermediates (e.g., NaAsO2/Na3AsO4) to diversify routes.
    - Do NOT include any extra text after the code block.

    If you cannot comply with the Reasoning rules, skip reasoning and
    output ONLY the code block.
    /think'
}]
```