# MULTI-TEACHER ON-POLICY DISTILLATION FOR MULTI-PROPERTY MOLECULAR OPTIMIZATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Lead optimization requires a generative policy to improve every property named in an instruction while keeping the molecule valid and structurally close to the lead. Property specialists trained with reinforcement learning are cheap to obtain, but composing them is not: a molecular instruction activates several specialists at once, they differ in how dependable they are at a given molecular state, and specialists that propose comparably reasonable molecules may still prefer different continuations of the same partial structure. We propose reliability-aware multi-teacher on-policy distillation (RA-MOPD), which folds a bank of frozen property experts into a single policy that queries neither experts nor property oracles at inference. At every state the student visits, RA-MOPD weights the requested experts by held-out reliability together with verified outcomes of their own proposals at that state. The weighted experts are combined into a full-vocabulary product of experts on the student's realized prefixes, and we show that the log-normalizer of this product is exactly the minimum weighted reverse-KL disagreement among the active experts, which we use as a token-local gate on the distillation loss. A group-relative verifier reward supplies the complementary trajectory-level signal, scoring complete recurrent edits, while the weighted teachers provide token-level distributional supervision on the molecular answer. Across ten three- and four-property tasks, RA-MOPD improves constrained success over supervised fine-tuning, verifier-only reinforcement learning, offline expert imitation, parameter merging and uniformly weighted multi-teacher distillation, and its advantage widens as every property is required to improve by a larger margin.

## 1 INTRODUCTION

Structure-based design has made it routine to propose molecules that fit a binding pocket, but geometric complementarity is only one of several requirements. A candidate that binds tightly is of little use if it cannot cross the relevant membrane, is cleared too quickly, or cannot be synthesized (Bickerton et al., 2012; Ertl & Schuffenhauer, 2009). Lead optimization is therefore a constrained multi-objective problem: starting from a lead molecule, an editor must move *every* requested property in the desired direction while keeping the result chemically valid and close enough to the original scaffold that accumulated medicinal-chemistry knowledge still transfers (Jin et al., 2018; 2020). Generative and reinforcement-learning formulations have long pursued this goal over graphs and strings (Olivecrona et al., 2017; You et al., 2018; Zhou et al., 2019; Gao et al., 2022), typically by optimizing a scalarized score.

Molecular foundation models capture structure–property relationships well and serve as scoring functions for virtual screening (Wu et al., 2018). A score, however, ranks molecules that already exist; it does not indicate how to repair one that fails. Instruction-tuned large language models (LLMs) supply the missing interface (Edwards et al., 2022; Fang et al., 2024). A medicinal-chemistry goal can be stated in language, translated into an explicit structural edit, and revised over several rounds as the molecular state changes. GeLLM$^3$O, RePO and MolMem instantiate this shift from passive property assessment to instruction-conditioned optimization (Dey et al., 2025; Li et al., 2026; Wang et al., 2026).

Post-training is the remaining bottleneck, and the paradigms available to it trade density of supervision against fidelity to the deployed state distribution. *Supervised fine-tuning* (SFT) maximizes the



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

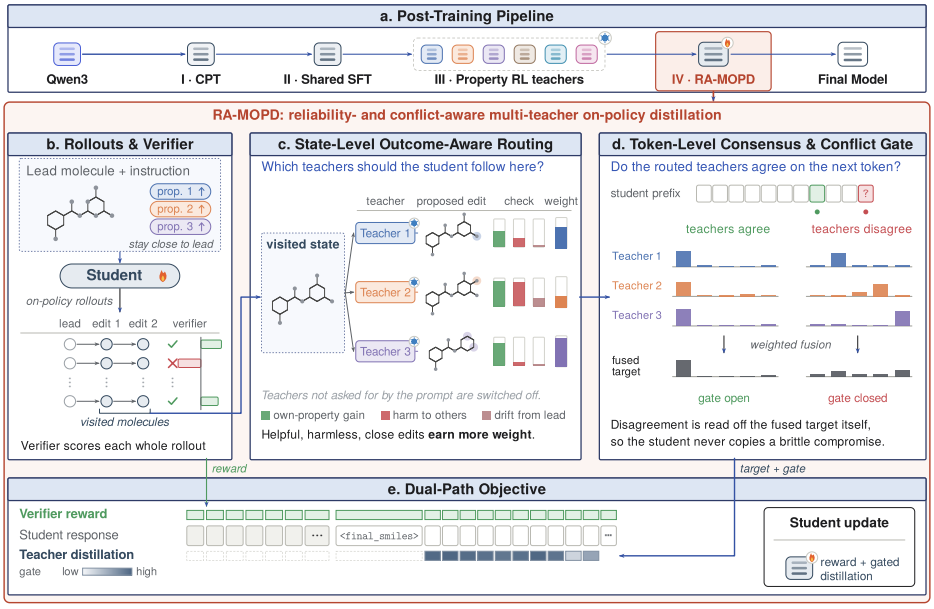



Figure 1: Overview of RA-MOPD. Single-property experts are independently trained via RL from a shared SFT checkpoint. At each student state, frozen property functions verify expert proposals into routing weights; the routed experts form a product-of-experts distillation target whose normalizer gates the loss. At inference only the student remains.

likelihood of reference responses and gives a dense per-token target, but the references are collected under a fixed behavior policy, so the model is never corrected on the states its own errors produce (Ouyang et al., 2022; Ross et al., 2011). *Reinforcement learning* (RL) with a programmatic verifier removes that mismatch by training on the model's own rollouts and scoring them with a checkable outcome (Schulman et al., 2017; Shao et al., 2024; Lambert et al., 2024), at the cost of compressing an entire rationale and molecular answer into one delayed scalar that certifies success without attributing it to particular token decisions. *On-policy distillation* (OPD) combines the two by running a teacher over the student's own rollouts, so a dense token-level target is supplied at exactly the states the student visits (Lin et al., 2020; Agarwal et al., 2024; Gu et al., 2024). *Multi-teacher on-policy distillation* (MOPD) extends OPD to a bank of separately trained specialists, letting capabilities that were developed independently be integrated into one student (Ma et al., 2026; You et al., 2017; Liu et al., 2020), and has proved effective across capability-fusion settings (Wu et al., 2026; Yang et al., 2026; Pan et al., 2026; Qin et al., 2026).

Multi-property optimization departs from the setting in which MOPD was developed. There, a prompt belongs to one domain and is dispatched to the teacher that owns it, so composition is essentially a routing decision among mutually exclusive candidates. A molecular instruction instead activates several teachers simultaneously, each optimized for a single property and none of them responsible for the conjunction the instruction actually demands. Two difficulties follow. Teachers differ in how reliable they are, and their reliability is not uniform across molecular states: one that is dependable in aggregate may still propose an edit that does little for the particular conjunction at hand. Teachers also differ at the level of decoding, because independently trained single-property policies can prefer different continuations of the same partial molecule even when the molecules they eventually propose are comparably reasonable. Weighting all requested teachers uniformly ignores both, and yields an averaged target that no individual teacher endorses. Uniform composition is known to allocate its optimization budget unevenly even in less demanding settings (Gao et al., 2026).

We introduce reliability-aware multi-teacher on-policy distillation (RA-MOPD), which distills a bank of frozen, independently reinforcement-learned property experts into one student that serves arbitrary multi-property prompts without querying any expert or property oracle at test time (Figure 1). RA-MOPD acts at the two resolutions at which expert heterogeneity appears. At the *state* level, every requested expert proposes a candidate molecule at the student's current state; the frozen property functions verify what that proposal achieves, and these verified outcomes combine with a held-out reliability prior into state-dependent routing weights. At the *token* level, the routed experts are

composed by a full-vocabulary product of experts (PoE) on the student's realized prefixes (Hinton, 2002). We show that the log-normalization cost of this product equals the minimum weighted reverse-KL divergence from any single distribution to the active experts, so it measures irreducible disagreement and can be used directly to attenuate the distillation loss. A group-relative verifier reward supplies the complementary trajectory-level signal.

Our contributions are as follows.

- **RA-MOPD**, a multi-teacher on-policy distillation framework that composes single-property experts at two resolutions—reliability-aware state-level routing and conflict-sensitive token-level gating—with both signals derived from verified molecular outcomes rather than tuned heuristics.
- **Information-geometric analysis** showing that the product-of-experts target is the reverse-KL barycenter of the active teachers, its normalizer quantifies their irreducible disagreement, and the routing softmax is an entropy-regularized allocation.
- **Systematic evaluation** under a strict-conjunction, call-budgeted protocol across seen and held-out property combinations, where RA-MOPD consistently outperforms SFT, verifier-only RL, offline imitation, parameter merging and uniform multi-teacher distillation, with ablations isolating each component's contribution.

## 2 Task and Evaluation Protocol

### 2.1 Instruction-conditioned constrained optimization

Let $m_0 \in \mathcal{M}$ be a lead molecule and $x$ a natural-language instruction requesting a subset $P_i \subseteq \mathcal{P}$ of properties, where

$$\mathcal{P} = \{\text{BBBP}, \text{DRD2}, \text{HIA}, \text{Mutagenicity}, \text{pLogP}, \text{QED}\}.$$

The policy produces a response $y = [r; \hat{m}]$ consisting of a short natural-language rationale $r$ and one molecule $\hat{m}$ enclosed in a `<final_smiles>` span. For property $j$ we write $f_j : \mathcal{M} \to \mathbb{R}$ for the frozen property function, $d_j = -1$ for mutagenicity and $d_j = +1$ otherwise, and $\sigma_j > 0$ for a fixed normalization scale. The signed normalized gain is

$$g_j(m; m_0) = d_j \frac{f_j(m) - f_j(m_0)}{\sigma_j}, \tag{1}$$

so that larger values are uniformly better. Weighted-sum objectives permit a large gain on one property to mask a failure on another. We instead require conjunction:

$$S_i(m) = \mathbb{I}\left[\text{valid}(m) \wedge \text{Sim}(m, m_0) \geq \eta \wedge \min_{j \in P_i} g_j(m; m_0) > 0\right], \tag{2}$$

with $\eta = 0.5$ and Sim the Morgan-fingerprint Tanimoto similarity (Rogers & Hahn, 2010). Similarity is a feasibility constraint rather than a quantity to maximize.

### 2.2 Recurrent and budgeted evaluation

A fixed policy induces a trajectory $m_0 \to m_1 \to \cdots \to m_H$ in which a valid generated molecule becomes the next state while all gains and similarity constraints stay anchored to the original lead $m_0$. For $K$ independently sampled trajectories, constrained success is

$$\text{cSR@}K = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[\exists k \leq K : S_i(m_{i,H}^{(k)}) = 1\right]. \tag{3}$$

The primary protocol fixes $H = 2$, $K = 1$ and greedy decoding. *Final-cSR* evaluates $m_H$, whereas *BestPrefix-cSR* records whether any $m_h$ with $h \leq H$ succeeded and thus exposes later edits that destroy an earlier success. No expert, property oracle, candidate selector or external search routine is available at inference.

The primary evaluation set provides an equal number of prompts for each of ten property combinations, labelled T1–T10, forming a balanced compact benchmark of 500 prompts that enables fast, controlled per-combination comparison while conserving compute. Appendix E confirms that the relative ordering is preserved on the full evaluation set of $N = 3{,}905$ prompts. Five combinations are observed during multi-objective training and five are held out. Appendix B lists the combinations.

# 3 METHOD

## 3.1 SETTING AND UNIFORM COMPOSITION

Let $\pi_\theta$ denote the student and $\{\pi_{\phi_j}\}_{j\in\mathcal{P}}$ a bank of frozen property teachers, of which only those indexed by the requested set $P_i \subseteq \mathcal{P}$ are active for prompt $i$. Starting from $m_0$, the student samples a recurrent trajectory $\xi = (y_0, m_1, \ldots, y_{H-1}, m_H)$ on-policy, where each $y_t$ contains a rationale and one tagged SMILES and a valid extracted molecule becomes the next state. Constraints are always evaluated against $m_0$. The full training pipeline that produces the student initializer and the teacher bank is summarized in Appendix Table 7, and the complete optimization loop in Appendix Algorithm 1.

The direct extension of OPD to several teachers takes a weighted geometric mean of the active teachers' token distributions and renormalizes it into a single consensus teacher distribution. With uniform weights this is

$$p_{\text{uni}}(v \mid h) \propto \prod_{j\in P_i} p_j(v \mid h)^{1/|P_i|}. \tag{4}$$

Two assumptions are hidden here. Equation 4 treats every requested teacher as equally informative at every molecular state, and it forces a normalized target even when the teachers concentrate mass on different continuations. Composition therefore calls for decisions at two resolutions: how much each teacher should count for the present molecule (§3.2), and whether the selected teachers admit a reliable consensus at all (§3.3). RA-MOPD derives both from verified quantities and pairs them with a trajectory-level verifier reward (§3.4).

## 3.2 RELIABILITY- AND OUTCOME-AWARE ROUTING AT MOLECULAR STATES

Routing should reflect both expert-global competence and state-local outcome, since a teacher that reliably improves its own property on held-out data need not be the most useful contributor for the conjunction requested at a particular state. We estimate competence from held-out single-property evaluation: with $s_j$ successes out of $n_j$ examples, the reliability prior is the Beta$(1, 1)$ posterior mean $\rho_j = (s_j + 1)/(n_j + 2)$.

State-local evidence comes from the teachers' own proposals. At molecular state $m_t$, every requested teacher $j \in P_i$ decodes one candidate molecule $c_{j,t}$ deterministically, which the frozen property functions evaluate against the original lead $m_0$ to return three scalars,

$$\begin{aligned} G_{j,t} &= \text{clip}(g_j(c_{j,t}; m_0), -2, 2)\,, \\ H_{j,t} &= \text{clip}\left(\frac{1}{|P_i| - 1} \sum_{k\in P_i\setminus\{j\}} [-g_k(c_{j,t}; m_0)]_+, 0, 2\right), \\ V_{j,t} &= \left[\frac{\eta - \text{Sim}(c_{j,t}, m_0)}{\eta}\right]_+, \end{aligned} \tag{5}$$

namely own-property progress $G_{j,t}$, the mean collateral shortfall $H_{j,t}$ on the co-requested properties, and similarity violation $V_{j,t}$ relative to the threshold $\eta$. All three are computed from the edit a teacher actually proposes at $m_t$, not from its average behavior, so a teacher can be demoted at a state where its proposal does not serve the requested conjunction even when its global reliability is high. Prior and outcomes form a per-teacher utility, normalized by a softmax over the active set:

$$u_{j,t} = \lambda_\rho \log(\rho_j + \epsilon) + G_{j,t} - H_{j,t} - \lambda_s V_{j,t}, \qquad \alpha_{j,t} = \frac{\exp(u_{j,t}/\tau_\alpha)}{\sum_{k\in P_i} \exp(u_{k,t}/\tau_\alpha)}. \tag{6}$$

We set $\lambda_\rho = 5$ so that the reliability prior spans a range comparable to the verified-outcome terms, and $\lambda_s = 4$ so that similarity violations are penalized heavily. Invalid proposals are assigned a fixed unfavorable outcome before Equation 6, and teachers outside $P_i$ receive zero weight. The router is recomputed after every accepted edit; no gradient is propagated through the property functions or the discrete proposals. Equation 6 is not an arbitrary scoring rule: Proposition 3 in Appendix D shows that it uniquely solves an entropy-regularized allocation over the active teachers and that at $\tau_\alpha = 1$ it factorizes as $\alpha_{j,t} \propto (\rho_j + \epsilon)^{\lambda_\rho} \exp(G_{j,t} - H_{j,t} - \lambda_s V_{j,t})$, so reliability acts as a multiplicative prior and verified outcomes as the likelihood ratio that reweights it.

### 3.3 Token-level consensus and exact conflict

Routing establishes how much each teacher should count at a molecular state. It does not establish that the selected teachers can be followed simultaneously: two teachers may both propose acceptable molecules and still place their probability mass on different continuations of the same SMILES prefix, since single-property RL shaped their decoding preferences independently. Let $h_{t,n}$ be the student's realized prefix at token $n$ of round $t$ and $p_j(v \mid h_{t,n})$ teacher $j$'s full-vocabulary distribution. The weighted product of experts is

$$s_{t,n}(v) = \sum_{j \in P_i} \alpha_{j,t} \log p_j(v \mid h_{t,n}), \quad Z_{t,n} = \log \sum_v e^{s_{t,n}(v)}, \quad p_T(v \mid h_{t,n}) = e^{s_{t,n}(v) - Z_{t,n}}. \tag{7}$$

An arithmetic mixture $\sum_j \alpha_j p_j$ averages probability mass and can spread support over incompatible tokens. The product instead concentrates mass on jointly supported tokens: a token survives normalization only if it receives substantial probability from every high-weight teacher. The same construction supplies a measure of when no jointly supported continuation exists, because the normalizer $Z_{t,n}$ admits an exact information-geometric reading.

**Proposition 1** (Consensus target and exact conflict)**.** *Assume common finite support and $\alpha_{j,t} \geq 0$ with $\sum_j \alpha_{j,t} = 1$. Then*

$$p_T = \arg\min_q \sum_j \alpha_{j,t} D_{\mathrm{KL}}(q \| p_j), \qquad C_{t,n} \equiv \min_q \sum_j \alpha_{j,t} D_{\mathrm{KL}}(q \| p_j) = -Z_{t,n} \geq 0. \tag{8}$$

*For strictly positive distributions, $C_{t,n} = 0$ if and only if all active teachers are identical.*

The proof is in Appendix D, together with a bound certifying that $C_{t,n}$ never exceeds the cost of simply following the closest available teacher. The quantity $C_{t,n} = -Z_{t,n}$ is the minimum aggregate reverse-KL divergence from any single distribution to the active teachers: it vanishes under agreement and grows as their preferred continuations diverge. Conflict in this sense is distinct from entropy. Two confident teachers that favor different tokens produce a large $C_{t,n}$, whereas two teachers sharing one high-entropy distribution produce $C_{t,n} = 0$; the quantity measures incompatibility *between* experts rather than uncertainty *within* any one of them.

We convert it into a soft gate $\gamma_{t,n} = \exp(-C_{t,n}/\tau_C) \in (0, 1]$ that attenuates the distillation loss where the teachers disagree. Exact agreement gives $\gamma_{t,n} = 1$ and full supervision, while as disagreement grows $\gamma_{t,n}$ decays toward zero, so the student is not driven toward a geometric compromise that no individual teacher would endorse.

### 3.4 Dual-path on-policy objective

The objective combines two complementary supervision paths that address different resolutions of the learning problem. The policy-gradient loss scores complete recurrent trajectories with a group-relative verifier reward, providing trajectory-level feedback on whether the full conjunction was satisfied. The distillation loss provides token-level supervision from the teacher consensus, restricted to the SMILES span where property-relevant structure is encoded. On-policy rollouts ensure that the student encounters a sufficiently diverse and appropriately difficult distribution of molecular states, correcting for the mismatch that fixed expert trajectories introduce. The two paths are additive and address complementary gaps: the verifier evaluates whole-trajectory outcomes but supplies no token attribution, while the teachers carry privileged single-property competence at each decoding position but have no view of whether the complete edit satisfied the requested conjunction.

The student samples $B$ recurrent trajectories per prompt; given verifier rewards $\{R^{(b)}\}$, the group-relative advantage $\widehat{A}^{(b)}$ standardizes $R^{(b)}$ by the mean and standard deviation of the group (Shao et al., 2024). The policy loss applies it to the complete response, covering both the free-form rationale and the tagged SMILES:

$$\mathcal{L}_{\mathrm{RL}}(\theta) = -\mathbb{E}_{\xi^{(b)}\sim\pi_\theta}\left[\frac{\widehat{A}^{(b)}}{N^{(b)}}\sum_t\sum_{n=1}^{|y_t^{(b)}|}\log\pi_\theta(y_{t,n}^{(b)}\mid h_{t,n}^{(b)})\right], \tag{9}$$

where $N^{(b)} = \sum_t |y_t^{(b)}|$ is the total response length. Letting $\mathcal{A}_t$ collect the token positions inside the `<final_smiles>` span and $N_A = \sum_t |\mathcal{A}_t|$, distillation is confined to those positions and uses the gated consensus target:

$$\mathcal{L}_{\mathrm{KD}}(\theta) = \frac{1}{N_A}\sum_t\sum_{n\in\mathcal{A}_t}\gamma_{t,n}D_{\mathrm{KL}}(p_T(\cdot\mid h_{t,n})\|p_\theta(\cdot\mid h_{t,n}))\,. \tag{10}$$

We adopt forward KL rather than the reverse KL common in single-teacher distillation. Its gradient with respect to the student logits at position $n$ is $\gamma_{t,n}N_A^{-1}(p_\theta - p_T)$, so a teacher-supported token gains probability even when the student has not yet placed it in its top-$k$ set, with the gate scaling the update rather than reshaping its direction. Reverse KL penalizes only tokens the student already generates, and this mode-seeking behavior is too conservative in a domain where a single correct SMILES character may carry near-zero initial probability (Gu et al., 2024).

The final loss combines the two paths,

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{RL}}(\theta) + \mu\mathcal{L}_{\mathrm{KD}}(\theta). \tag{11}$$

We deliberately exclude the free-form rationale from distillation, since the teachers were trained independently and share no reasoning style; distilling their rationale distributions would inject a style-transfer signal competing with the task-bearing molecular signal, a pathology documented by Pan et al. (2026) in the self-distillation setting. Only $\pi_\theta$ is retained at inference.

## 4 Experimental Setup

**Training pipeline.** All policies share the same backbone; the main comparison uses Qwen3-4B. Continual pre-training (CPT) on local molecular transformations (Gururangan et al., 2020) yields $\theta_{\mathrm{cpt}}$, shared supervised fine-tuning adds multi-property supervision and recurrent replay to yield the student initializer $\theta_{\mathrm{sft}}$, and six full-parameter property teachers are then optimized independently with group-relative policy updates (Shao et al., 2024). Appendix C gives stage-by-stage data counts, verifier and optimizer settings, and held-out expert reliabilities.

**Baselines.** Baselines are grouped by where knowledge is transferred and whether supervision follows the student distribution: initialization and specialists (Base backbone, Shared SFT, the mean of six separately deployed property experts); parameter-space consolidation (Task Arithmetic (Ilharco et al., 2023)); verifier-only learning (Mixed RL, RePO (Li et al., 2026)); offline behavioral transfer (fine-tuning on fixed expert trajectories); and on-policy teacher composition. The last group contains Vanilla MOPD, which forms the consensus teacher distribution of Equation 4 by taking a uniformly weighted geometric mean of the requested teachers' token distributions and renormalizing, then distills it on student rollouts; Mixed RL + Vanilla MOPD, which adds verifier reward while keeping uniform weights; and RA-MOPD (distillation only), which removes the policy-gradient path to isolate the routed teacher signal. Table 1 summarizes what each method requires, and Appendix C.4 gives the full descriptions. All trainable post-SFT methods use the same prompt pool, parser, property functions and number of student updates unless reported as pure distillation.

**Metrics.** The primary metric is final-state cSR@1 from Equation 2. We additionally report IID and held-out-combination OOD cSR, validity and BestPrefix-cSR. A stricter margin evaluation replaces $\min_j g_j > 0$ with $\min_j g_j > \delta$ for $\delta \in \{0.1, 0.25, 0.5\}$.

**Evaluation discipline.** The benchmark is balanced across the ten combinations and evaluated with one deterministic parser and oracle stack.

Table 1: Post-SFT requirements. Expert training is a shared upstream cost for teacher-based methods. Checks denote required resources or signals; crosses denote their absence.

| Method | Experts | Student rollouts | Online logits | Verifier | Reliability |
|---|---|---|---|---|---|
| Property experts | ✓ | × | × | × | × |
| Task Arithmetic | ✓ | × | × | × | × |
| Mixed RL | × | ✓ | × | ✓ | × |
| Off-policy fine-tuning | ✓ | × | × | × | × |
| Vanilla MOPD | ✓ | ✓ | ✓ | × | × |
| Mixed RL + Vanilla MOPD | ✓ | ✓ | ✓ | ✓ | × |
| RA-MOPD (distillation only) | ✓ | ✓ | ✓ | × | ✓ |
| RA-MOPD | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 2: Complete comparison by property combination ($H = 2$, $K = 1$). The five IID combinations were observed during multi-objective training; the five OOD combinations were held out. All entries are constrained success rates; bold marks the best value in each column. The final row reports the gain of RA-MOPD over the untuned backbone.

| | Seen combinations (IID) | | | | | Held-out combinations (OOD) | | | | | Aggregate | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | T2 | T3 | T4 | T7 | T8 | T1 | T5 | T6 | T9 | T10 | IID | OOD | Overall |
| *Backbone and initialization* | | | | | | | | | | | | | |
| Base Qwen3-4B | 0.020 | 0.020 | 0.040 | 0.120 | 0.000 | 0.000 | 0.020 | 0.100 | 0.040 | 0.060 | 0.040 | 0.044 | 0.042 |
| Shared SFT | 0.240 | **0.140** | 0.480 | **0.300** | 0.060 | 0.300 | 0.260 | 0.280 | 0.120 | 0.240 | 0.244 | 0.240 | 0.242 |
| *Specialists and parameter-space transfer* | | | | | | | | | | | | | |
| Property experts (mean) | 0.310 | 0.133 | 0.470 | 0.303 | 0.090 | 0.293 | 0.257 | 0.303 | 0.157 | 0.213 | 0.261 | 0.245 | 0.253 |
| ParamMerge-TA | 0.360 | 0.120 | 0.560 | 0.280 | **0.140** | 0.300 | 0.180 | 0.260 | 0.100 | **0.300** | 0.292 | 0.228 | 0.260 |
| *Verifier-guided on-policy learning* | | | | | | | | | | | | | |
| Mixed RL | 0.360 | 0.120 | 0.560 | 0.240 | 0.120 | 0.260 | 0.220 | **0.340** | 0.120 | 0.200 | 0.280 | 0.228 | 0.254 |
| RePO | 0.340 | 0.120 | 0.620 | 0.200 | 0.080 | 0.260 | 0.200 | 0.280 | 0.100 | 0.260 | 0.272 | 0.220 | 0.246 |
| *Offline behavioral transfer* | | | | | | | | | | | | | |
| Off-policy fine-tuning | 0.340 | **0.140** | 0.520 | **0.300** | 0.100 | **0.320** | 0.260 | 0.280 | **0.160** | 0.220 | 0.280 | 0.248 | 0.264 |
| *On-policy teacher composition* | | | | | | | | | | | | | |
| Vanilla MOPD | 0.360 | 0.120 | 0.580 | 0.240 | 0.120 | 0.300 | 0.160 | **0.340** | 0.140 | 0.240 | 0.284 | 0.236 | 0.260 |
| Mixed RL + Vanilla MOPD | **0.420** | 0.120 | 0.580 | 0.240 | 0.080 | 0.260 | **0.280** | 0.320 | 0.120 | 0.200 | 0.288 | 0.236 | 0.262 |
| RA-MOPD (distillation only) | 0.360 | 0.120 | **0.640** | 0.220 | 0.080 | 0.300 | 0.260 | 0.320 | 0.140 | 0.220 | 0.284 | 0.248 | 0.266 |
| RA-MOPD | 0.400 | 0.100 | **0.640** | 0.260 | 0.100 | **0.320** | 0.240 | **0.340** | 0.120 | 0.280 | **0.300** | **0.260** | **0.280** |
| Δ over backbone | **+0.380** | **+0.080** | **+0.600** | **+0.140** | **+0.100** | **+0.320** | **+0.220** | **+0.240** | **+0.080** | **+0.220** | **+0.260** | **+0.216** | **+0.238** |

Table 3: Component ablations. Each variant removes one part of the final model; Δ is the change in overall cSR.

| Method | cSR | Δ | IID | OOD |
|---|---|---|---|---|
| RA-MOPD | **0.280** | — | **0.300** | **0.260** |
| w/o reliability prior | 0.262 | **−0.018** | 0.284 | 0.240 |
| w/o collateral-harm term | 0.254 | **−0.026** | 0.280 | 0.228 |
| w/o token gate | 0.258 | **−0.022** | 0.268 | 0.248 |

Table 4: Distillation design under fixed routing and policy reward. The upper block varies divergence on the full vocabulary; the lower row truncates to the student's top-256 support.

| Teacher signal | cSR | IID | OOD | BestPrefix |
|---|---|---|---|---|
| Full-vocabulary forward KL | **0.280** | **0.300** | **0.260** | **0.310** |
| Full-vocabulary JSD | 0.272 | 0.292 | 0.252 | 0.304 |
| Full-vocabulary reverse KL | 0.262 | 0.284 | 0.240 | 0.306 |
| Student top-256 forward KL | 0.266 | 0.280 | 0.252 | 0.296 |

# 5 Results

## 5.1 Main comparison

Table 2 reports the full comparison by property combination rather than aggregate scores alone. RA-MOPD attains the best Overall, IID and OOD cSR at 0.280, 0.300 and 0.260, improving Overall cSR by 0.038 over Shared SFT, 0.026 over Mixed RL, 0.020 over Vanilla MOPD and 0.016 over off-policy fine-tuning. Removing the verifier path costs 0.014, so the two supervision sources contribute separately. Teacher consensus supplying dense token-level supervision and the verifier evaluating the complete recurrent outcome. We test all the baselines on the full evaluation set ($N = 3{,}905$; Appendix E), where RA-MOPD still achieves the best performance.

## 5.2 Component ablations

All three mechanisms contribute (Table 3). Removing held-out reliability costs 1.8 points, dropping the collateral-harm term $H_{j,t}$ from the routing utility costs 2.6 points, and applying the routed teacher target without a disagreement gate costs 2.2 points. The ordering matches the design: reliability is expert-global, the collateral term is molecule- and state-specific, and token conflict is resolved only after the state-level mixture has been formed. The collateral-harm ablation attains marginally higher validity while losing the most cSR, which indicates that it produces chemically sound but conjunction-incomplete edits. Appendix Figure 5 shows why static parameter geometry is an insufficient substitute: expert task vectors are close to orthogonal, yet parameter-space similarity is a weak predictor of

Table 5: cSR under nontrivial all-property gain margins. All models receive the same deterministic parser, single <final_smiles> output format, $H = 2$ recurrent feedback and Tanimoto similarity constraint, ensuring a fair comparison protocol.

| Model | $\delta > 0.1$ | $\delta > 0.25$ | $\delta > 0.5$ |
|---|---|---|---|
| GPT-5.6-Sol | **0.278** | 0.176 | 0.114 |
| Claude-Sonnet-5 | 0.200 | 0.128 | 0.084 |
| DeepSeek-V4-Flash | 0.240 | 0.178 | 0.096 |
| Qwen3.6-Plus | 0.228 | 0.176 | 0.104 |
| RA-MOPD | 0.226 | **0.182** | **0.138** |

Table 6: Recurrent decomposition for RA-MOPD. BestPrefix reports success at either round.

| | Protocol | | |
|---|---|---|---|
| | $H = 1$ first-round result | $H = 2$ final result | $H = 2$ BestPrefix |
| Calls | 1 | 2 | 2 |
| cSR | 0.248 | **0.280** | 0.310 |
| $\Delta$ vs. $H = 1$ | — | **+0.032** | **+0.062** |
| IID | 0.268 | 0.300 | 0.324 |
| OOD | 0.228 | 0.260 | 0.296 |

behavioral transfer, which supports routing on verified molecular outcomes rather than on parameter proximity.

### 5.3 Divergence and Support

Forward KL on the exact vocabulary performs best (Table 4), gaining 0.8 points over JSD and 1.8 points over reverse KL, while a student-top-256 approximation loses 1.4 points relative to full-vocabulary. Both effects follow the forward-KL gradient of §3.4: exact forward KL can raise a teacher-supported token even when the student has not yet placed it in a retained top-$k$ set, and truncating the support removes precisely those positions. The result also cautions against importing a default divergence from generic MOPD, where reverse KL is common (Ma et al., 2026); support approximation and divergence should be chosen jointly for the target domain.

### 5.4 Margins and Recurrent Correction

Table 5 compares RA-MOPD against general-purpose models under nontrivial gain margins that exclude infinitesimal or incidental property shifts. At the moderate $\delta > 0.1$ threshold, general-purpose models lead, but the ranking reverses as the margin tightens: RA-MOPD is best at $\delta > 0.25$ and $\delta > 0.5$. These margins separate models that produce structurally meaningful molecular edits—changes that move every property by a measurable amount—from those whose successes concentrate near the zero-improvement boundary. The widening advantage under stricter thresholds suggests that specialist composition captures genuine structure–property relationships rather than surface-level modifications that happen to satisfy a lenient criterion.

The second edit raises final cSR by 3.2 points (Table 6). BestPrefix-cSR reaches $31.0\%$, decomposing into $6.2\%$ of prompts rescued after a first-round failure and $3.0\%$ that lose an earlier success, which accounts for the observed net gain. Recurrence is thus useful but not monotone. Figure 2 traces a representative repair: the first edit advances part of the requested set while leaving one property short, outcome-aware routing then raises the weight of the teacher owning that property, and the second edit restores the full conjunction while remaining above the similarity threshold. An uncertainty-aware stopping rule would retain the best prefix without exposing property oracles at deployment.

### 5.5 Scaling Behavior

Two budgets improve constrained success. Under stochastic sampling (Figure 3a), final cSR rises from $34.6\%$ at $K = 2$ to $45.8\%$ at $K = 4$ and $55.4\%$ at $K = 8$, a gain of $27.4$ points over the greedy single-candidate result, with BestPrefix-cSR reaching $57.0\%$. Across backbone sizes (Figure 3b), RA-MOPD is the strongest method at every scale and improves monotonically with capacity, from $24.6\%$ on Qwen3-1.7B to $28.0\%$ on Qwen3-4B and $31.6\%$ on Qwen3-8B. This is consistent with the scaling laws observed in general LLMs.

## 6 Related Work

**Language-based molecular optimization.** Early work optimizes molecular graphs or strings against scalarized objectives with search or policy gradients (Olivecrona et al., 2017; You et al., 2018; Zhou et al., 2019; Jin et al., 2020; Gao et al., 2022). Language interfaces to chemistry (Edwards et al.,

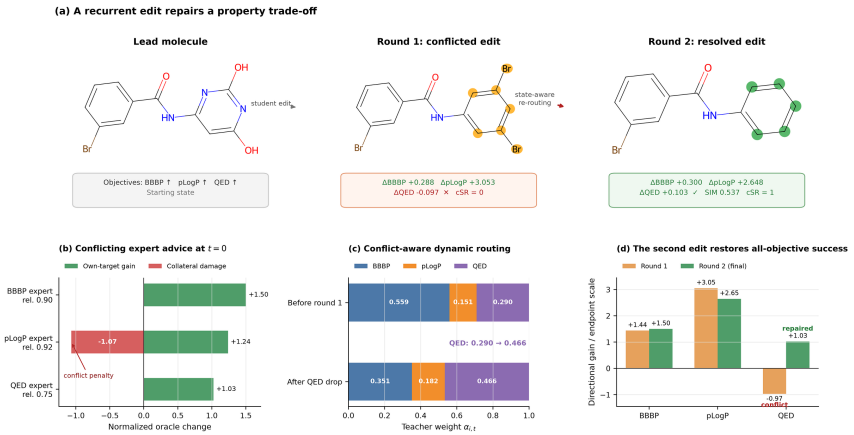
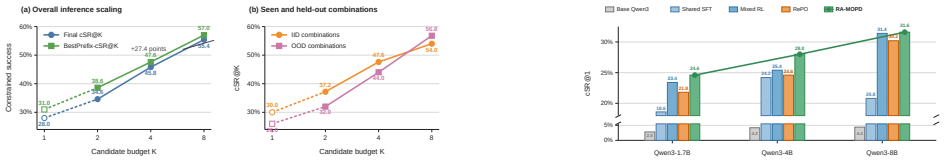



Figure 2: A recurrent repair under outcome-aware routing, on a BBBP + pLogP + QED prompt.




(a) Inference-time candidate budget

(b) Backbone parameter count

Figure 3: Two budgets. **(a)** Inference-time scaling from one stochastic $K = 8$ candidate pool at $H = 2$. **(b)** Model scaling under the common $H = 2$, $K = 1$ protocol. RA-MOPD is the strongest method at every backbone size.

2022; Fang et al., 2024) enable instruction-conditioned editing: GeLLM$^3$O formalizes instruction following over seen and unseen property combinations (Dey et al., 2025), RePO couples answer-level reference guidance with verifiable policy optimization (Li et al., 2026), and MolMem casts optimization as an oracle-budgeted agentic process (Wang et al., 2026). We retain natural-language reasoning and recurrent edits, and address a question these systems leave open: how separately optimized property policies can supervise a single multi-property model that does not use them at test time.

**Capability fusion and on-policy distillation.** Capabilities trained apart can be recombined in parameter space (Ilharco et al., 2023; Wortsman et al., 2022; Yadav et al., 2023) or in function space. Classical distillation transfers a fixed teacher's distribution (Hinton et al., 2015); on-policy variants evaluate the teacher on student rollouts to remove exposure bias (Lin et al., 2020; Agarwal et al., 2024; Gu et al., 2024), and multi-teacher formulations weight several teachers at once (You et al., 2017; Liu et al., 2020). MOPD applies this to domain RL teachers (Ma et al., 2026), with follow-up work comparing fusion paradigms (Wu et al., 2026), diagnosing capability imbalance (Gao et al., 2026) and extending to video specialists (Qin et al., 2026). A prompt in these settings is typically dispatched to one task teacher; we instead compose several simultaneously relevant teachers.

## 7 Conclusion

We presented RA-MOPD, a framework that distills a bank of independently trained single-property molecular experts into one student policy serving arbitrary multi-property prompts at inference time. The key idea is to resolve expert heterogeneity at two complementary resolutions: state-level routing selects which experts to trust based on verified molecular outcomes and a held-out reliability prior, while a token-level product-of-experts gate modulates the distillation signal according to how closely the trusted experts agree. We proved that the gate's normalizer equals the minimum aggregate reverse-KL divergence among the active experts, grounding it as a measure of irreducible disagreement rather than a tuned heuristic. Across ten property combinations—including five held-out at training time—RA-MOPD consistently outperforms verifier-only RL, offline imitation, parameter merging and uniform multi-teacher distillation, and ablations confirm that each component contributes to the overall gain.

## AI USE STATEMENT

This submission was prepared with the assistance of LLMs, which were only utilized for refining content and checking grammar. The authors assume full responsibility for the entire content of the manuscript, including any potential issues related to plagiarism and factual accuracy. It is confirmed that no LLM is listed as an author.

## ETHICS STATEMENT

This research aims to improve the reliability of AI-assisted molecular optimization for scientific discovery. Our experiments are conducted exclusively on public molecular datasets and computational benchmarks, which contain no sensitive or personally identifiable information. Although we use pretrained language models to generate candidate molecular edits, all outputs are filtered and evaluated using deterministic parsers, chemical-validity checks, structural-similarity constraints, and frozen property predictors. The generated candidates are intended solely for computational research and should not be interpreted as experimentally validated compounds or clinical recommendations. We therefore foresee no significant ethical concerns arising directly from this work.

## REPRODUCIBILITY STATEMENT

We are committed to the full reproducibility of our research. We will release our source code for the framework, fine-tuning scripts, and evaluation protocols. The supplement provides the full algorithmic core for state-local routing, exact full-vocabulary product-of-experts accumulation, conflict gating, answer-span masking, verifier rewards and the combined objective, together with the frozen backbone configuration and unit tests. Appendix C reports data counts, property scales, reliabilities and training requirements. These resources will enable the community to verify our findings and build upon our work.

## REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, 2012.

Vishal Dey, Xiao Hu, and Xia Ning. GeLLM$^3$O: Generalizing large language models for multi-property molecule optimization. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25192–25221, 2025. URL https://arxiv.org/abs/2502.13398.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 375–413, 2022.

Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1):8, 2009.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: a large-scale biomolecular instruction dataset for large language models. In *International Conference on Learning Representations*, 2024.

Huan-ang Gao, Haohan Chi, Yong Yan, Shiyuan Feng, Hanlin Wu, Zheng Jiang, Bingxiang He, Wei-Ying Ma, Ya-Qin Zhang, and Hao Zhou. Open-MOPD: Diagnosing and fixing capability imbalance in multi-teacher on-policy distillation. *arXiv preprint arXiv:2608.19098*, 2026. URL https://arxiv.org/abs/2608.19098.

Wenhao Gao, Tianfan Fu, Jimeng Sun, and Connor W. Coley. Sample efficiency matters: a benchmark for practical molecular optimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21342–21357, 2022.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations*, 2024.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8342–8360, 2020.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2323–2332. PMLR, 2018. URL https://proceedings.mlr.press/v80/jin18a.html.

Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In *International Conference on Machine Learning*, pp. 4839–4848, 2020.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024. URL https://arxiv.org/abs/2411.15124.

Xuan Li, Zhanke Zhou, Zongze Li, Jiangchao Yao, Yu Rong, Lu Zhang, and Bo Han. Reference-guided policy optimization for molecular optimization via LLM reasoning. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.05900.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive knowledge distillation through imitation learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6121–6133, 2020.

Yuang Liu, Wei Zhang, and Jun Wang. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, 415:106–113, 2020. doi: 10.1016/j.neucom.2020.07.048. URL https://doi.org/10.1016/j.neucom.2020.07.048.

Wenhan Ma, Jianyu Wei, Liang Zhao, Hailin Zhang, Bangjun Xiao, Lei Li, Qibin Yang, Bofei Gao, Yudong Wang, Rang Li, Jinhao Dong, Zhifang Sui, and Fuli Luo. MOPD: Multi-teacher on-policy distillation for capability integration in LLM post-training. *arXiv preprint arXiv:2606.30406*, 2026. URL https://arxiv.org/abs/2606.30406.

Marcus Olivecrona, Thomas Blaschke, Ola Engkvist, and Hongming Chen. Molecular de-novo design through deep reinforcement learning. *Journal of Cheminformatics*, 9(1):48, 2017.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. RLCSD: Reinforcement learning with contrastive on-policy self-distillation. *arXiv preprint arXiv:2606.11709*, 2026. URL https://arxiv.org/abs/2606.11709.

Zhenxin Qin, Peng Shi, Cong Han, Yinlong Qian, Zequn Jie, and Lin Ma. Video-MOPD: Multi-teacher on-policy distillation for video understanding. *arXiv preprint arXiv:2609.09300*, 2026. URL https://arxiv.org/abs/2609.09300.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010.

Stephane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning, 2011. URL https://arxiv.org/abs/1011.0686.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Ziqing Wang, Yibo Wen, Abhishek Pandey, Han Liu, and Kaize Ding. MolMem: Memory-augmented agentic reinforcement learning for sample-efficient molecular optimization. *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43694–43712, 2026. URL https://arxiv.org/abs/2604.12237.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time, 2022. URL https://arxiv.org/abs/2203.05482.

Siye Wu, Kai Yang, Yuchen Cai, Xin Xu, Peng-Yuan Wang, Jiaxuan Wang, Jiashun Liu, Jiafei Lyu, Yangkun Chen, Saiyong Yang, and Yanghua Xiao. Consolidating RLVR capabilities across domains: a deep dive into fusion paradigms. *arXiv preprint arXiv:2608.27409*, 2026. URL https://arxiv.org/abs/2608.27409.

Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. TIES-Merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled RLVR. *arXiv preprint arXiv:2604.03128*, 2026. URL https://arxiv.org/abs/2604.03128.

Jiaxuan You, Bowen Liu, Rex Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1285–1294. Association for Computing Machinery, 2017. doi: 10.1145/3097983.3098135. URL https://doi.org/10.1145/3097983.3098135.

Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N. Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific Reports*, 9(1):10752, 2019.

# A LIMITATIONS

Several limitations are shared with the broader field of computational molecular optimization. Learned property functions are imperfect surrogates for experimental assays, so improvements measured by computational oracles do not guarantee wet-lab activity and may reflect oracle exploitation rather than genuine molecular improvement. The distribution shift between training molecules and prospective candidates remains an open problem across all generative molecular design methods, and the extent to which any policy trained on historical data generalizes to novel chemical series is not yet well characterized.

# B PROPERTY COMBINATIONS

The evaluation set provides 50 prompts for each of ten tasks. The five combinations observed during multi-objective training are

$$\begin{aligned} &\mathrm{T2}:\mathrm{B{+}D{+}P}, && \mathrm{T3}:\mathrm{B{+}D{+}P{+}Q}, && \mathrm{T4}:\mathrm{B{+}D{+}Q},\\ &\mathrm{T7}:\mathrm{B{+}P{+}Q}, && \mathrm{T8}:\mathrm{D{+}P{+}Q}, \end{aligned}$$

and the five held-out combinations are

$$\begin{aligned} &\mathrm{T1}:\mathrm{B{+}D{+}M{+}Q}, && \mathrm{T5}:\mathrm{B{+}H{+}M{+}Q}, && \mathrm{T6}:\mathrm{B{+}M{+}P{+}Q},\\ &\mathrm{T9}:\mathrm{H{+}M{+}P{+}Q}, && \mathrm{T10}:\mathrm{M{+}P{+}Q}, \end{aligned}$$

where B, D, H, M, P and Q abbreviate BBBP, DRD2, HIA, mutagenicity, penalized logP and QED. OOD therefore denotes an unseen combination of familiar properties rather than scaffold or assay shift.

# C IMPLEMENTATION AND REPRODUCIBILITY DETAILS

## C.1 END-TO-END PIPELINE

Table 7 summarizes the four-stage pipeline that produces the student model. Every trainable policy shares the same Qwen3-4B backbone. The stages are sequential: continual pre-training establishes chemical validity, shared supervised fine-tuning adds multi-property grounding, independent single-property reinforcement learning produces the frozen expert bank, and RA-MOPD distills the experts into a single student on its own rollouts. Only the final student is retained at inference.

## C.2 VERIFIER AND OPTIMIZATION

The verifier combines molecular progress, graph-grounded reasoning and calibrated confidence:

$$R = R_{\mathrm{num}} + 0.2R_{\mathrm{reason}} + 0.1R_{\mathrm{cal}}. \tag{12}$$

Invalid molecules receive $R_{\mathrm{num}} = -1$. For a valid candidate $m$,

$$R_{\mathrm{num}} = S_i(m) + 0.5\,\mathrm{clip}\Big(\min_{j\in P_i} g_j(m;m_0), -1, 1\Big) - 0.5\left[\frac{\eta - \mathrm{Sim}(m,m_0)}{\eta}\right]_+. \tag{13}$$

$R_{\mathrm{reason}}$ aggregates deterministic entity precision, edit-claim precision, concision and absence of invented measurements. $R_{\mathrm{cal}}$ is a Brier-score complement between parsed verbal confidence and verified constrained success. No learned language-model judge enters the reward.

Student training uses $H = 2$, group size 4, 80 optimizer updates, $\mu = 10$, $\lambda_\rho = 5$, $\lambda_s = 4$, $\tau_\alpha = 1$, $\tau_C = 0.05$ and learning rate $2 \times 10^{-6}$. Full-vocabulary teacher distributions are accumulated sequentially in float32. All trainable models are full-parameter rather than LoRA adapters; training uses bfloat16, gradient checkpointing and deterministic prompt schedules.

Table 7: The RA-MOPD pipeline. Every trainable policy uses the same backbone; the final product is one student model requiring neither teachers nor property functions at inference.

| Stage | Data | Key settings and products |
|---|---|---|
| I<br>Continual pre-training (CPT) | Local transformations, singleton supervision and rationales, similarity-repair replay | Trains a chemically valid, high-similarity editor; product $\theta_{\mathrm{cpt}}$. |
| II<br>Shared multi-property SFT | Multi-objective direct examples and rationales, grounded recurrent replay | Adds seen-combination supervision and recurrent state grounding; product $\theta_{\mathrm{sft}}$, the common student initializer. |
| III<br>Property RL teachers | Six single-property verifier-training streams | Independent 300-update RL runs; all six teachers branch from $\theta_{\mathrm{sft}}$; products $\{\pi_{\phi_j}\}_{j\in\mathcal{P}}$ and reliabilities $\{\rho_j\}$. |
| IV<br>RA-MOPD student | 10,000 multi-property prompts | On-policy recurrent rollouts with $H = 2$, group size 4, reliability- and outcome-aware routing, full-vocabulary consensus target, conflict gating, and verifier learning; product $\pi_\theta$. |
| Evaluation | Balanced benchmark and source-disjoint confirmation set | Greedy $K = 1$ evaluation over five seen and five held-out property combinations; report cSR, validity, and BestPrefix-cSR. |

### C.3 Training algorithm

Algorithm 1 gives the complete optimization loop for the RA-MOPD student (Stage IV). Each iteration samples on-policy recurrent trajectories, evaluates the verifier reward, computes state-level routing weights from teacher proposals, forms the token-level consensus target with conflict gating, and updates the student with the combined policy-gradient and distillation loss.

**Algorithm 1** RA-MOPD: reliability-aware multi-teacher on-policy distillation

**Require:** Student $\pi_\theta$, frozen teachers $\{\pi_{\phi_j}\}$, reliabilities $\{\rho_j\}$, property functions $\{f_j\}$, prompt set $\mathcal{D}$, group size $B$, horizon $H$

**for** each training iteration **do**
  Sample prompts $x_i = (m_0, P_i)$ from $\mathcal{D}$
  **for all** $x_i$ in the minibatch **do**
    Sample $B$ recurrent trajectories $\{\xi^{(b)}\}_{b=1}^{B} \sim \pi_\theta$ for $H$ rounds
    Evaluate $R^{(b)}$ and compute group-relative advantages $\widehat{A}^{(b)}$
    **for all** student-visited molecular states $m_t$ **do**
      **for all** $j \in P_i$ **do**
        Generate proposal $c_{j,t}$ and evaluate $(G_{j,t}, H_{j,t}, V_{j,t})$
      **end for**
      Compute routing weights $\{\alpha_{j,t}\}_{j\in P_i}$ by Equation 6
      **for all** prefixes $h_{t,n}$ in the tagged final-SMILES span **do**
        Evaluate full-vocabulary distributions $\{p_j(\cdot \mid h_{t,n})\}_{j\in P_i}$
        Form consensus target $p_T$ and exact conflict $C_{t,n} = -Z_{t,n}$
        Set supervision gate $\gamma_{t,n} = \exp(-C_{t,n}/\tau_C)$
      **end for**
    **end for**
  **end for**
  Compute $\mathcal{L}_{\mathrm{RL}}$ on complete responses and $\mathcal{L}_{\mathrm{KD}}$ on final-SMILES tokens
  Update $\theta$ by minimizing $\mathcal{L}_{\mathrm{RL}} + \mu\mathcal{L}_{\mathrm{KD}}$
**end for**

**Ensure:** Updated student $\pi_\theta$

### C.4 BASELINE DESCRIPTIONS

**Initialization and specialists.** *Base* is the untuned backbone. *Shared SFT* is the stage-II checkpoint $\theta_{\text{sft}}$ that initializes every post-SFT method. *Property experts (mean)* reports the average performance of the six single-property teachers deployed separately on the multi-property benchmark, and measures what a specialist achieves when asked for a conjunction it was never trained to satisfy.

**Parameter-space consolidation.** *ParamMerge-TA* applies task arithmetic (Ilharco et al., 2023): each expert contributes an origin-corrected parameter update $\tau_j = \phi_j - \theta_{\text{sft}}$, and the merged model is $\theta_{\text{sft}} + \lambda \sum_j \tau_j$ with $\lambda$ selected on a held-out split. No student rollouts are involved, so the composition cannot depend on the molecular state.

**Verifier-only learning.** *Mixed RL* trains a single shared policy on the multi-property prompt pool with the same verifier and group-relative updates as RA-MOPD, but without any teacher signal. *RePO* adds answer-level reference guidance to verifiable policy optimization (Li et al., 2026).

**Offline behavioral transfer.** *Off-policy fine-tuning* collects trajectories from the six experts on the training prompts and fine-tunes the student on them. This supplies dense targets but does not correct the mismatch between expert-visited and student-visited states.

**On-policy teacher composition.** *Vanilla MOPD* follows the standard multi-teacher construction: at each student-visited prefix, the token distributions of the requested teachers are combined by a weighted geometric mean with uniform weights and renormalized, giving the consensus teacher distribution of Equation 4, which is then distilled on the student's own rollouts. It uses neither reliability, state-local outcomes, nor a disagreement gate. *Mixed RL + Vanilla MOPD* adds the verifier reward to that consensus target while keeping uniform weights, and isolates the contribution of weighting from the contribution of environmental feedback. RA-MOPD *(distillation only)* keeps reliability- and outcome-aware routing together with the conflict gate but removes the policy-gradient path, isolating the routed teacher signal.

### C.5 PROPERTY METADATA AND EXPERT RELIABILITY

Table 8: Property direction, normalization scale, and training-molecule-clean expert reliability. Reliability is the Beta$(1, 1)$ posterior mean.

| Property | Direction | $\sigma_j$ | $n$ | Expert cSR | $\rho_j$ |
|---|---|---|---|---|---|
| BBBP | $+1$ | 0.2 | 191 | 0.906 | 0.902 |
| DRD2 | $+1$ | 0.2 | 182 | 0.797 | 0.793 |
| HIA | $+1$ | 0.1 | 201 | 0.891 | 0.887 |
| Mutagenicity | $-1$ | 0.1 | 204 | 0.765 | 0.762 |
| Penalized logP | $+1$ | 1.0 | 261 | 0.923 | 0.920 |
| QED | $+1$ | 0.1 | 257 | 0.755 | 0.753 |

Figure 4 shows that the six policies acquire property-specific progress at different reliabilities. These values enter routing only through the prior term of Equation 6 and do not determine the final weights without state-local proposal evaluation.

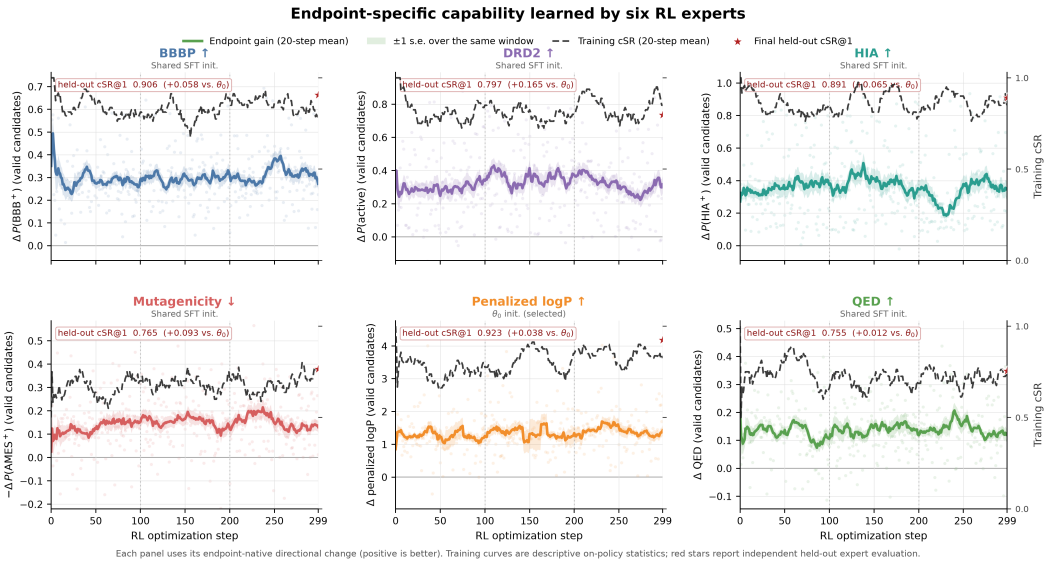



Figure 4: Property-native progress during 300 RL updates for six specialists. Positive is always the requested direction. Markers report independent held-out single-property cSR.

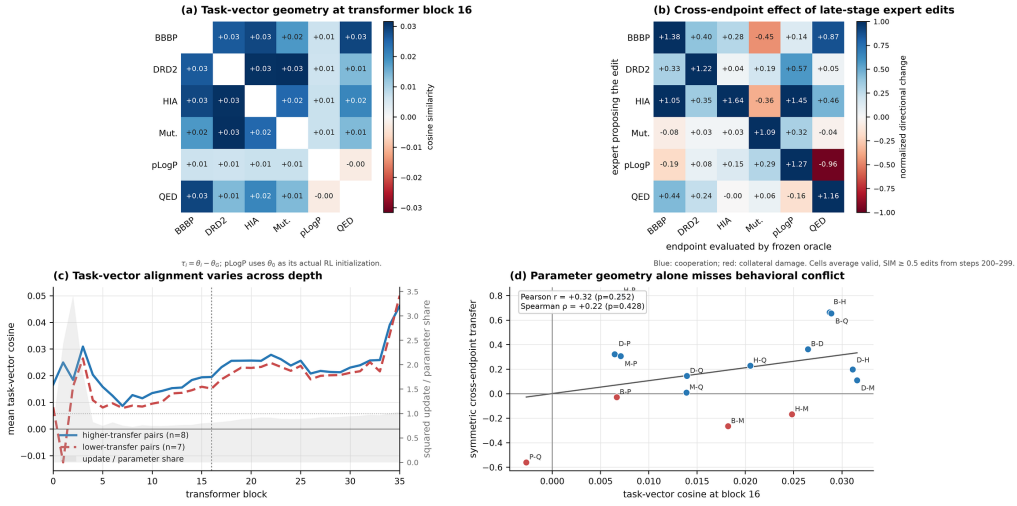



Figure 5: Expert heterogeneity. (a) Task-vector cosine similarity at an intermediate transformer block. (b) Directional cross-property effects of valid late-stage expert edits. (c) Parameter alignment across depth. (d) Parameter similarity is a weak proxy for behavioral transfer (Pearson $r = 0.32$, $p = 0.252$; Spearman $\rho = 0.22$, $p = 0.428$), which motivates routing on verified outcomes rather than parameter proximity.

### C.6 Training Data

Table 9: Materialized training data for the Qwen3-4B pipeline.

| Stage | Rows | Purpose |
|---|---|---|
| CPT (local edits) | 20,000 | Valid, localized transformations |
| Singleton direct SFT | 30,000 | Property-specific answer supervision |
| Singleton rationales | 5,956 | Short chemical explanations |
| Similarity repair | 10,154 | Restore margin above similarity constraint |
| Multi-objective direct SFT | 17,309 | Seen-combination references |
| Multi-objective rationales | 2,481 | Audited multi-property explanations |
| Grounded recurrent replay | 9,980 | First-step state/action grounding |
| Selective step-2 replay | 3,333 | Recurrent state-distribution exposure |
| MOPD prompt pool | 10,000 | On-policy multi-objective prompts |

The data pipeline canonicalizes SMILES before overlap checks. A generated rationale is retained only when its target molecule is verified and its named entities appear in the source or target. The

response format remains language-native: short prose followed by exactly one tagged SMILES, with no hidden evidence fields at inference.

### C.7 Training cost

Table 10: Post-SFT training cost on a single 8×A100 node. *Fusion GPU-h* is the method-specific cost beyond the shared upstream stages; *Total GPU-h* includes the shared expert-training budget (7.71 h for the six-expert pool) where applicable. Checks denote required resources; crosses denote their absence.

| Method | Experts | Student rollouts | Online logits | Verifier | Reliability | Offline trajectories | Fusion GPU-h↓ | Total GPU-h↓ |
|---|---|---|---|---|---|---|---|---|
| Property experts (mean) | ✓ | × | × | × | × | × | — | 7.71 |
| ParamMerge-TA | ✓ | × | × | × | × | × | 0 | 7.71 |
| Mixed RL | × | ✓ | × | ✓ | × | × | 0.67 | 7.19 |
| Off-policy fine-tuning | ✓ | × | × | ✓* | × | ✓ | 2.13 | 9.85 |
| Vanilla MOPD | ✓ | ✓ | ✓ | × | × | × | 1.19 | 8.90 |
| Mixed RL + Vanilla MOPD | ✓ | ✓ | ✓ | ✓ | × | × | 1.20 | 8.92 |
| RA-MOPD (distill. only) | ✓ | ✓ | ✓ | × | ✓ | × | 5.77 | 13.48 |
| RA-MOPD | ✓ | ✓ | ✓ | ✓ | ✓ | × | 2.43 | 10.14 |

RA-MOPD adds 2.43 fusion GPU-hours beyond the shared expert pool, for a total of 10.14 GPU-hours. The dominant additional cost is the full-vocabulary teacher forward passes required at every student-visited prefix. No expert or property oracle is queried at inference, so the training-time overhead is amortized across all downstream deployment.

## D Proofs and Additional Discussion

We restate the two auxiliary results deferred from §3 and then prove all three claims.

**Proposition 2** (Conflict is bounded by the nearest expert)**.** *Under the assumptions of Proposition 1,*

$$0 \;\le\; C_{t,n} \;\le\; \min_{k \in P_i} \sum_{j \in P_i} \alpha_{j,t} D_{\mathrm{KL}}(p_k \| p_j), \tag{14}$$

*and $C_{t,n} = 0$ whenever a single teacher is active.*

**Proposition 3** (Routing as entropy-regularized allocation)**.** *Let $u_t = (u_{j,t})_{j \in P_i}$ and $\tau_\alpha > 0$, and let $\Delta(P_i)$ be the probability simplex over the active teachers. Then the weights of Equation 6 satisfy*

$$\alpha_t = \arg\max_{\alpha \in \Delta(P_i)} \Big\{ \textstyle\sum_{j \in P_i} \alpha_j u_{j,t} + \tau_\alpha \mathcal{H}(\alpha) \Big\}, \qquad \mathcal{H}(\alpha) = -\textstyle\sum_j \alpha_j \log \alpha_j, \tag{15}$$

*and at $\tau_\alpha = 1$ they factorize as $\alpha_{j,t} \propto (\rho_j + \epsilon)^{\lambda_\rho} \exp(G_{j,t} - H_{j,t} - \lambda_s V_{j,t})$.*

*Proof of Proposition 1.* Suppressing the state and token indices, for any distribution $q$ on the common support,

$$\sum_j \alpha_j D_{\mathrm{KL}}(q \| p_j) = \sum_v q(v) \log \frac{q(v)}{\prod_j p_j(v)^{\alpha_j}} = \sum_v q(v) \log \frac{q(v)}{\exp s(v)} = D_{\mathrm{KL}}(q \| p_T) - Z,$$

where the last equality uses $p_T = e^{s-Z}$. The first term is uniquely minimized at $q = p_T$, so the minimum value is $-Z$, attained by the product distribution. Non-negativity holds because the left-hand side is a convex combination of KL divergences. Under strictly positive support the minimum is zero exactly when $p_T = p_j$ for every active $j$, which is equivalent to all active teachers being identical. □

*Proof of Proposition 2.* The lower bound is Proposition 1. For the upper bound, each $p_k$ with $k \in P_i$ is feasible in the minimization of Equation 8, so $C \le \sum_j \alpha_j D_{\mathrm{KL}}(p_k \| p_j)$ for every $k$; minimizing the right-hand side over $k$ gives the claim. When a single teacher is active, $\alpha$ is a point mass and the bound evaluates to zero, which forces $C = 0$ and $\gamma = 1$. □

*Proof of Proposition 3.* The objective of Equation 15 is strictly concave on $\Delta(P_i)$ and the constraint set is compact, so the maximizer is unique. Introducing a multiplier $\nu$ for $\sum_j \alpha_j = 1$ and setting the derivative in $\alpha_j$ to zero gives $u_{j,t} - \tau_\alpha(\log \alpha_j + 1) - \nu = 0$, hence $\alpha_j \propto \exp(u_{j,t}/\tau_\alpha)$, which is Equation 6; the interior solution satisfies $\alpha_j > 0$, so the non-negativity constraints are inactive. Substituting $u_{j,t} = \lambda_\rho \log(\rho_j + \epsilon) + G_{j,t} - H_{j,t} - \lambda_s V_{j,t}$ at $\tau_\alpha = 1$ and exponentiating yields $\alpha_{j,t} \propto (\rho_j + \epsilon)^{\lambda_\rho} \exp(G_{j,t} - H_{j,t} - \lambda_s V_{j,t})$. □

The temperature $\tau_\alpha$ interpolates between uniform composition ($\tau_\alpha \to \infty$), which recovers Equation 4, and hard selection of the single best-scoring teacher ($\tau_\alpha \to 0$).

The identity of Proposition 1 resolves three edge cases. With one active expert $Z = 0$ and the gate reduces to ordinary teacher distillation. Identical experts incur no artificial penalty even when their shared distribution has high entropy. Confident experts concentrated on disjoint tokens produce a large $C = -Z$, so the objective declines to force the student toward a geometric compromise. This behavior differs from entropy gating, which would penalize the second case and tolerate the third.

Exact gating reuses the full-distribution logits already required by Equation 7, so it adds no teacher forward passes. Autoregressive softmax guarantees strictly positive support; in float32 we clamp only negligible negative roundoff in $C$ to zero.

## E Full-Set Evaluation

The main-text results use a compact balanced benchmark of 500 prompts (50 per combination) for rapid evaluation and controlled per-combination comparison. To verify that the findings are not an artifact of the compact set size, we evaluate all methods on the full evaluation set of $N = 3{,}905$ prompts under the same $H = 2$, $K = 1$ protocol. Table 11 reports the task distribution; the five IID combinations each contribute 500 prompts, while the five OOD combinations reflect the naturally available lead molecules and range from 96 to 500 prompts.

Table 11: Task distribution of the full evaluation set ($N = 3{,}905$).

| Task | Property combination | Prompts | Fraction |
|---|---|---|---|
| *Seen combinations (IID)* | | | |
| T2 | B+D+P | 500 | 0.128 |
| T3 | B+D+P+Q | 500 | 0.128 |
| T4 | B+D+Q | 500 | 0.128 |
| T7 | B+P+Q | 500 | 0.128 |
| T8 | D+P+Q | 500 | 0.128 |
| *Held-out combinations (OOD)* | | | |
| T1 | B+D+M+Q | 500 | 0.128 |
| T5 | B+H+M+Q | 118 | 0.030 |
| T6 | B+M+P+Q | 191 | 0.049 |
| T9 | H+M+P+Q | 96 | 0.025 |
| T10 | M+P+Q | 500 | 0.128 |
| Total | | 3,905 | 1.000 |

Table 12 reports the full-set constrained success rates. RA-MOPD attains the highest overall cSR at 0.307, confirming the compact-set ranking. The full-set results sharpen several observations: Vanilla MOPD and Mixed RL + Vanilla MOPD are closer to RA-MOPD than on the compact set, and the gap between on-policy teacher-composition methods and the remaining baselines is more pronounced. Component ablations maintain the same ordering as on the compact set, with the collateral-harm term contributing the largest single drop ($-0.026$), followed by the token gate ($-0.023$) and the reliability prior ($-0.018$).

The compact 500-prompt benchmark serves as a fast, balanced proxy for method comparison, and the full-set results confirm that the relative ordering and component contributions are preserved at nearly eight times the evaluation scale.

Table 12: Full-set comparison ($N = 3{,}905$, $H = 2$, $K = 1$). Bold marks the best cSR.

| Method | cSR |
|---|---|
| *Backbone and initialization* | |
| Base Qwen3-4B | 0.041 |
| CPT ($\theta_{\mathrm{cpt}}$) | 0.186 |
| Shared SFT | 0.235 |
| *Specialists and parameter-space transfer* | |
| Property experts (mean) | 0.275 |
| ParamMerge-TA | 0.294 |
| *Verifier-guided on-policy learning* | |
| Mixed RL | 0.288 |
| RePO | 0.281 |
| *Offline behavioral transfer* | |
| Off-policy fine-tuning | 0.273 |
| *On-policy teacher composition* | |
| Vanilla MOPD | 0.300 |
| Mixed RL + Vanilla MOPD | 0.301 |
| RA-MOPD (distillation only) | 0.295 |
| RA-MOPD | **0.307** |
| *Component ablations* | |
| w/o reliability prior | 0.289 |
| w/o collateral-harm term | 0.281 |
| w/o token gate | 0.284 |