TRAINING SPECIALIST MODELS WITHOUT REASONING TRAJECTORIES FOR DOMAIN EXPERT DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Specialist distillation effectively transfers domain expertise to student models via teacher-generated reasoning trajectories. However, when these specialists are trained solely on question–answer pairs without explicit reasoning supervision, what governs the trajectories they generate? In this work, we show that specialist optimization implicitly selects from this latent trajectory space. To isolate and observe this latent distribution, we leverage student distillation not as a downstream goal, but as an agnostic probe—since students inherit no parameterization or optimization constraints from the specialist, inheriting only the sampled trajectories themselves. Through this probe, our empirical analysis unveils a tight governing relationship: across 27 specialist–student pairings, their specialization–generalization profiles correlate exceptionally strongly. Crucially, explicitly controlling the specialist's distributional drift systematically shifts both the teacher and its distilled student along a controllable trade-off between domain precision and general-capability retention. Across chemistry, physics, and multilingual settings, distilled students systematically reflect these specialist-induced profiles, even across divergent model families. Our findings establish a new view of specialist training: when gold reasoning is absent, tuning choices directly control the latent supervision passed to downstream models. [1]

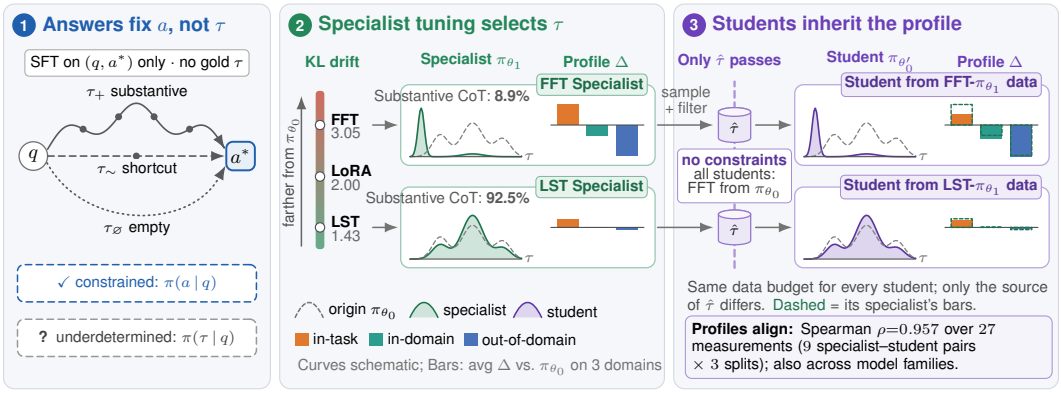


Figure 1: **Specialist optimization selects the latent supervision that students inherit.** **(1)** SFT on (q, a^*) constrains $\pi(a \mid q)$ but leaves $\pi(\tau \mid q)$ underdetermined. **(2)** Tuning resolves this via the specialist's drift from π_{θ_0}: unconstrained FFT drifts furthest and collapses trajectories; Structural constraint of *Layer Selective Tuning* (LST) keeps it close. **(3)** Students, all FFT-trained from π_{θ_0} on equal budgets and differing only in $\hat{\tau}$, reproduce their specialist's profile: specialization (in-task gain) vs. generalization (in-domain and out-of-domain retention). Bars: Tab. 2; KL drift: Fig. 2b.

[1] The code is available at an anonymous repository `https://anonymous.4open.science/r/specialist-distillation-iclr-D629/`. Trained model artifacts and all inference data will be released publicly upon acceptance.

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.

1 INTRODUCTION

Specialist distillation (Hinton et al., 2015; Fu et al., 2023; Hsieh et al., 2023; Ho et al., 2023; Magister et al., 2023; Yao et al., 2021; Liu et al., 2024) transfers domain expertise through an intermediate specialist model. A general-purpose *origin* model π_θ, e.g., Qwen3-8B (Yang et al., 2025), is first adapted to a target domain, and the resulting *specialist* then generates reasoning trajectories that serve as supervision for a downstream student (Mukherjee et al., 2023; Xu et al., 2024). Yet in most specialized domains, the specialist itself is never explicitly taught how to reason. Domain datasets typically provide (q, a) question–answer (QA) pairs (Uesato et al., 2022; Chan et al., 2022; Lightman et al., 2024; Turpin et al., 2023) but no gold trajectories (Zelikman et al., 2022; Gülçehre et al., 2023; Singh et al., 2024), because expert reasoning is difficult to obtain and verify. Since the specialist is optimized only against final answers, what governs the trajectories it generates?

Let τ denote a free-running reasoning trajectory generated for question q. We consider representative behaviors including a substantive, answer-consistent path (τ_+; Creswell & Shanahan, 2022; Turpin et al., 2023), a shortcut ($\tau_\sim$; Geirhos et al., 2020), and an empty one ($\tau_\varnothing$). Standard teacher-forced QA training directly optimizes answer likelihood and does not explicitly supervise these trajectories. At generation time, however, the trained model induces a distribution over them. Conceptually,

$$\log \pi_\theta(a \mid q) = \log \sum_{\tau \to a} \pi_\theta(\tau \mid q), \tag{1}$$

where $\tau \to a$ denotes a trajectory compatible with answer a. Eq. 1 characterizes generation-time behavior rather than the implemented SFT objective. Since answer-level supervision provides no direct preference among $\{\tau_+, \tau_\sim, \tau_\varnothing, \cdots\}$, multiple trajectory distributions may remain compatible with the same supervised answer (Fig. 1, panel 1).

To isolate and observe this latent selection, we repurpose student distillation as an agnostic probe rather than treating it only as a downstream goal. A student inherits neither the specialist's parameters nor the optimization constraints used to obtain them; it receives only the specialist-generated supervision. In our controlled pipeline, students share the same initialization and training configuration and are trained on equal amounts of sampled and filtered data. Differences among students thus expose what these reasoning trajectories actually carry (Fig. 1, panel 3), even though students never receive the specialist's parameters or optimization constraints directly.

Through this diagnostic probe, we uncover a tight link between specialists and their distilled students. Across chemistry, physics, and multilingual benchmarks, the balance between domain specialization and general capability in specialists systematically transfers to their downstream students. Across nine specialist-student pairs covering 27 distinct evaluations, these performance profiles align remarkably well (Spearman $\rho = 0.9573$, permutation test $p - \text{value} = 0.0093$, see § 4.2 for details). This pattern holds even when teacher and student models belong to completely different model families, showing that shared teacher–student parameterization is not necessary for the observed transfer. Distilled students therefore expose how the latent supervision selected by specialist optimization shapes downstream capability profiles.

Crucially, this selection process is both observable and controllable. Limiting how far a specialist drifts from its original base model directly recalibrates its trade-off between domain mastery and general capabilities, guiding the student model along with it. Layer-Selective Tuning (LST; Gao et al., 2025) provides an implicit anchor, yielding lower behavioral Kullback-Leibler (KL) divergence and better preserving reasoning structures than standard full fine-tuning (Fig. 1, panel 2). For explicit control, Anchored Supervised Fine-Tuning (ASFT; Zhu et al., 2026) provides a complementary explicit intervention: varying its anchoring strength systematically moves both specialist and student along the same trade-off. Together, these results identify distributional drift as a controllable axis of the latent supervision passed downstream.

Our main contributions are:

- **Specialist optimization is the key design variable for distillation data.** Under QA-only training, reasoning trajectories remain underdetermined by answer labels; the specialist's optimization procedure selects the trajectory distribution from which downstream supervision is generated.
- **Distilled students reveal the supervision selected by their specialists.** Using student distillation as an agnostic probe, we uncover a strong correspondence between specialist and student specialization–generalization profiles across domains and model families. This inheritance shows

that the effects of specialist optimization are encoded in the generated trajectories and transferred downstream, rather than remaining confined to the specialist's parameters.
- **Distributional drift governs the specialization–generalization trade-off.** Through controlled distillation experiments, trajectory-quality analysis, and comparisons between unconstrained tuning, implicit (LST) and explicit (ASFT) drift controls, we identify distributional drift as a governing axis of latent supervision. Varying this drift steers both specialists and their students between domain precision and general-capability retention.

2 Related Works

Specialist Distillation and Domain Adaptation. Prior work has shown that domain expertise can be transferred through specialist distillation and domain adaptation (Yao et al., 2021; Saad-Falcon et al., 2023; Liu et al., 2024; Cai et al., 2025), while recent LLM pipelines similarly construct specialized experts and distill their capabilities back into more general models (Li et al., 2024; DeepSeek-AI, 2025; Qwen-Team, 2026). Parameter-efficient adaptation can itself induce different specialization–retention trade-offs; for example, LoRA often preserves general capabilities better than full fine-tuning despite weaker target-domain adaptation (Biderman et al., 2024). In contrast, we study an upstream question specific to QA-only specialist distillation: how the specialist optimization shapes the reasoning trajectories later exposed to the student, and whether the resulting specialization–generalization profile is subsequently inherited through distillation.

Self-Distillation and Self-Generated Reasoning. Prior work has established that useful supervision can emerge without human-annotated reasoning trajectories through self-generated rationales and teacher-generated chains of thought (Zelikman et al., 2022; Huang et al., 2023; Magister et al., 2023; Li et al., 2023; Wang et al., 2023; Hsieh et al., 2023). Recent studies further show that the choice and construction of reasoning trajectories materially affect distillation: trajectory suitability predicts downstream student performance (Yang et al., 2026), while answer-conditioned generation can alter rationale properties and degrade distilled reasoning (Lee et al., 2026). Self-distillation likewise exploits a model's own outputs or distributions across reasoning, continual learning, multilingual transfer, and code generation (Yang et al., 2024; Shenfeld et al., 2026; Zhao et al., 2026; Zhang et al., 2024; 2026). Thus, our contribution is not that useful distillation is possible without gold rationales. Rather, we investigate how QA-only specialist optimization influences which latent reasoning supervision is generated in the first place, and how these optimization-induced differences propagate to downstream students.

3 QA-only Specialist Distillation as Trajectory Selection

Specialist distillation typically refers to a two-phase training pipeline. Starting from an origin model π_{θ_0}, one first obtains an intermediate model π_{θ_1} that is adapted to a target domain. This intermediate model is then used to generate training data for the final model $\pi_{\theta_0'}$, often initialized from the same origin model π_{θ_0}. In this section, we study specialist distillation from a trajectory learning perspective (in § 3.1). Surprisingly, we find that properly controlling specialist optimization can induce high-quality reasoning trajectories under QA-only supervision (in § 3.2).

3.1 From QA-only supervision to trajectory-distribution learning

QA-only supervision underdetermines the induced trajectory distribution. In many domain-specific tasks, datasets $\mathcal{D}$ contain only question–answer pairs (q, a) without reasoning trajectories τ. In our implementation, the specialist is trained with standard teacher-forced SFT on question–answer pairs (q, a); it does not explicitly optimize or marginalize over latent reasoning trajectories. We instead use a trajectory-distribution abstraction to characterize the behavior induced by such answer-level supervision. Let $\pi_{\theta_1}(\tau \mid q)$ denote the distribution over reasoning trajectories generated by the resulting specialist. At this abstraction level, the probability assigned to an answer can be viewed as aggregating over trajectories compatible with that answer:

$$\log \pi_{\theta_1}(a \mid q) = \log \sum_{\tau \to a} \pi_{\theta_1}(\tau \mid q). \tag{2}$$

where $\tau \to a$ denotes trajectories consistent with the correct answer a (or its equivalents). Consequently, multiple trajectory distributions may be equally consistent with the same QA supervision by inducing the same answer likelihood.

Optimization implicitly selects a trajectory distribution among many valid ones. Although multiple trajectory distributions are valid with the same QA supervision, optimization does not treat them equally. To understand how optimization resolves this ambiguity, we examine the gradient of the objective:

$$\nabla \log \pi_{\theta_1}(a \mid q) = \sum_{\tau \to a} p_\theta(\tau \mid q, a) \nabla \log \pi_{\theta_1}(\tau \mid q), \quad p_{\theta_1}(\tau \mid q, a) = \frac{\pi_{\theta_1}(\tau \mid q)}{\pi_{\theta_1}(a \mid q)}. \tag{3}$$

Eq. 3 shows that trajectories with larger posterior weight $p_{\theta_1}(\tau \mid q, a)$ contribute more strongly to the gradient update. Since $p_{\theta_1}(\tau \mid q, a) \propto \pi_{\theta_1}(\tau \mid q)$ among answer-consistent trajectories, high-probability trajectories dominate the gradient update. Repeated optimization therefore resolves the underdetermination by selecting a particular trajectory distribution $\pi_{\theta_1}(\tau \mid q)$ from many valid ones.

Distillation probes the latent trajectory distribution induced by specialist optimization. The latent trajectory distribution induced by specialist optimization is difficult to characterize directly: it spans a vast space of variable-length reasoning sequences, and individual samples reveal only partial information about its structure and value as supervision. Distillation provides an operational probe of this distribution by examining what a student learns from its sampled trajectories. Crucially, the student receives these trajectories without inheriting the specialist's adapted parameters or optimization constraints. Under controlled student training, downstream differences therefore provide evidence of how specialist optimization shapes transferable supervision. Formally, when trajectories sampled from π_{θ_1} are used to train $\pi_{\theta_0'}$, the resulting distillation objective is:

$$\mathcal{L}(\theta_0') = \mathbb{E}_{\tau \sim \pi_{\theta_1}(\cdot \mid q)} \big[-\log \pi_{\theta_0'}(\tau \mid q) \big]. \tag{4}$$

This quantity corresponds to the cross-entropy between π_{θ_1} and $\pi_{\theta_0'}$ and admits the decomposition

$$\begin{aligned} \mathbb{E}_{\tau \sim \pi_{\theta_1}(\cdot \mid q)} \big[-\log \pi_{\theta_0'}(\tau \mid q) \big] &= \mathbb{E}_{\tau \sim \pi_{\theta_1}(\cdot \mid q)} \left[\log \frac{\pi_{\theta_1}(\tau \mid q)}{\pi_{\theta_0'}(\tau \mid q)} \right] - \mathbb{E}_{\tau \sim \pi_{\theta_1}(\cdot \mid q)} \big[\log \pi_{\theta_1}(\tau \mid q) \big] \\ &= \mathrm{KL}\big(\pi_{\theta_1}(\cdot \mid q) \,||\, \pi_{\theta_0'}(\cdot \mid q) \big) + \mathcal{H}\big(\pi_{\theta_1}(\cdot \mid q) \big). \end{aligned} \tag{5}$$

where $\mathcal{H}(\pi_{\theta_1}(\cdot \mid q)) = -\mathbb{E}_{\tau \sim \pi_{\theta_1}(\cdot \mid q)} [\log \pi_{\theta_1}(\tau \mid q)]$ denotes the entropy of $\pi_{\theta_1}(\cdot \mid q)$ and is independent of $\pi_{\theta_0'}$. Therefore, optimizing w.r.t. $\pi_{\theta_0'}$ is equivalent to minimizing $\mathrm{KL}\big(\pi_{\theta_1}(\cdot \mid q) \,||\, \pi_{\theta_0'}(\cdot \mid q) \big)$, which drives the distilled model $\pi_{\theta_0'}$ to approximate π_{θ_1} in the trajectory space. Consequently, once the trajectory distribution π_{θ_1} is fixed, the behavior of the $\pi_{\theta_0'}$ is largely determined by π_{θ_1}. Distillation thus probes the downstream consequences of trajectory selection without requiring an explicit characterization of the full trajectory distribution.

3.2 Specialist Training as a Key Design Variable for Distillation Data

The preceding analysis connects specialist training to downstream data design: different training strategies can induce different trajectory distributions under identical QA supervision, thereby changing the supervision available for distillation. The specialist's training procedure is therefore a key design variable for shaping what the student learns. This raises a practical question: how can specialist adaptation be controlled to shape the resulting distillation data?

One approach is to regulate distributional drift from the origin model. Such control allows domain-specific QA supervision to reshape the trajectory distribution while constraining its departure from the origin model's behavior. A canonical formulation of this principle is a KL-constrained objective:

$$\max_{\pi_{\theta_1}} \quad \mathbb{E}_{(q,a) \sim \mathcal{D}} [\log \pi_{\theta_1}(a \mid q)] \quad \text{s.t.} \quad \mathrm{KL}\big(\pi_{\theta_1}(\cdot \mid q) \,||\, \pi_{\theta_0}(\cdot \mid q) \big) \le \delta, \tag{6}$$

where a small $\delta > 0$ limits how far π_{θ_1} can drift from the original model π_{θ_0}. Under this constraint, the induced trajectory distribution can be viewed as a reweighted version of the π_{θ_0}:

$$\pi_{\theta_1}(\tau \mid q) \propto \pi_{\theta_0}(\tau \mid q) \exp\left(\frac{w(\tau, q, a)}{\lambda} \right), \tag{7}$$

Table 1: Summary of training setups for our QA-only Specialist–Distillation pipeline, additional explicit KL-drift control analysis with ASFT, and two *self-training* baselines.

Model Setting	Tuning Data	Implementation Brief
FFT Specialist π_{θ_1}	(q, a^*)	FFT, LoRA, and LST share the same hyperparameters, except for those specific to LoRA and LST. See App. C.1 for details.
LoRA Specialist π_{θ_1}		Rank $r = 64$, $\alpha = 2r$, `lora_dropout` $= 0.05$, `target:all`. We provide ablation study on LoRA configs in App. B.6.
LST Specialist π_{θ_1}		Only update selected layers. We adopt Gao et al. (2025)'s best-performing config for π_{θ_0}: update bottom 4 and top 16 layers.
ASFT Specialist π_{θ_1} explicit KL-drift control		Explicitly regularizes π_{θ_1} toward π_{θ_0} with a KL penalty; $\lambda \in \{0.05, 0.2, 0.5\}$ controls anchoring strength. See § 5.2 for details.
FFT Distilled $\pi_{\theta'_0}$	$(q, \hat{\tau}, \hat{a})$	All $\pi_{\theta'_0}$'s are FFT-trained identically from π_{θ_0} Qwen3-8B, varying only in the rationale data generated by their corresponding π_{θ_1}.
Self-Distill $\pi_{\theta'_0}$ $q \to \hat{\tau}, \hat{a}$	$(q, \hat{\tau}, \hat{a})$	π_{θ_0} is prompted with only the question q to generate both the reasoning path $\hat{\tau}$ and the prediction answer $\hat{a}$.
Self-Rationalize $\pi_{\theta'_0}$ $q, a^* \to \hat{\tau}$	$(q, \hat{\tau}, a^*)$	π_{θ_0} is prompted with both the question q and the ground-truth answer a^*, and is tasked to generate the rationale $\hat{\tau}$ that leads to a^*.

where $w(\tau, q, a)$ is an implicit quantity reflecting how a trajectory contributes to increasing $\log \pi(a \mid q)$, and $\lambda > 0$ is a Lagrange multiplier controlling the strength of the constraint. In our setting, the KL–constrained formulation serves only as a characterization of the induced trajectory reweighting behavior. We use standard QA-only supervised fine-tuning without explicit KL regularization, so neither $w(\tau, q, a)$ nor λ is explicitly specified during training. The derivation of Eq. 7 is provided in App. A. Substituting Eq. 7 into the distillation objective Eq. 4 yields:

$$\mathcal{L}(\theta'_0) \propto \mathbb{E}_{\tau \sim \pi_{\theta_0}} \Bigg[\underbrace{\exp\left(\frac{w(\tau, q, a)}{\lambda}\right)}_{\text{quality-aware reweighting}} \cdot (-\log \pi_{\theta'_0}(\tau \mid q)) \Bigg], \tag{8}$$

This form shows that the fine-tuned trajectory distribution $\pi_{\theta_1}(\tau \mid q)$ is obtained by reweighting the base distribution $\pi_{\theta_0}(\tau \mid q)$. Under this characterization, trajectories that contribute more strongly to the answer-level objective receive greater relative weight while the overall distribution remains anchored to the origin model.

4 Experiments

4.1 Experimental Setup

We summarize in Tab. 1 the overarching experimental setup, including required data formats, training objectives, and key implementation details. Comprehensive training, inference, rationale filtration, and evaluation protocols are deferred to Apps. C and D. The cores are described below.

Specialist Fine-Tuning Strategies. We deploy Qwen3-8B as our origin model π_{θ_0}. To investigate how different fine-tuning strategies affect the quality of generated rationales, we train the specialist models π_{θ_1} starting from π_{θ_0} using three methods as in Tab. 1: (1) Full Fine-Tuning (FFT), (2) LoRA (Hu et al., 2022), and (3) Layer-Selective Tuning (LST, Gao et al., 2025). We additionally evaluate Anchored Supervised Fine-Tuning (ASFT; Zhu et al., 2026) as an explicit KL-based drift-control mechanism in § 5.2.

Distillation Pipeline. Once trained, each specialist π_{θ_1} variant generates candidate chain-of-thoughts (CoT; Wei et al., 2022) and answers $(\hat{\tau}, \hat{a})$, which are *filtered* such that $\hat{a}$ is equivalent to ground-truth answer a^* and *complete* CoT to construct valid $(q, \hat{\tau}, \hat{a})$ pool (see § 5.4 and App. D for details). Crucially, to isolate the impact of data quality from the *student* model's learning capacity, all *distilled* models $\pi_{\theta'_0}$ are FFT-trained on an equal amount of subsampled data and identical configurations, regardless of their corresponding *specialist* model π_{θ_1} tuning strategy.

Table 2: **Performance comparison across Chemistry, Physics, and Multilingualism.** The parenthesized, e.g., ID (4), are the number of benchmarks (subsets) we use for this category. Chemistry (SMol) IT score aggregate its 14 subtasks, with its computation detailed in App. D.1.1. All other metrics are macro–averages over the corresponding categories. The same five OOD benchmarks (5) are used across all three *target* domain settings. Bold denotes the best per column among models derived from Qwen3-8B ; -14/32B are excluded.

	Model	Chemistry			Physics			Multilingualism		
		IT (14)	ID (4)	OOD (5)	IT (1)	ID (4)	OOD (5)	IT (16)	ID (16)	OOD (5)
	Qwen3-8B π_{θ_0}	19.34	60.19	42.22	90.21	68.61	42.22	29.48	44.99	**42.22**
	Qwen3-14B	23.80	63.67	48.95	91.33	73.06	48.95	32.31	48.43	48.95
	Qwen3-32B	25.35	68.09	51.23	92.45	77.86	51.23	33.78	49.13	51.23
Tuned on π_{θ_0} Qwen3-8B	$q \to \hat{\tau}, \hat{a}$ Self-Distill $\pi_{\theta'_0}$	21.06	59.28	42.04	91.11	67.97	**43.54**	29.30	**45.94**	39.66
	$q, a^* \to \hat{\tau}$ Self-Rationalize $\pi_{\theta'_0}$	17.92	58.78	33.14	87.89	68.64	41.78	6.02	27.26	27.71
	FFT *Specialist* π_{θ_1}	**59.65**	57.72	33.60	90.73	68.43	16.31	**35.93**	23.86	6.90
	↳ FFT *Distilled* $\pi_{\theta'_0}$	40.55	46.14	25.20	91.85	68.51	14.09	30.77	27.78	15.08
	LoRA *Specialist* π_{θ_1}	28.57	60.36	42.65	92.83	68.98	38.03	20.47	43.16	41.41
	↳ FFT *Distilled* $\pi_{\theta'_0}$	29.80	59.31	34.89	91.93	69.87	38.26	22.97	30.87	35.89
	LST *Specialist* π_{θ_1}	29.61	60.20	**44.03**	**93.65**	70.13	38.50	35.68	43.99	39.08
	↳ FFT *Distilled* $\pi_{\theta'_0}$	29.54	**60.47**	42.11	92.75	**70.84**	40.80	34.17	43.87	41.23

Baselines. We introduce two *self-training* baselines (Tabs. 1 and 2) derived directly from π_{θ_0}: (1) **Self-Distill**: unconditional rationale generation and (2) **Self-Rationalize**: answer-conditioned rationale generation. We further report larger models Qwen3-{14,32}B for scale-based comparisons.

Training Data and Evaluation Suites. We organize our datasets and benchmarks into 4 categories (see App. D for all benchmarks we evaluate): (1) **Training** data utilized for fine-tuning within each target domain; (2) **In-Task (IT)** benchmarks that share the same domain and task formulation as the training data, using the official test split when available, otherwise a distribution-wise closely matched benchmark; (3) **In-Domain (ID)** benchmarks that remain within the target domain but differ in task distribution and difficulty level, evaluating robustness under domain shift (Farahani et al., 2020); and (4) **Out-of-Domain (OOD)** benchmarks drawn from domains entirely different from the target domain, assessing broader generalization. We use 1 benchmark for complex reasoning (Kazemi et al., 2025), 2 for mathematics (Zhang, 2026), and 2 for coding (Jain et al., 2025).

Domains and Rationale Filtrations. We study three *target* domains for Training, IT, and ID: (1) **Chemistry** (CHEM). We train on SMol (Yu et al., 2024), covering molecular understanding and generation across 14 subtasks, whose official test split serves as the IT. For ID, we evaluated on general chemistry benchmarks. During rationale filtration, we apply subtask-specific criteria (App. D.1.1) to accommodate its diverse output formats and evaluation protocols. (2) **Physics** (PHYS). We train on the university-level physics subset of MegaScience (Fan et al., 2025). For IT, we evaluate on university-level subsets from PHYSICS benchmark (Feng et al., 2025); for ID, on high-school physics benchmarks. Rationales are retained only for answers that pass rule-based symbolic and numerical verification. (3) **Low-Resource Multilingualism** (LRM). We train on OPUS (Tiedemann, 2012) for *bi*-directional English – 8 low-resource languages translation. For IT, we evaluate the same 16 translation task using Flores-101 (Goyal et al., 2022); for ID, we assess general reasoning in these 8 languages, beyond translation. Rationales are ranked by sentence-level `spBLEU` (Papineni et al., 2002; Goyal et al., 2022) and the top-performing 20% subset is retained.

4.2 MAIN RESULTS

Tab. 2 compares the origin model, larger same-family models, untuned self-training baselines, and our specialist–distillation pipeline across Chemistry, Physics, and Multilingualism under IT, ID, and OOD evaluation. Benchmark composition and metric computation details are provided in App. D.

QA-only specialist distillation consistently improves target-domain performance. Across FFT, LoRA, and LST tuning strategies and three target domains, our key observations from Tab. 2 demonstrate that even without gold rationales, explicitly training a QA-only *specialist* to induce

domain-specific reasoning traces, and subsequently transferring them to a *distilled* model, consistently yields robust and significant target-domain growth:

- **Substantial target-domain improvements.** Across majority of configurations, both the *specialist* models π_{θ_1} and their downstream *distilled* models $\pi_{\theta'_0}$ achieve massive *in-task* (IT) capability gains compared to the origin Qwen3-8B π_{θ_0} and two *self-training* baselines.
- **Bridging a $4\times$ parameter gap.** Our pipeline enables an 8B model to "punch above its weight class" without relying on human-annotated rationales. For example, our LST-distilled model ($29.54/92.75/34.17$ for all IT) surpass the zero-shot performance of the $4\times$ larger Qwen3-32B ($25.35/92.45/33.78$). This highlights that extracting latent reasoning paths from a specialist is a highly parameter-efficient paradigm compared to merely scaling up generalist models.
- **Toxic post-hoc rationalization.** The untrained *Self-Rationalize* baseline severely degrades performance, notably plummeting LRM IT from 29.48 to 6.02, which corroborates findings that post-hoc reasoning on translation is often spurious and unreliable (Li et al., 2026). Forcing weak models to rationalize answers induces hallucination or shortcuts that poison distillation.

Different QA-only supervision methods implicitly select different trajectory distributions and lead to distinct specialization-generalization trade-offs. Overall, across all 3 domains, LST consistently improves IT performance over the origin, achieves modest gains on ID, and remains on par on OOD. While FFT yields impressive IT gains in specific domains such as chemistry, it severely sacrifices both the performance of ID and OOD, suffering from *catastrophic forgetting* (Kumar et al., 2022). LoRA exhibits a trade-off pattern similar to LST but generally falls short of LST across all 9 splits. These distinct performance profiles confirm that different QA-only tuning strategies implicitly select different rationale distributions. Thus, the optimal tuning method depends on the desired specialization–generalization trade–off rather than a universal optimum.

Distillation reliably inherits and refines specialist capabilities. Beyond absolute metrics, an intriguing phenomenon observed in Tab. 2 is the strong *rank* correlation between the capabilities of π_{θ_1} and its downstream $\pi_{\theta'_0}$: across all 27 experimental measurements, Spearman's rank correlation reaches $\rho = 0.9573$ (Fig. 2a). To account for the dependence among IT, ID, and OOD measurements from the same specialist–student pair, we additionally conduct a domain-blocked, model-profile-level exact permutation test, which yields $p-\text{value} = 2/215 \approx 0.0093$. The relative performance ranking among tuning strategies (e.g., LST > LoRA > FFT in PHYS IT) is therefore strongly reflected in the downstream *distilled* models across the evaluated splits. This consistency supports the effectiveness of our variable-controlled pipeline: because all $\pi_{\theta'_0}$ models are fine-tuned using the exact same FFT configuration, differences in their performance are closely associated with the rationale supervision $\hat{\tau}$ generated by their corresponding π_{θ_1}. Consequently, stronger *specialists* generally tend to produce supervision that leads to stronger *distilled* models. Furthermore, in nearly half ($12/27$) of the splits, the *distilled* model even exceeds its corresponding *specialist*, indicating that distillation does not merely copy teacher behavior but can further refine the supervision induced by specialist-generated rationales. More broadly, distilled models consistently outperform the two untuned self-training baselines, suggesting that effective rationale supervision is better obtained from explicitly trained *specialist* models than from the origin model alone or post-hoc self-rationalization. Together, these results support specialist training as an effective approach for producing rationale data useful for downstream distillation within the evaluated Qwen3-8B-based pipeline.

5 ANALYSIS

Beyond the analyses presented below, we provide supplementary studies on *data scaling* (App. B.1), *cross-model-family transfer* App. B.2, *cross-scale distillation* (App. B.3), and LoRA *ablations* (App. B.6). We further verify that the specialization–generalization trade-off and rank preservation persist when scaling π_{θ_0} to Qwen3-14B (App. B.5).

5.1 POST-TRAINING STRATEGIES INDUCE DIFFERENT DEGREES OF KL DRIFT.

To characterize how specialist post-training changes model behavior with respect to the origin model, we measure the token-level $\mathrm{KL}(\pi_\bullet \,||\, \pi_{\theta_0}), \forall \pi_\bullet \in \{\pi_{\theta_1}, \pi_{\theta'_0}\}$, across different tuning strate-

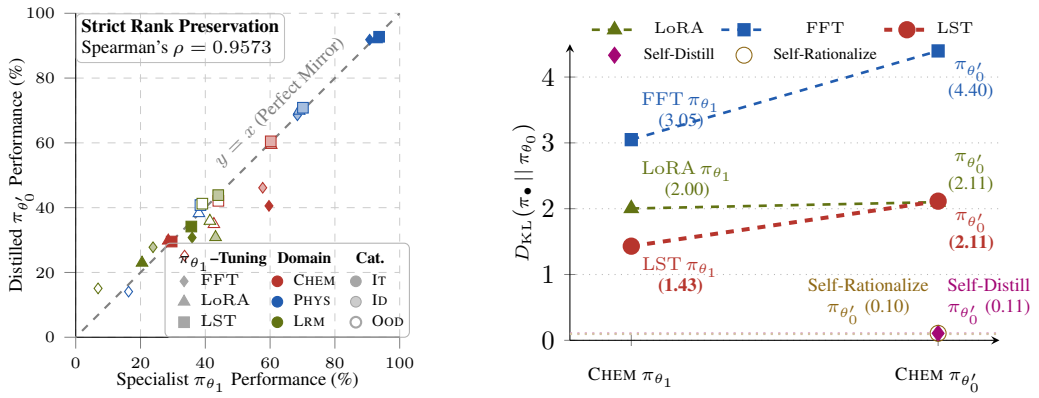


Figure 2: **Specialist–student (a) inheritance and (b) distributional drift. Left (a):** Rank preservation from π_{θ_1} to $\pi_{\theta_0'}$ across all evaluated specialist–student pairs. Each point is one $(\pi_{\theta_1}, \pi_{\theta_0'})$ performance pair as in Tab. 2. **Right (b):** KL divergence of Chem π_{θ_1} and $\pi_{\theta_0'}$ models with respect to the origin model π_{θ_0} Qwen3-8B. LST models exhibit lower KL divergence than their FFT and LoRA counterparts.

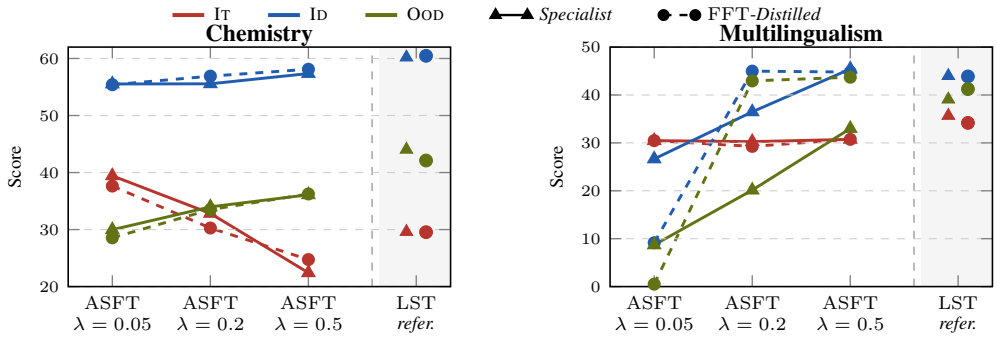


Figure 3: **Explicit KL anchoring with ASFT.** Increasing the KL coefficient λ strengthens anchoring to the origin model, shifting both specialists and their distilled students from stronger task specialization toward better Id/Ood retention. The first three x-axis settings correspond to ASFT with $\lambda \in \{0.05, 0.2, 0.5\}$. The shaded LST *refer.* column shows the corresponding LST specialist and distilled student as an implicit drift-control reference (Tab. 2) rather than an ASFT setting. **Left (a)**: Chemistry. **Right (b)**: Multilingualism.

gies on Chem (Fig. 2b). We use KL divergence as a behavioral diagnostic of distributional drift rather than as the explicit tuning objective. FFT exhibits the largest divergence from the origin model for both the specialist π_{θ_1} and its distilled student $\pi_{\theta_0'}$. LoRA produces substantially smaller drift, while LST yields the lowest KL divergence among the adapted models evaluated. These results show an empirical association between stronger constraints on specialist adaptation and reduced behavioral drift from the origin model. In particular, LST is consistent with an *implicit anchoring effect*: although it contains no explicit KL regularization, its restricted update space yields a specialist that remains closer to the origin-model distribution while acquiring target-domain capability. Importantly, we do not claim that LST optimizes a KL-constrained objective, nor that a particular layer-selection configuration corresponds to a specific KL radius. The lower-drift profile induced by LST is also reflected in its distilled student, motivating the question of whether explicitly controlling KL drift can systematically move the specialization–generalization operating point (§ 5.2).

5.2 Explicit KL Drift Control with ASFT

Our analysis suggests that the specialization–generalization trade-off is governed more broadly by how strongly specialist optimization controls distributional drift, rather than by LST specifically. To test this interpretation beyond structural constraints, we additionally evaluate Anchored Supervised Fine-Tuning (ASFT; Zhu et al., 2026), which explicitly regularizes the specialist toward the origin model through a KL penalty. We repeat the specialist generation and FFT-distillation pipeline with the coefficient $\lambda \in \{0.05, 0.2, 0.5\}$, where larger λ imposes stronger anchoring. Stronger anchoring systematically shifts both specialists and their distilled students from task specialization toward im-

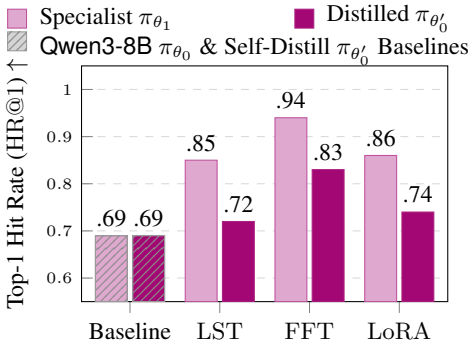


Figure 4: HR@1 under *teacher-forcing* on unseen CHEM SMol queries. Higher (↑) is better. Hatched are the origin model π_{θ_0} and *self-distill* baselines.

Table 3: **Rationale quality on Chemistry**. Statistics over 150K generated trajectories from 50K SMol queries with 3 samples each, covering two *self-training* modes of π_{θ_0} and three specialists (FFT, LoRA, and LST). "Comp." is the percentage of structurally *complete* trajectories. "Empty" is the permille of empty rationales among *complete* trajectories. "Words" denotes average word count ± std within *complete* rationales.

Model Generation	Comp.(%)	Empty(‰)	Words($\mu \pm \sigma$)
π_{θ_0} $q \to \hat{\tau}, \hat{a}$	87.93	0.00	2476 ± 1624
π_{θ_0} $q, a^* \to \hat{\tau}$	16.18	0.08	2049 ± 1443
FFT Spec. π_{θ_1}	8.90	1.10	233 ± 450
LoRA Spec. π_{θ_1}	41.39	0.00	1378 ± 902
LST Spec. π_{θ_1}	92.50	0.02	2334 ± 1435

proved ID/OOD retention (Fig. 3), confirming that the trade-off is not unique to LST: explicit KL regularization provides a tunable mechanism for controlling specialist drift, whose resulting capability profile is inherited through distillation. However, ASFT introduces additional computational and implementation overhead from explicit KL regularization, whereas LST provides implicit drift control through its restricted update space alone.

5.3 STRUCTURAL CONSTRAINTS IMPROVE PREDICTION WITHOUT TRAJECTORY COLLAPSE.

To analyze how different fine-tuning strategies distribute probability mass over valid trajectories (§ 3.1), we evaluate next-token top-1 hit rate (HR@1) via *teacher-forcing*, where each next token is predicted conditioned on the gold prefix, on $2,000$ unseen SMol samples (App. D.1.1). Suppressing intermediate CoT steps isolates the ability to sequentially predict ground-truth answers. While all tuning methods improve HR@1 over the baselines (Fig. 4), unconstrained FFT induces extreme specialist confidence (HR@1 $= 0.94$), suggesting trajectory-distribution collapse toward answer likelihood (Eq. 2). In contrast, parameter-efficient LST and LoRA yield more calibrated improvements (HR@1 ≈ 0.85). Consistent with our framework (§ 3.2), these structural constraints induce smoother reweighting of the base policy (Eq. 7) rather than severely distorting the generation manifold, facilitating more robust transfer to distilled students. Next-token probability and rank show consistent trends (App. B.4).

5.4 STRUCTURAL CONSTRAINTS PRESERVE REASONING-TRAJECTORY INTEGRITY.

We analyze 150K Chemistry trajectories (3 samples for each of 50K SMol queries), defining a trajectory as complete when it contains exactly one valid `<think>...</think>` block. As Tab. 3 shows, FFT collapses to only 8.90% complete trajectories with much shorter traces (233 vs. 2476 words for the origin), whereas LST preserves 92.50% completeness and origin-like trace lengths; LoRA lies between them. Furthermore, the *self-rationalize* baseline drops to 16.18% completion, largely due to the generation of multiple reasoning blocks. This reinforces our claim in § 4.2 that forcing weak models into post-hoc rationalization fundamentally destabilizes the generation manifold and acts as toxic supervision.

6 CONCLUSION

We frame QA-only specialist distillation as a trajectory-distribution selection problem: answer supervision leaves reasoning trajectories underdetermined, while specialist optimization determines the latent supervision passed downstream. Across chemistry, physics, and multilingual tasks, distilled students closely inherit their specialists' specialization–generalization profiles, even across model families. We further identify distributional drift as a controllable axis of this transfer: unconstrained FFT drives stronger specialization at the cost of generalization, whereas LST better preserves general capabilities while retaining domain gains. Together, these results establish specialist optimization as a key design variable for distillation when gold trajectories are unavailable.

AI USE STATEMENT

We used generative AI tools to polish the manuscript's writing and to provide reviewer-style feedback and suggestions for revision. These tools did not contribute to the development of the ideas, methodology, experiments, analysis, conclusions, or citations presented in this work. All AI-assisted edits were reviewed by the authors. We, the authors, shall take full responsibility for the final content of this manuscript.

ETHICS STATEMENT

This work studies language-model training and distillation using publicly available models, datasets, and evaluation benchmarks. It does not involve human subjects, crowdsourcing, or the collection of personally identifiable information. Third-party assets are used with attribution and in accordance with their applicable licenses and terms of use. While our work does not target high-stakes deployment, the studied adaptation and distillation techniques may inherit or amplify errors and biases present in the underlying models or training data; consequently, our results should not be interpreted as establishing reliability or safety for such applications.

REPRODUCIBILITY STATEMENT

We provide detailed information necessary to reproduce our experiments throughout the paper. The overall experimental setup and controlled specialist–distillation pipeline are described in § 4.1, while comprehensive training and inference configurations, including hyperparameters and computational resources, are provided in App. C. Dataset construction, trajectory generation and filtering, benchmark composition, and evaluation protocols are documented in App. D.

Our code is available at an anonymous repository `https://anonymous.4open.science/r/specialist-distillation-iclr-D629/`. Trained model artifacts and all inference data will be released publicly upon acceptance.

REFERENCES

Lei Bai, Zhongrui Cai, Yuhang Cao, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, Yongkang Chen, Yu Cheng, Pei Chu, Tao Chu, Erfei Cui, Ganqu Cui, Long Cui, Ziyun Cui, Nianchen Deng, Ning Ding, Nanqing Dong, Peijie Dong, Shihan Dou, Sinan Du, Haodong Duan, Caihua Fan, Ben Gao, Changjiang Gao, Jianfei Gao, Songyang Gao, Yang Gao, Zhangwei Gao, Jiaye Ge, Qiming Ge, Lixin Gu, Yuzhe Gu, Aijia Guo, Qipeng Guo, Xu Guo, Conghui He, Junjun He, Yili Hong, Siyuan Hou, Caiyu Hu, Hanglei Hu, Jucheng Hu, Ming Hu, Zhouqi Hua, Haian Huang, Junhao Huang, Xu Huang, Zixian Huang, Zhe Jiang, Lingkai Kong, Linyang Li, Peiji Li, Pengze Li, Shuaibin Li, Tianbin Li, Wei Li, Yuqiang Li, Dahua Lin, Junyao Lin, Tianyi Lin, Zhishan Lin, Hongwei Liu, Jiangning Liu, Jiyao Liu, Junnan Liu, Kai Liu, Kaiwen Liu, Kuikun Liu, Shichun Liu, Shudong Liu, Wei Liu, Xinyao Liu, Yuhong Liu, Zhan Liu, Yinquan Lu, Haijun Lv, Hongxia Lv, Huijie Lv, Qitan Lv, Ying Lv, Chengqi Lyu, Chenglong Ma, Jianpeng Ma, Ren Ma, Runmin Ma, Runyuan Ma, Xinzhu Ma, Yichuan Ma, Zihan Ma, Sixuan Mi, Junzhi Ning, Wenchang Ning, Xinle Pang, Jiahui Peng, Runyu Peng, Yu Qiao, Jiantao Qiu, Xiaoye Qu, Yuan Qu, Yuchen Ren, Fukai Shang, Wenqi Shao, Junhao Shen, Shuaike Shen, Chunfeng Song, Demin Song, Diping Song, Chenlin Su, Weijie Su, Weigao Sun, Yu Sun, Qian Tan, Cheng Tang, Huanze Tang, Kexian Tang, Shixiang Tang, Jian Tong, Aoran Wang, Bin Wang, Dong Wang, Lintao Wang, Rui Wang, Weiyun Wang, Wenhai Wang, Jiaqi Wang, Yi Wang, Ziyi Wang, Ling-I Wu, Wen Wu, Yue Wu, Zijian Wu, Linchen Xiao, Shuhao Xing, Chao Xu, Huihui Xu, Jun Xu, Ruiliang Xu, Wanghan Xu, GanLin Yang, Yuming Yang, Haochen Ye, Jin Ye, Shenglong Ye, Jia Yu, Jiashuo Yu, Jing Yu, Fei Yuan, Yuhang Zang, Bo Zhang, Chao Zhang, Chen Zhang, Hongjie Zhang, Jin Zhang, Qiaosheng Zhang, Qiuyinzhe Zhang, Songyang Zhang, Taolin Zhang, Wenlong Zhang, Wenwei Zhang, Yechen Zhang, Ziyang Zhang, Haiteng Zhao, Qian Zhao, Xiangyu Zhao, Xiangyu Zhao, Bowen Zhou, Dongzhan Zhou, Peiheng Zhou, Yuhao Zhou, Yunhua Zhou, Dongsheng Zhu, Lin Zhu, and Yicheng Zou. Intern-S1: A scientific multimodal foundation model. *CoRR*, abs/2508.15763, 2025. doi: 10.48550/ARXIV.2508.15763. URL `https://doi.org/10.48550/arXiv.2508.15763`.

Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare R. Voss (eds.), *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pp. 65–72. Association for Computational Linguistics, 2005. URL https://aclanthology.org/W05-0909/.

Dan Biderman, Jacob P. Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John Patrick Cunningham. Lora learns less and forgets less. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=aloEru2qCG.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 7432–7439. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6239. URL https://doi.org/10.1609/aaai.v34i05.6239.

Xunxin Cai, Chengrui Wang, Qingqing Long, Yuanchun Zhou, and Meng Xiao. Knowledge hierarchy guided biological-medical dataset distillation for domain LLM training. *CoRR*, abs/2501.15108, 2025. doi: 10.48550/ARXIV.2501.15108. URL https://doi.org/10.48550/arXiv.2501.15108.

Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan, Shaoliang Nie, Xiaochang Peng, Xiang Ren, and Hamed Firooz. UNIREX: A unified learning framework for language model rationale extraction. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Proceedings of Machine Learning Research, pp. 2867–2889. PMLR, 2022. URL https://proceedings.mlr.press/v162/chan22a.html.

OpenCompass Contributors. OpenCompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

Antonia Creswell and Murray Shanahan. Faithful reasoning using large language models. *CoRR*, abs/2208.14271, 2022. doi: 10.48550/ARXIV.2208.14271. URL https://doi.org/10.48550/arXiv.2208.14271.

DeepSeek-AI. DeepSeek-V3.2: Pushing the frontier of open large language models. *CoRR*, abs/2512.02556, 2025. doi: 10.48550/ARXIV.2512.02556. URL https://doi.org/10.48550/arXiv.2512.02556.

Run-Ze Fan, Zengzhi Wang, and Pengfei Liu. Megascience: Pushing the frontiers of post-training datasets for science reasoning. *CoRR*, abs/2507.16812, 2025. doi: 10.48550/ARXIV.2507.16812. URL https://doi.org/10.48550/arXiv.2507.16812.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. A brief review of domain adaptation. *CoRR*, abs/2010.03978, 2020. URL https://arxiv.org/abs/2010.03978.

Henri A Favre and Warren H Powell. *Nomenclature of Organic Chemistry: IUPAC Recommendations and Preferred Names 2013 (Blue Book)*. Royal Society of Chemistry, 2013.

Kaiyue Feng, Yilun Zhao, Yixin Liu, Tianyu Yang, Chen Zhao, John Sous, and Arman Cohan. Physics: Benchmarking foundation models on university-level physics problem solving. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pp. 11717–11743. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-acl.610/.

The International Organization for Standardization. ISO 639:2023(en)code for individual languages and language groups, 2023. URL https://www.iso.org/obp/ui/en/#iso:std:iso:639:ed-2:v1:en.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. Specializing smaller language models towards multi-step reasoning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pp. 10421–10430. PMLR, 2023. URL https://proceedings.mlr.press/v202/fu23d.html.

Changjiang Gao, Zixian Huang, Jingyang Gong, Shujian Huang, Lei Li, and Fei Yuan. LLaMAX2: Your translation-enhanced model also performs well in reasoning. *CoRR*, abs/2510.09189, 2025. doi: 10.48550/ARXIV.2510.09189. URL https://doi.org/10.48550/arXiv.2510.09189.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11):665–673, 2020. doi: 10.1038/S42256-020-00257-Z. URL https://doi.org/10.1038/s42256-020-00257-z.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Trans. Assoc. Comput. Linguistics*, 10:522–538, 2022. doi: 10.1162/TACL_A_00474. URL https://doi.org/10.1162/tacl_a_00474.

Çaglar Gülçehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. Reinforced self-training (rest) for language modeling. *CoRR*, abs/2308.08998, 2023. doi: 10.48550/ARXIV.2308.08998. URL https://doi.org/10.48550/arXiv.2308.08998.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL http://arxiv.org/abs/1503.02531.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 14852–14882. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.830. URL https://doi.org/10.18653/v1/2023.acl-long.830.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, volume ACL 2023 of *Findings of ACL*, pp. 8003–8017. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.507. URL https://doi.org/10.18653/v1/2023.findings-acl.507.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 1051–1068. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.67. URL https://doi.org/10.18653/v1/2023.emnlp-main.67.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. BenchMAX: A comprehensive multilingual evaluation suite for large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 16751–16774. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-emnlp.909/.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

Damjan Kalajdzievski. A rank stabilization scaling factor for fine-tuning with lora. *CoRR*, abs/2312.03732, 2023. doi: 10.48550/ARXIV.2312.03732. URL https://doi.org/10.48550/arXiv.2312.03732.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K. Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V. Le, and Orhan Firat. BIG-Bench extra hard. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 26473–26501. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.1285/.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=UYneFzXSJWh.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Jason Flinn, Margo I. Seltzer, Peter Druschel, Antoine Kaufmann, and Jonathan Mace (eds.), *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Greg Landrum, Paolo Tosco, Brian Kelley, Ricardo Rodriguez, David Cosgrove, Riccardo Vianello, sriniker, Peter Gedeck, Gareth Jones, Eisuke Kawashima, NadineSchneider, Dan Nealschneider, tadhurst cdd, Andrew Dalke, Matt Swain, Brian Cole, Samo Turk, Aleksandr Savelev, Niels Maeder, Rachel Walker, Alain Vaucher, Maciej Wójcikowski, Hussein Faara, Ichiru Take, Vincent F. Scalfani, Yakov Pechersky, Kazuya Ujihara, Daniel Probst, Jeremy Monat, and Juuso Lehtivarjo. rdkit/rdkit: 2026_03_1 (q1 2026) release, March 2026. URL https://doi.org/10.5281/zenodo.19250388.

Jungseob Lee, Seungyoon Lee, Suhyune Son, Dongyub Jude Lee, Sungbin Han, Sugyeong Eo, and Heuiseok Lim. Answer-conditioned chains of thought degrade verifiable-reasoning distillation in large language models. *CoRR*, abs/2607.14552, 2026. doi: 10.48550/ARXIV.2607.14552. URL https://doi.org/10.48550/arXiv.2607.14552.

Jiawei Li, Xiaoang Xu, and Yang Gao. METEOR: evolutionary journey of large language models from guidance to self-growth. *CoRR*, abs/2411.11933, 2024. doi: 10.48550/ARXIV.2411.11933. URL https://doi.org/10.48550/arXiv.2411.11933.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 2665–2679. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.150. URL https://doi.org/10.18653/v1/2023.acl-long.150.

Zihao Li, Shaoxiong Ji, and Jörg Tiedemann. Test-time scaling of reasoning models for machine translation. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2902–2917, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.133. URL https://aclanthology.org/2026.eacl-long.133/.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Jiaheng Liu, Chenchen Zhang, Jinyang Guo, Yuanxing Zhang, Haoran Que, Ken Deng, Zhiqi Bai, Jie Liu, Ge Zhang, Jiakai Wang, Yanan Wu, Congnan Liu, Jiamang Wang, Lin Qu, Wenbo Su, and Bo Zheng. DDK: distilling domain knowledge for efficient large language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/b206d54ffbb803b5c51d85f405d422e4-Abstract-Conference.html.

Llama-Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

M-A-P-Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, King Zhu, Minghao Liu, Yiming Liang, Xiaolong Jin, Zhenlin Wei, Chujie Zheng, Kaixin Deng, Shawn Gavin, Shian Jia, Sichao Jiang, Yiyan Liao, Rui Li, Qinrui Li, Sirun Li, Yizhi Li, Yunwen Li, David Ma, Yuansheng Ni, Haoran Que, Qiyao Wang, Zhoufutu Wen, Siwei Wu, Tyshawn Hsing, Ming Xu, Zhenzhu Yang, Zekun Moore Wang, Junting Zhou, Yuelin Bai, Xingyuan Bu, Chenglin Cai, Liang Chen, Yifan Chen, Chengtuo Cheng, Tianhao Cheng, Keyi Ding, Siming Huang, Yun Huang, Yaoru Li, Yizhe Li, Zhaoqun Li, Tianhao Liang, Chengdong Lin, Hongquan Lin, Yinghao Ma, Tianyang Pang, Zhongyuan Peng, Zifan Peng, Qige Qi, Shi Qiu, Xingwei Qu, Shanghaoran Quan, Yizhou Tan, Zili Wang, Chenqing Wang, Hao Wang, Yiya Wang, Yubo Wang, Jiajun Xu, Kexin Yang, Ruibin Yuan, Yuanhao Yue, Tianyang Zhan, Chun Zhang, Jinyang Zhang, Xiyue Zhang, Xingjian Zhang, Yue Zhang, Yongchi Zhao, Xiangyu Zheng, Chenghua Zhong, Yang Gao, Zhoujun Li, Dayiheng Liu, Qian Liu, Tianyu Liu, Shiwen Ni, Junran Peng, Yujia Qin, Wenbo Su, Guoyin Wang, Shi Wang, Jian Yang, Min Yang, Meng Cao, Xiang Yue, Zhaoxiang Zhang, Wangchunshu Zhou, Jiaheng Liu, Qunshu Lin, Wenhao Huang, and Ge Zhang. SuperGPQA: Scaling LLM evaluation across 285 graduate disciplines. *CoRR*, abs/2502.14739, 2025. doi: 10.48550/ARXIV.2502.14739. URL https://doi.org/10.48550/arXiv.2502.14739.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adámek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki

Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 1773–1781. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-SHORT.151. URL https://doi.org/10.18653/v1/2023.acl-short.151.

Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Martiño Ríos-García, Benedict Emoekabu, Aswanth Krishnan, Tanya Gupta, Mara Schilling-Wilhelmi, Macjonathan Okereke, Anagha Aneesh, Mehrdad Asgari, Juliane Eberhardt, Amir Mohammad Elahi, Hani M. Elbeheiry, María Victoria Gil, Christina Glaubitz, Maximilian Greiner, Caroline T. Holick, Tim Hoffmann, Abdelrahman Ibrahim, Lea C. Klepsch, Yannik Köster, Fabian Alexander Kreth, Jakob Meyer, Santiago Miret, Jan Matthias Peschel, Michael Ringleb, Nicole C. Roesner, Johanna Schreiber, Ulrich S. Schubert, Leanne M. Stafast, A. D. Dinga Wonanke, Michael Pieler, Philippe Schwaller, and Kevin Maik Jablonka. A framework for evaluating the chemical knowledge and reasoning abilities of large language models against the expertise of chemists. *Nature Chemistry*, 17(7):1027–1034, May 2025. ISSN 1755-4349. doi: 10.1038/s41557-025-01815-x. URL http://dx.doi.org/10.1038/s41557-025-01815-x.

H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, 1965. doi: 10.1021/c160017a018. URL https://doi.org/10.1021/c160017a018.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. Orca: Progressive learning from complex explanation traces of GPT-4. *CoRR*, abs/2306.02707, 2023. doi: 10.48550/ARXIV.2306.02707. URL https://doi.org/10.48550/arXiv.2306.02707.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pp. 311–318. ACL, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Qwen-Team. Qwen3.5-omni technical report. *CoRR*, abs/2604.15804, 2026. doi: 10.48550/ARXIV.2604.15804. URL https://doi.org/10.48550/arXiv.2604.15804.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022, 2023. doi: 10.48550/ARXIV.2311.12022. URL https://doi.org/10.48550/arXiv.2311.12022.

Jon Saad-Falcon, Omar Khattab, Keshav Santhanam, Radu Florian, Martin Franz, Salim Roukos, Avirup Sil, Md Sultan, and Christopher Potts. UDAPDR: Unsupervised domain adaptation via LLM prompting and distillation of rerankers. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11265–11279, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.693. URL https://aclanthology.org/2023.emnlp-main.693/.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. *CoRR*, abs/2601.19897, 2026. doi: 10.48550/ARXIV.2601.19897. URL https://doi.org/10.48550/arXiv.2601.19897.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron T. Parisi, Abhishek Kumar, Alexander A. Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Fathy Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. Beyond human data: Scaling self-training for problem-solving with language models. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=lNAyUngGFK.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Findings of ACL, pp. 13003–13051. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.824. URL https://doi.org/10.18653/v1/2023.findings-acl.824.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pp. 2214–2218. European Language Resources Association (ELRA), 2012. URL http://www.lrec-conf.org/proceedings/lrec2012/summaries/463.html.

Yilei Tu, Andrew Xue, and Freda Shi. Blessing of multilinguality: A systematic analysis of multilingual in-context learning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pp. 6213–6248. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-acl.323/.

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. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/ed3fea9033a80fea1376299fa7863f4a-Abstract-Conference.html.

Jonathan Uesato, Nate Kushman, Ramana Kumar, H. Francis Song, Noah Y. Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *CoRR*, abs/2211.14275, 2022. doi: 10.48550/ARXIV.2211.14275. URL https://doi.org/10.48550/arXiv.2211.14275.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. SCOTT: self-consistent chain-of-thought distillation. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 5546–5558. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.304. URL https://doi.org/10.18653/v1/2023.acl-long.304.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract-Datasets_and_Benchmarks_Track.html.

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. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005. URL https://doi.org/10.1021/ci00057a005.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pp. 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-DEMOS.6. URL https://doi.org/10.18653/v1/2020.emnlp-demos.6.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. Wizardlm: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=CfXh93NDgH.

Wenda Xu, Rujun Han, Zifeng Wang, Long T. Le, Dhruv Madeka, Lei Li, William Yang Wang, Rishabh Agarwal, Chen-Yu Lee, and Tomas Pfister. Speculative knowledge distillation: Bridging the teacher-student gap through interleaved sampling. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=EgJhwYR2tB.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

Yuming Yang, Mingyoung Lai, Wanxu Zhao, Xiaoran Fan, Zhiheng Xi, Mingqi Wu, Chiyue Huang, Jun Zhao, Haijun Lv, Jian Tong, Yunhua Zhou, Yicheng Zou, Qipeng Guo, Tao Gui, Qi Zhang, and Xuanjing Huang. Which reasoning trajectories teach students to reason better? A simple metric of informative alignment. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2026, San Diego, California, United States, July 2-7, 2026*, pp. 42123–42150. Association for Computational Linguistics, 2026. doi: 10.18653/V1/2026.ACL-LONG.1950. URL https://doi.org/10.18653/v1/2026.acl-long.1950.

Zhaorui Yang, Tianyu Pang, Haozhe Feng, Han Wang, Wei Chen, Minfeng Zhu, and Qian Liu. Self-distillation bridges distribution gap in language model fine-tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 1028–1043. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.58. URL https://doi.org/10.18653/v1/2024.acl-long.58.

Yunzhi Yao, Shaohan Huang, Wenhui Wang, Li Dong, and Furu Wei. Adapt-and-distill: Developing small, fast and effective pretrained language models for domains. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, Findings of ACL, pp. 460–470. Association for Computational Linguistics, 2021. doi: 10.18653/V1/2021.FINDINGS-ACL.40. URL https://doi.org/10.18653/v1/2021.findings-acl.40.

Botao Yu, Frazier N. Baker, Ziqi Chen, Xia Ning, and Huan Sun. LlaSMol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. In *First Conference on Language Modeling, COLM 2024, Philadelphia, PA, USA, October 7-9, 2024*, 2024. URL https://openreview.net/forum?id=lY6XTF9tPv.

Fei Yuan, Yinquan Lu, Wenhao Zhu, Lingpeng Kong, Lei Li, Yu Qiao, and Jingjing Xu. Lego-MT: Learning detachable models for massively multilingual machine translation. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, Findings of ACL, pp. 11518–11533. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.731. URL https://doi.org/10.18653/v1/2023.findings-acl.731.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/639a9a172c044fbb64175b5fad42e9a5-Abstract-Conference.html.

Ruixiang Zhang, Richard He Bai, Huangjie Zheng, Navdeep Jaitly, Ronan Collobert, and Yizhe Zhang. Embarrassingly simple self-distillation improves code generation. *CoRR*, abs/2604.01193, 2026. URL https://arxiv.org/abs/2604.01193.

Yifan Zhang. American invitational mathematics examination (AIME) 2025, 2025. URL https://huggingface.co/datasets/math-ai/aime25.

Yifan Zhang. American invitational mathematics examination (AIME) 2026, 2026. URL https://huggingface.co/datasets/math-ai/aime26.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11189–11204, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.603. URL https://aclanthology.org/2024.acl-long.603/.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. *CoRR*, abs/2601.18734, 2026. doi: 10.48550/ARXIV.2601.18734. URL https://doi.org/10.48550/arXiv.2601.18734.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In Yixin Cao, Yang Feng, and Deyi Xiong (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 400–410. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-DEMOS.38. URL https://doi.org/10.18653/v1/2024.acl-demos.38.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. AGIEval: A human-centric benchmark for evaluating foundation models. In Kevin Duh, Helena Gómez-Adorno, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, Findings of ACL, pp. 2299–2314. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-NAACL.149. URL https://doi.org/10.18653/v1/2024.findings-naacl.149.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *CoRR*, abs/2311.07911, 2023. doi: 10.48550/ARXIV.2311.07911. URL https://doi.org/10.48550/arXiv.2311.07911.

He Zhu, Junyou Su, Peng Lai, Ren Ma, Wenjia Zhang, Linyi Yang, and Guanhua Chen. Anchored supervised fine-tuning. In *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*. OpenReview.net, 2026. URL https://openreview.net/forum?id=PORko7QT64.

A KL CONSTRAINTS ENABLE CONTROLLED REWEIGHTING

To prevent uncontrolled trajectory shift, we formulate learning as a drift-constrained optimization problem:

$$\max_{\pi} \quad \mathbb{E}_{(q,a)}[\log \pi(a \mid q)] \qquad \text{s.t.} \qquad \mathbb{E}_q \left[\mathrm{KL}\big(\pi(\cdot \mid q) \,||\, \pi_{\theta_0}(\cdot \mid q)\big) \right] \leq \delta, \tag{9}$$

where $\delta > 0$ controls the average allowed deviation from the base policy π_{θ_0}.

The objective is defined at the answer level:

$$\log \pi(a \mid q) = \log \sum_{\tau \to a} \pi(\tau \mid q), \tag{10}$$

which depends on a marginalization over latent trajectories and does not explicitly specify how probability mass is distributed across trajectories.

To make this structure explicit, we introduce a trajectory-level view. For any distribution $p(\tau \mid q, a)$ supported on trajectories satisfying $\tau \to a$, Jensen's inequality gives

$$\log \pi(a \mid q) \geq \mathbb{E}_{\tau \sim p(\cdot \mid q,a)} \left[\log \pi(\tau \mid q) - \log p(\tau \mid q, a) \right]. \tag{11}$$

This lower bound shows that improving $\log \pi(a \mid q)$ corresponds to increasing probability mass on trajectories that produce the correct answer.

Under this trajectory-level perspective, the KL constraint restricts how much the trajectory distribution can deviate from the base policy. As a result, updates are realized through a controlled reweighting of trajectories, favoring those that support the correct answer while remaining close to π_{θ_0}.

Introducing a Lagrange multiplier $\lambda > 0$ for the KL constraint in Eq. 9, the resulting solution takes the form

$$\pi(\tau \mid q) \propto \pi_{\theta_0}(\tau \mid q) \exp\left(\frac{w(\tau, q, a)}{\lambda} \right), \tag{12}$$

where $w(\tau, q, a)$ is an implicit advantage-like quantity induced by the objective, reflecting the relative contribution of trajectory τ to increasing $\log \pi(a \mid q)$.

Thus, KL-constrained updates achieve improvement through controlled reweighting of trajectories, rather than unconstrained redistribution of probability mass.

B SUPPLEMENTARY ANALYSIS

B.1 DATA SCALING

To test whether the gains of specialist-generated supervision persist with more distillation data, we study scaling on CHEM SMol using rationales generated by the LST specialist π_{θ_1}, followed by FFT distillation. SMol contains roughly 3M examples in total; after generating $(\hat{\tau}, \hat{a})$ pairs and filtering them by answer correctness and completeness of the reasoning traces, we obtain a curated pool of 360K training examples. We scale the training set from 50K to the full 360K in 50K increments (Fig. 5). Increasing the amount of filtered distillation data leads to clear gains on IT (from 19.3 to 32.5), while ID remains virtually unchanged and OOD degrades only modestly overall. These results suggest that distillation from π_{θ_1} rationales has not saturated with a moderate amount of data and exhibits favorable scaling behavior with additional high-quality data.

B.2 CROSS-MODEL TRANSFERABILITY

To test whether specialist–student profile inheritance depends on a shared model family, we distill rationales from Intern-S1-mini (S1m; Bai et al., 2025), a science-specialized 8B model, into a Qwen3-8B student (Tab. 4). Crucially, the distilled student strictly inherits the behavioral profile of its specific teacher. While direct distillation from S1m already improves the student's IT substantially (19.34 $\rightarrow$ 31.33), distilling from the FFT-S1m-specialist maximizes student IT but

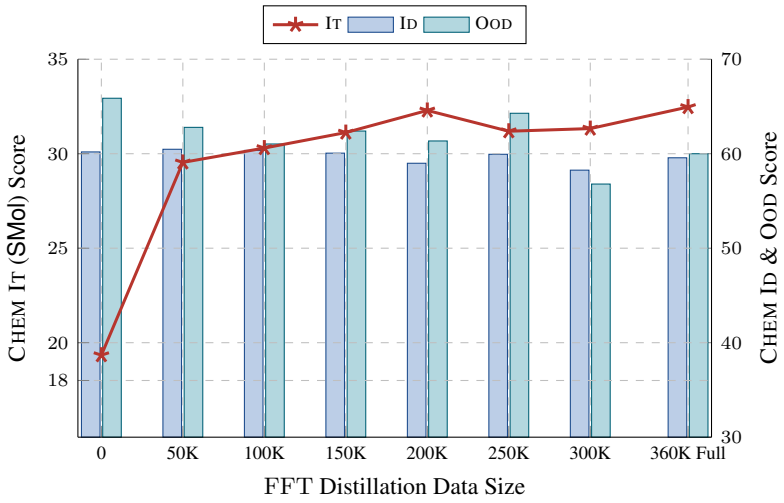


Figure 5: **Data scaling on SMol.** Rationales are collected from the LST *Specialist* π_{θ_1} and used for FFT distillation. Increasing the amount of filtered training data substantially improves IT performance, while leaving ID largely stable and causing only minimal degradation on OOD.

Table 4: **Cross-model Transferability on Chemistry.** FFT / LST-tuned Intern-S1-mini (π_{ϕ_0} -S1m) serves as the specialist π_{ϕ_1}, and π_ϕ-generated $(q, \hat{\tau}, \hat{a})$ are distilled into Qwen3-8B (3-8B).

Models on Chemistry	**IT** (14)	**ID** (4)	**OOD** ($5^{=}$)
Origin π_{θ_0} 3-8B	19.34	60.19	**42.22**
Origin π_{ϕ_0} -S1m	34.80	**62.51**	37.77
↳ FFT-*Distill* to $\pi_{\theta'_0}$ 3-8B	31.33	60.25	33.79
FFT *Specialist* π_{ϕ_1} -S1m	**42.83**	54.86	22.43
↳ FFT-*Distill* to $\pi_{\theta'_0}$ 3-8B	36.91	56.99	25.90
LST *Specialist* π_{ϕ_1} -S1m	33.11	60.31	36.21
↳ FFT-*Distill* to $\pi_{\theta'_0}$ 3-8B	32.72	58.79	37.46

severely degrades ID/OOD, exactly mirroring the FFT-teacher's extreme overfitting. Conversely, an LST-S1m-specialist imparts a balanced profile: the student preserves robustness comparable to S1m while still delivering a sizable IT gain (+13.38) from Qwen3-8B. These results confirm that structural constraints (LST) versus unconstrained tuning (FFT) fundamentally shape the rationale distributions, dictating downstream generalization even across model families. Beyond cross-family transfer, App. B.3 shows that our 8B specialist actually imparts better performance to an 8B student than a larger 14B teacher, hence specialist supervision surpasses the model scale.

B.3 CROSS-SCALE TEACHER–STUDENT GAP

We test whether a stronger teacher alone is sufficient for effective distillation. For CHEM, we use Qwen3-14B to generate $(q, \hat{\tau}, \hat{a})$ data, apply the same filtering procedure, and subsample to the same training size as *Self-Distill*, before continuing FFT on Qwen3-8B. This cross-scale baseline achieves 25.82/56.27/37.26 on CHEM IT/ID/OOD. Fig. 6 shows that cross-scale distillation (14B$\xrightarrow{d}$8B) improves IT by +6.48, but reduces ID and OOD by −3.92 and −4.96. In contrast, specialist distillation from the same 8B base yields better overall trade-offs, suggesting that teacher strength alone is insufficient and echoing that effective distillation also depends on the compatibility between generated supervision and the student learner (Xu et al., 2025).

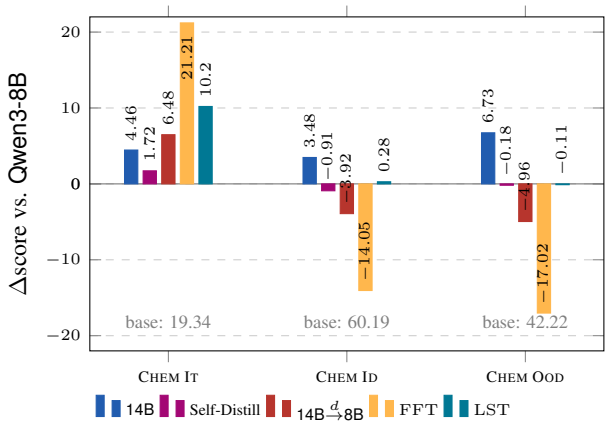
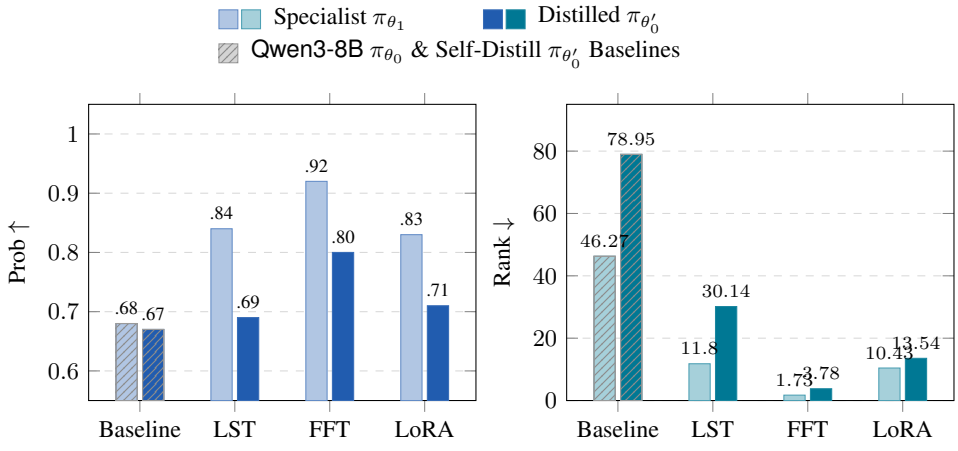


Figure 6: Delta performance on Chemistry relative to the Qwen3-8B origin model π_{θ_0}. The gray labels the absolute baseline score (as in Tab. 2) of π_{θ_0} for each category. 14B$\xrightarrow{d}$8B denotes the *cross-scale distillation* in contrast to *self-distill* baseline.



(a) Next-Token Probability **(b)** Rank

Figure 7: Comparison of Chemistry specialist π_{θ_1} and distilled $\pi_{\theta'_0}$ models across *teacher-forcing* next-token Probability (Higher ↑ is better) and Rank (Lower ↓ is better) metrics. Baseline groups (π_{θ_0} / $\pi_{\theta'_0}$) are shown with hatched overlay.

B.4 Additional Teacher-forcing Next-token Prediction Metrics: Probability and Rank

In addition to the Top-1 Hit Rate discussed in § 5.3, we also evaluate the next-token probability and rank using *teacher-forcing* on the same $2,000$ unseen samples.

As shown in Fig. 7, both Probability and Rank exhibit behaviors highly consistent with the Top-1 Hit Rate. Specifically, the unconstrained FFT specialist achieves near-perfect next-token probability (0.92) and an extremely low rank (1.73). This extreme over-confidence highlights a severe collapse of its trajectory distribution, indicating that the model arbitrarily distorts its generation manifold to overfit the answer marginal likelihood rather than learning a generalizable reasoning process.

Conversely, parameter-efficient architectures, namely LST and LoRA, demonstrate a much more controlled and calibrated improvement in both metrics. For instance, their specialist probabilities are anchored around 0.83–0.84, and their ranks are maintained at approximately 10.4–11.8. This empirical observation aligns well with the implicit structural regularization provided by parameter-efficient tuning. By restricting the optimization space, LST and LoRA prevent the trajectory dis-

tribution from degenerating into rote memorization. Consequently, these structural constraints enable a smoother, more robust transfer of rationale quality to the distilled student models ($\pi_{\theta_0'}$), effectively avoiding the downstream generalizability degradation typically caused by extreme overspecialization.

B.5 Model Scaling: Using Qwen3-14B as Origin Model

Table 5: **Model Scaling on Chemistry.** Performance of specialist models and their corresponding distilled models using Qwen3-14B as the origin model π_{θ_0}.

Models on Chemistry		IT (14)	ID (4)	OOD ($5^{=}$)
Origin π_{θ_0} Qwen3-14B		23.80	63.67	48.95
Tuned on π_{θ_0} 14B	FFT *Specialist* π_{θ_1}	59.10	37.50	37.14
	$\hookrightarrow$ FFT-*Distilled* $\pi_{\theta_0'}$	59.06	27.53	3.46
	LST *Specialist* π_{θ_1}	35.76	63.80	48.94
	$\hookrightarrow$ FFT-*Distilled* $\pi_{\theta_0'}$	33.40	60.68	29.54

To investigate whether the phenomena observed in our main experiments hold for larger models, we scale our origin model π_{θ_0} from Qwen3-8B to Qwen3-14B and replicate the QA-only specialist distillation pipeline on the **Chemistry** domain. We compare the FFT and LST tuning strategies. As shown in Tab. 5, scaling up the origin model reveals trends that are perfectly consistent with our findings in § 4.2:

- **Consistent Target-Domain Improvement:** Both FFT and LST specialists (π_{θ_1}), as well as their downstream distilled models ($\pi_{\theta_0'}$), significantly outperform the zero-shot Qwen3-14B origin model on the IT benchmark. This reinforces that QA-only specialist distillation remains highly effective for larger-scale models.
- **Specialization–Generalization Trade-offs:** The distinct trade-off profiles induced by different tuning strategies persist. FFT achieves extreme IT gains ($23.80 \rightarrow 59.10$) but suffers from severe catastrophic forgetting on ID and OOD. Alarmingly, the FFT-distilled model inherits and amplifies this toxicity, plummeting to 3.46 on OOD. In contrast, LST strikes a much healthier balance, providing robust IT improvements ($23.80 \rightarrow 35.76$) while tightly preserving both ID and OOD capabilities in the specialist, which translates safely into the distilled model.
- **Monotonic Rank Correlation:** The relative capabilities of the 14B-based specialists are strictly mirrored in their distilled counterparts. FFT yields a higher IT but lower OOD than LST in π_{θ_1}, and exactly the same ranking applies to $\pi_{\theta_0'}$. This further corroborates that the distilled model's behavior is deterministically governed by the trajectory distribution selected by the specialist tuning strategy.

B.6 Ablation Study on LoRA Configurations

We also analyze the effect of LoRA-specific design choices in the LRM setting. All experiments are evaluated on the same IT, ID, and OOD benchmark groups as in the main experiments.

Rank sensitivity. Tab. 6 reports the LRM performance of LoRA Specialist models trained with ranks $r \in \{16, 32, 64, 128\}$. The results show that LoRA performance is relatively stable for ranks 16, 32, and 64 on ID and OOD evaluation, while increasing the rank to 128 leads to a noticeable OOD drop. Rank 32 obtains the highest IT score, whereas rank 64 obtains the best ID score and remains nearly tied with ranks 16 and 32 on OOD. We therefore use rank 64 in the main experiments as a capacity-balanced default rather than tuning the rank to maximize a single IT score.

Standard LoRA vs. RsLoRA. We also compare standard LoRA with rank-stabilized LoRA (RsLoRA, Kalajdzievski, 2023) at rank 64. As shown in Tab. 7, RsLoRA performs poorly at the Specialist stage, especially on IT and OOD. Although FFT distillation from the RsLoRA Specialist partially recovers IT performance, its OOD score remains close to zero, indicating that the generated

Table 6: Rank sensitivity of LoRA Specialist π_{θ_1} on LRM using 7M LRM QA examples.

LRM Model	IT	ID	OOD
LoRA Specialist π_{θ_1}, $r = 16$	23.02	42.84	**41.45**
LoRA Specialist π_{θ_1}, $r = 32$	**23.47**	43.03	**41.45**
LoRA Specialist π_{θ_1}, $r = 64$	20.47	**43.16**	41.41
LoRA Specialist π_{θ_1}, $r = 128$	21.77	42.07	38.90

rationales do not provide transferable supervision. In contrast, standard LoRA yields a much more balanced Specialist model and produces rationales that lead to a substantially stronger FFT Distilled model on ID and OOD. This suggests that, in our LRM setting, RsLoRA changes the optimization dynamics in a way that is harmful to trajectory selection and downstream distillation quality.

Table 7: Comparison between standard LoRA and RsLoRA at rank 64 on LRM.

Model	IT	ID	OOD
RsLoRA Specialist π_{θ_1}, $r = 64$	0.51	19.36	0.66
LoRA Specialist π_{θ_1}, $r = 64$	**20.47**	**43.16**	**41.41**
FFT Distilled π'_{θ_0} from RsLoRA	**27.23**	21.94	0.83
FFT Distilled π'_{θ_0} from LoRA	22.97	**30.87**	**35.89**

C Experimental Setup Details

C.1 Training

All fine-tuning experiments are implemented using `LlamaFactory` (Zheng et al., 2024) on clusters of NVIDIA & AMD high-end GPUs. We uniformly employ `bfloat16` precision to accelerate training. Unless otherwise specified or when multiple runs with different seeds are required for averaging, we fix the random seed to 42 across all experiments to ensure reproducibility. We maintain a consistent effective batch size of 128 across all runs. Implementation-wise, we constrain

$$\begin{aligned}&\texttt{per_device_train_batch_size} \times \texttt{world_size} \\ &\times \texttt{gradient_accumulation_steps} \equiv 128.\end{aligned} \tag{13}$$

The exact combination varies dynamically to accommodate different computational requirements of the training data. The cutoff lengths are $16,384$ for CHEM, $15,384$ for PHYS, and $8,192$ for LRM. The specific configurations for each FT strategy are detailed below:

- FFT. We train models for 1 epoch by `AdamW` (Loshchilov & Hutter, 2019) with a 1% weight decay. The learning rate (lr) follows a cosine decay schedule, ascending to a peak of $\mathrm{lr}_{\max} = 1\times10^{-5}$ with a warmup ratio of 3% and subsequently decaying to a minimum of $\mathrm{lr}_{\min} = 2\times10^{-6}$.
- LST. Following Gao et al. (2025), we adopt the configuration reported to achieve the best performance on Qwen3-8B. We restrict trainable parameters to the bottom 4 and top 16 transformer layers with all other layers frozen. All other hyperparameters remain identical to the FFT setting.
- LoRA. We set the intrinsic rank $r = 64$, the scaling parameter $\alpha = 2r$, and a dropout rate of 5%. The peak $\mathrm{lr}_{\max}$ increases to 2×10^{-4} without a $\mathrm{lr}_{\min}$ constraint. The remaining align with the FFT setup.

C.2 Test-time Decoding

We use `vLLM` (Kwon et al., 2023) and `HuggingFace Transformers` (Wolf et al., 2020) for CoT generation and IT evaluation, enabling thinking mode with `max_seq_len` matching the training cutoff lengths (see App. C.1) and sampling parameters $T = 0.6, \texttt{TopP} = 0.95, \texttt{TopK} = 20, \texttt{MinP} = 0$ in accordance with the *Best Practices*[2] of Qwen3 series. For a sufficient pool of

[2] `https://huggingface.co/Qwen/Qwen3-8B#best-practices`

Table 8: **Overview of evaluation suites** across the three studied *target* domains (CHEM, PHYS and LRM) under *Training*, IT and ID settings, and three OOD benchmarks (complex reasoning for BBEH, math for AIME and coding for LCB). Bracketed terms (e.g., SuperGPQA[Chemistry, $\cdots$]) indicate that only these domain-relevant subset(s) are evaluated, rather than the entire benchmark. The numbers in parentheses in the table headers denote the number of evaluated subsets (or variants) for each *target* domain under the respective benchmark category. Note that the OOD benchmarks are shared across all three *target* domains. Notably, SMol officially consists of 14 subtasks (see App. D.1.1). For the LRM domain, we select 8 low-resource languages (see App. D.3) and evaluate bidirectional translation with English (en$\rightleftarrows$xx), yielding 16 subsets in total.

Domain	**Training** $(14/1/16)$	**IT** $(14/1/16)$	**ID** $(4/4/(2\times 8))$	**OOD** $(\underline{5})$
Chemistry (CHEM)	SMolInstruct [Train] (SMol) (Yu et al., 2024)	SMol[Test]	ChemBench (Mirza et al., 2025)	BIG-Bench Extra Hard (BBEH) (Kazemi et al., 2025) AIME 2025[a] & 2026[b] LCB V5[d] & V6[e]
			MMLU-Pro[Chemistry] (Wang et al., 2024)	
			SuperGPQA[Chemistry, Chemical Engineering and Technology] (M-A-P-Team et al., 2025)	
Physics (PHYS)	MegaScience [Physics] (Fan et al., 2025)	PHYSICS[Undergraduate /Postgraduate (Physics Major)] (Feng et al., 2025)	PIQA (Bisk et al., 2020)	
			AGIEval[Gaokao Physics] (Zhong et al., 2024)	
			MMLU[High School Physics] (Hendrycks et al., 2021)	
			MMLU-Pro[Physics]	
Low-Resource Multilingual -ism(LRM)	OPUS[en$\rightleftarrows${bn,cs,hu,sr, sw,te,th,vi}] (Tiedemann, 2012)	Flores-101[en$\rightleftarrows${bn,cs,hu,sr, sw,te,th,vi}] (Goyal et al., 2022)	IFEval[c] (Zhou et al., 2023)	
			GPQA[c] (Rein et al., 2023)	

[a] The American Invitational Mathematics Examination (AIME). https://huggingface.co/datasets/math-ai/aime25

[b] https://huggingface.co/datasets/math-ai/aime26

[c] These two ID benchmarks in the LRM domain are originally in English. We use their multilingual translations provided by BenchMAX (Huang et al., 2025), and evaluate on 8 non-English low-resources languages: {bn, cs, hu, sr, sw, te, th, vi} (see Tab. 10 in App. D.3 for their names and properties), rather than the original English versions.

[d] Problems released between May 2023 and January 2025, containing 880 problems. See https://github.com/livecodebench/livecodebench#dataset-versions for details.

[e] Problems released between May 2023 and April 2025, containing $1,055$ problems.

high-quality CoT rationales for distilled models, $\texttt{n} = 3$ candidate responses are generated for each query in CHEM and PHYS domains, and $\texttt{n} = 1$ for LRM.

D Details of Training Data and Evaluation Suites

Tab. 8 provides a comprehensive overview of all evaluation suites used across the three target domains (CHEM, PHYS, and LRM) under the Training, IT, and ID settings, as well as the shared OOD benchmarks spanning complex reasoning, mathematics, and coding.

We employ the widely adopted frameworks OpenCompass (Contributors, 2023), LM-Eval-Harness (Gao et al., 2024), and BenchMAX (Huang et al., 2025) to ensure a standardized, reproducible, and fair evaluation for ID and OOD, with the maximum sequence lengths $\texttt{max_seq_len} = 40,960$, i.e., the max_position_embeddings of Qwen3 series, and temperature $T = 0.6$.

D.1 CHEMISTRY

D.1.1 CHEMISTRY TRAINING AND IT DATA — SMOLINSTRUCT

SMolInstruct is a large-scale instruction tuning dataset crafted for CHEM domain.

Subtasks Composition. SMolInstruct comprises 14 instruction-following molecular (sub-)tasks: forward synthesis (FS), retrosynthesis (RS), molecule captioning (MC), molecule generation (MG), name conversion-i2f (I2F), name conversion-i2s (I2S), name conversion-s2f (S2F), name conversion-s2i (S2I), property prediction-esol (ESOL), property prediction-lipo (Lipo), property prediction-bbbp (BBBP), property prediction-clintox (ClinTox), property prediction-hiv (HIV), and property prediction-sider (SIDER). These tasks cover reaction prediction, molecular generation and understanding, conversion of molecular representations, and property prediction, providing a broad testbed in CHEM.

Subtask-Specific Evaluation Metrics. We adopt subtask-specific evaluation metrics according to Yu et al. (2024):

(1) **FS, RS and MG** These are molecular generation tasks. We evaluate them using *Morgan Fingerprint Tanimoto Similarity* (Morgan FTS, Morgan, 1965), which measures the similarity between the ground-truth molecule and the generated molecule. The score ranges from 0 to 1, where a larger value indicates higher structural similarity.

(2) **MC.** This task requires generating a textual description for a molecule. We evaluate the semantic similarity between the generated caption and the reference caption using METEOR score (Banerjee & Lavie, 2005), whose value lies in $[0.0, 1.0]$, the higher the more similar.

(3) **BBBP, ClinTox, HIV and SIDER.** These are binary classification tasks. We report naïve *accuracy* (Acc) as the evaluation metric.

(4) **I2F, I2S, S2F and S2I.** These four tasks require converting one molecular representation from another. We adopt variants of *exact match* (EM) accuracy, including *element match* and *split match*, to evaluate both element-order-independent exactness and partial structural consistency.

(5) **ESOL and Lipo.** These are numerical regression tasks. We use *Root Mean Square Error* (RSME) to measure the deviation between the predicted values and the ground-truth values. RSME ranges from 0 to $+\infty$, where a smaller value indicates better performance.

Overall Score. In Tab. 2, we report a unified overall score in the range $[0.00\%, 100.00\%]$ to summarize model performance across all 14 tasks. Since the raw metrics are heterogeneous and have different scales and optimization directions, we compute the SMol overall score as follows:

$$\begin{aligned}\text{Score}_{\text{SMol}} = 100\% \times \frac{1}{14}\Bigg\{ & \text{Morgan}_{\text{FS}} + \text{Morgan}_{\text{RS}} + \text{Morgan}_{\text{MG}} \\ & + \text{METEOR}_{\text{MC}} \\ & + \text{Acc}_{\text{BBBP}} + \text{Acc}_{\text{ClinTox}} + \text{Acc}_{\text{HIV}} + \text{Acc}_{\text{SIDER}} \\ & + \text{EM}_{\text{I2F}} + \text{EM}_{\text{I2S}} + \text{EM}_{\text{S2F}} + \text{EM}_{\text{S2I}} \\ & + \max\left(\frac{2 - \text{RMSE}_{\text{ESOL}}}{2}, 0\right) + \max\left(\frac{1.2 - \text{RMSE}_{\text{Lipo}}}{1.2}, 0\right)\Bigg\}.\end{aligned} \tag{14}$$

Stage-2 CoT Rationale Filtration Strategies. Unlike the testing evaluation protocols described above, which often allow for partial credit (e.g., structural similarity via Morgan FTS even if the generated molecule is not exactly identical to the ground-truth), our rationale filtration process *strictly* enforces equivalence to the ground-truth (GT) answers, if feasible. The guiding principle is to ensure the absolute high quality and correctness of the generated Chain-of-Thought (CoT) rationales. Based on this stringent equivalence principle and the distinct output formats, we re-categorize the 14 subtasks into 6 groups and apply tailored filtration rules:

I. **SMILES-based Tasks (FS, RS, MG, I2S).** The output format is *Simplified Molecular-Input Line-Entry System* (SMILES, Weininger, 1988), which represents molecular graphs as ASCII strings (multiple molecules are delimited by "`.`"). Since a single molecule can be legally represented by multiple valid SMILES strings, we utilize the standard `RDKit` (Landrum et al., 2026) Python library to convert both the generated and GT SMILES into their *canonical* forms. A rationale is retained only if its canonicalized SMILES perfectly matches the GT, ensuring absolute structural equivalence.

II. **Molecular Formula Tasks (S2F, I2F).** The output is a molecular formula (e.g., `C6H12O6`). We apply an *Element Match* rule: we extract the constituent elemental symbols and their corresponding counts from the response and compare them to the GT. The rationale is kept if the elements and their quantities are identical, completely disregarding the order in which the elements appear.

III. **IUPAC Naming Task (S2I).** The output is an *International Union of Pure and Applied Chemistry* (IUPAC, Favre & Powell, 2013) name (e.g., `2-methylpropane`). We employ a *Split Match* criterion. Both the generated name and the GT are tokenized by splitting at the hyphen (`-`). The rationale is preserved if the set of generated splits perfectly matches the set of GT splits, disregarding their relative order.

IV. **Binary Property Prediction (BBBP, ClinTox, HIV, SIDER).** The outputs are binary choices (e.g., `yes`/`no` or `True`/`False`). We parse the final predicted label via heuristic keyword matching and retain the rationale only if the parsed boolean value strictly aligns with the GT annotation.

V. **Numerical Property Prediction (ESOL, Lipo).** The outputs are continuous numerical values. We calculate the absolute difference between the parsed predicted value and the GT. Rationales are retained if and only if its absolute error is ≤ 1.0.

VI. **Text Generation (MC).** The output is a natural language description of a molecule. We calculate the METEOR score between the generated caption and the GT reference. A rationale is considered valid and thus retained if its METEOR score is ≥ 0.25.

Table 9: **Statistics of SMolInstruct across subtasks**, including the *downsampled* training split used for rapid experimentation, the *full* training split, and the official in-task (IT) test split. For the downsampled setting, each subtask is capped at 50K training instances; subtasks with fewer than 50K original examples are upsampled accordingly. For the full training setting, additional upsampling is applied to selected low-resource subtasks to mitigate data imbalance, resulting in a final rebalanced training set of ~3.6M instances.

SMol Subtasks	***Downsampled* Training Split**	***Full* Training Split**	**Test (IT) Split**
Total	$534,805$	$3,288,855 \xrightarrow{\text{upsampled}} 3,675,404$	$33,061$
FS	$50,000$	$971,809$	$4,062$
RS	$50,000$	$941,735$	$4,156$
MG	$50,000$	$56,498$	$2,493$
MC	$50,000$	$56,498$	$2,538$
BBBP	$1,569 \xrightarrow{5\times} 7,845$	$1,569 \xrightarrow{50\times} 78,450$	197
ClinTox	$1,144 \xrightarrow{5\times} 5,720$	$1,144 \xrightarrow{50\times} 57,200$	144
HIV	$32,864 \xrightarrow{\text{up to 50K}} 50,000$	$32,864$	$4,107$
SIDER	$22,820 \xrightarrow{\text{up to 50K}} 50,000$	$22,820 \xrightarrow{3\times} 68,460$	$2,860$
I2F	$50,000$	$300,000$	$2,993$
I2S	$50,000$	$299,890$	$2,993$
S2F	$50,000$	$299,890$	$2,993$
S2I	$50,000$	$299,890$	$2,993$
ESOL	$888 \xrightarrow{5\times} 4,440$	$888 \xrightarrow{50\times} 44,400$	112
Lipo	$3,360 \xrightarrow{5\times} 16,800$	$3,360 \xrightarrow{50\times} 168,000$	420

Data Scale and Preparation. SMolInstruct is constructed from a large-scale training corpus with substantial variation in data volume across subtasks. In its original form, the training split contains ~3.2M instances, while the test split contains ~33K instances. The exact number of examples for each subtask is reported in the third and fourth columns of Tab. 9.

To accelerate early-stage experimentation, ablation studies, and hyperparameter exploration, we additionally curate a downsampled training set, whose per-subtask sizes are summarized in the second column of Tab. 9. Specifically, for each subtask, we randomly sample up to 50K training instances. For six subtasks whose original training sets contain fewer than 50K examples, we retain all available instances and expand the corresponding subtask data to 50K through repeated replication (i.e., copying the full set 5 times at most) or, more generally, by upsampling to the target size.

We also note that the original training split exhibits data imbalance among subtasks. To alleviate this issue, we perform additional upsampling for five subtasks, as detailed in Tab. 9. After this rebalancing procedure, the final full training split used in our main experiments contains ~3.6M instances.

Prompt Template We formulate all tasks into a unified conversational format following the `ShareGPT` scheme. To elicit the rigorous systematic reasoning capabilities of the model, we employ a consistent system prompt across all tasks. This system prompt explicitly instructs the model to act as an expert reasoner and encapsulate its thinking process within `<think> ...</think>` tags. The user prompt consists of a task-specific instruction followed by the input question. The visual representation of our `ShareGPT`-style prompt template is illustrated in Fig. 8.

System

You are an expert reasoner with extensive experience in all areas. You approach problems through systematic thinking and rigorous reasoning. Your response should reflect deep understanding and precise logical thinking, making your solution path and reasoning clear to others. Please put your thinking process within `<think> ...</think>` tags.

User

[Task-Specific Instruction]
Question: *[Input Data]*

Figure 8: **The unified ShareGPT-style prompt template for SMolInstruct.** The model is prompted with a constant system message to encourage rigorous CoT reasoning. The user message provides the task-specific instruction alongside the input data.

The *[Task-Specific Instruction]* in Fig. 8 is specifically customized for each of the 14 subtasks to define the chemical context, specify the input format, and strictly constrain the desired output format (e.g., using specific XML tags or LaTeX `\boxed{}` commands). The exact text for each task's instruction is provided in Fig. 12.

D.1.2 Chemistry ID Benchmarks

To comprehensively evaluate the robustness and generalization capabilities of our model under distribution shifts within the chemical domain (i.e., the ID setting in Tab. 8), we employ three challenging, expert-level benchmarks. All evaluations in this section are conducted in a **zero-shot** setting using the `OpenCompass` (Contributors, 2023). We report **accuracy** (Acc) as the unified evaluation metric across all datasets. The detailed descriptions of these benchmarks are as follows:

- **MMLU-Pro [Chemistry]:** This is the chemistry-specific subset of MMLU-Pro (Wang et al., 2024). As an enhanced and more robust version of the original MMLU, MMLU-Pro significantly increases the task difficulty by expanding the number of distractor options and filtering out trivial questions, thereby providing a rigorous testbed for college- and professional-level chemical knowledge.
- **SuperGPQA [Chemistry, Chemical Engineering and Technology]:** This comprises two domain-relevant subsets derived from the highly challenging SuperGPQA benchmark suite (M-A-P-Team et al., 2025). The *Chemistry* subset specifically evaluates PhD-level theoretical understanding, experimental logic, and complex chemical problem-solving abilities that are difficult to bypass via simple search engines. The *Chemical Engineering and Technology* subset shifts the focus to practical engineering, assessing the model's capability to reason about industrial scaling, thermodynamics, applied materials science, and chemical processes.
- **ChemBench:** ChemBench (Mirza et al., 2025) is a comprehensive, domain-specific evaluation framework tailored for large language models in chemistry. It spans a wide range of subdisciplines (e.g., physical, organic, inorganic, and analytical chemistry) and evaluates models on their ability to understand chemical properties, follow reaction rules, and perform multi-step scientific reasoning.

D.2 Physics

D.2.1 Physics Training Data — MegaScience

MegaScience (Fan et al., 2025) is a large-scale, high-quality scientific reasoning dataset designed to enhance the specialized knowledge and logical reasoning capabilities of large language models across multiple disciplines, including physics, chemistry, and biology. During the construction of our training set, we extracted physics problems from the full corpus.

Data Generation Strategy. To enhance the diversity of the model's reasoning paths and improve the robustness of CoT generation, we executed three independent inferences for each physics problem using three different random seeds. The conversational prompt is illustrated as Fig. 9.

System

```
You are an expert physicist. Solve the following problem step by step.
After your reasoning, clearly state your final answer.
If the answer is a number, give the numeric value.
If the answer is an expression, write it in LaTeX.
Wrap your final answer in \\boxed{}.
```

User

[Question]

Figure 9: **Prompt template for MegaScience.**

Rule-Based Verification. To robustly compare the generated answer against the reference ground truth, the verifier sequentially applies five deterministic matching strategies:

(1) **Exact Match**: A direct, case-insensitive string comparison after basic LaTeX and whitespace normalization.

(2) **Numeric Evaluation**: Both answers are parsed as numerical values (seamlessly handling LaTeX scientific notation and fractions). They are deemed equivalent if they match within a 5% relative tolerance or a 10^{-8} absolute tolerance.

(3) **Symbolic Equivalence**: The framework utilizes `SymPy` to symbolically subtract the parsed LaTeX expressions and check for mathematical equivalence (i.e., simplifying to zero).

(4) **Normalized Matching**: If symbolic parsing fails, all formatting and non-alphanumeric characters (except basic operators) are stripped for a rigid structural comparison.

(5) **Substring Inclusion**: For remarkably short reference answers (e.g., under 60 characters), the prediction is accepted if it fully contains the reference string.

D.2.2 PHYSICS IT BENCHMARK — PHYSICS[UNDERGRADUATE/POSTGRADUATE(PHYSICS MAJOR)]

PHYSICS (Feng et al., 2025) is a comprehensive, large-scale, and bilingual (English and Chinese) dataset tailored to evaluate and enhance the physical reasoning capabilities of large language models. For this evaluation, we specifically focus on its most advanced subset to rigorously test expert-level physical reasoning.

Subtasks and Difficulty Composition. While the full PHYSICS dataset encompasses various educational stages, in this setting, we strictly isolate the Undergraduate/Postgraduate (Physics Major) difficulty level. This subset shifts the focus away from foundational concepts to highly complex, expert-level problem-solving. The evaluated problems span 5 major physics domains: Mechanics, Electromagnetism, Thermodynamics, Optics, and Modern Physics.

Answer-Type-Specific Evaluation Metrics (Rule+Model Framework). Given the diversity and complexity of physics answers, standard text matching alone is insufficient. We therefore adopt a hybrid **Rule+Model** evaluation framework consisting of a deterministic rule-based verifier followed, when necessary, by an LLM-as-a-judge fallback. The deterministic verifier applies answer-type-specific rules, including normalized exact matching, numerical comparison, symbolic equivalence checking, MCQ letter-set matching, and bilingual True/False normalization:

(1) **Numerical.** Numerical answers are extracted while accounting for scientific notation and fractions. They are considered equivalent when the prediction matches the reference within a 5% relative tolerance or a 10^{-8} absolute tolerance.

(2) **Expression and Equation.** Mathematical expressions and equations are first normalized to remove superficial LaTeX and formatting differences. We then attempt symbolic equivalence checking via `SymPy`, e.g., by simplifying the difference between the predicted and reference expressions. Normalized exact matching is also used when the two expressions can be directly matched after canonicalization.

(3) **Multiple Choice (MCQ) and True/False (T/F).** MCQ answers are evaluated through exact matching of the extracted option-letter set (e.g., sorting letters A–F before comparison). For T/F questions, case-insensitive normalization maps bilingual and abbreviated variants (e.g., "True", "正确", "yes", and "T") to a unified Boolean representation.

(4) **Interval.** Interval-valued answers (e.g., $[-1, 1]$) are evaluated using normalized string matching so that both boundary values and inclusion/exclusion symbols must agree with the reference.

(5) **Open-ended and Uncertain (LLM-as-a-Judge).** For open-ended answers, or whenever deterministic verification cannot confidently establish equivalence, the framework falls back to an LLM judge, specifically `gemini-3.1-flash-lite-preview`. Such cases commonly involve complex multi-variable LaTeX expressions, integrals, vectors, tensors, or other symbolic forms that cannot be reliably converted into standard `SymPy` representations without task-specific parsing rules. The judge determines whether the predicted and reference answers are mathematically or physically equivalent.

Fallback Frequency and Judge Configuration. The LLM fallback is frequently invoked because the Undergraduate/Postgraduate subset contains a large proportion of advanced symbolic and open-ended answers that cannot be robustly resolved by generic string or `SymPy`-based verification.

Across the 2,000 PHYSICS test problems (approximately 3,100–3,300 sub-answers per model), the fallback rates are highly consistent across evaluated model variants: 84.8% for the Qwen3-8B origin model (2,674 sub-answers), 84.9% for the FFT specialist (2,677), 85.9% for the LST specialist (2,707), 85.6% for the LoRA specialist (2,697), and 83.4% for Qwen3-14B (2,628). The similar fallback frequencies across model variants indicate that use of the LLM judge is primarily determined by the answer structure rather than by a particular model family.

For reproducibility, the LLM judge is queried with `temperature=0.0` using greedy decoding. Its system prompt restricts the output to a single binary token-level decision, either `CORRECT` or `INCORRECT`, as shown in Fig. 11. The same zero-shot judge, prompt, decoding configuration, and verification procedure are applied uniformly to all origin, baseline, specialist, and distilled models to ensure a consistent comparison.

Overall Score and Multi-Part Logic. Unlike standard QA benchmarks, many questions in PHYSICS are multi-part, requiring the model to generate a sequence of answers. We prompt the model to encapsulate each sub-answer in its own separate `\boxed{}`. We apply a strict **AND logic**: an item is considered fully correct (yielding an Accuracy of 1) if and only if *all* of its sub-answers are judged correct against their corresponding ground-truth references. Partial credits are tracked during evaluation, but the primary reported metric remains the strict overall Accuracy (Acc).

Prompt Template. We unify the evaluation under a consistent conversational format. As illustrated in Fig. 10, the system prompt explicitly establishes the identity of an expert physicist, guiding the model to reason step-by-step before finalizing answers in the required LaTeX `\boxed{}` format. For open-ended or uncertain cases requiring model-based verification, we use the dedicated LLM-as-a-judge template in Fig. 11, which instructs the judge to assess mathematical and physical equivalence and return only a binary verdict (`CORRECT` or `INCORRECT`).

System

You are an expert physicist. Solve the following problem step by step.
After your reasoning, clearly state your final answer(s).
- If the answer is a number, give the numeric value.
- If the answer is a mathematical expression, write it in LaTeX.
- If the question is multiple-choice (MCQ), state the correct option letter(s).
- If the question is True/False, state True or False.
Wrap EACH final answer in \\boxed{}.
If there are multiple sub-answers, put each one in its own separate \\boxed{}.

User

[Question]

Figure 10: **The unified ShareGPT-style prompt template for PHYSICS.**

D.2.3 PHYSICS ID BENCHMARKS

In order to thoroughly assess the model's proficiency in the physical sciences and its resilience to domain-specific distribution shifts, we select five rigorous benchmarks that span various dimensions of physical reasoning. All evaluations are implemented utilizing the `LM-Evaluation-Harness` framework (Gao et al., 2024). We adopt specific prompting strategies (i.e., zero-shot or 5-shot) and metrics tailored to the standard evaluation protocol of each respective dataset. The details of the selected benchmarks and their specific configurations are outlined below:

System

You are a strict answer-verification assistant for physics problems.
Determine whether the student's answer is mathematically/physically equivalent to the reference answer. Consider different notations, units, rounding, and equivalent expressions. Respond with ONLY the single word CORRECT or INCORRECT.

User

Reference Answer:*[ref]*
Student Answer: *[model_ans]*
Verdict:

Figure 11: **The LLM-as-a-judge prompt template for PHYSICS.**

- **PIQA:** Physical Interaction: Question Answering (PIQA) (Bisk et al., 2020) evaluates a model's physical commonsense reasoning. Instead of focusing on abstract academic formulas, it requires an understanding of the affordances of everyday objects and basic physical mechanics to deduce logical outcomes in real-world scenarios. We evaluate this dataset in a **zero-shot** setting and report the length-normalized accuracy (Acc_norm).
- **AGIEval[Gaokao Physics]:** This subset from the human-centric AGIEval benchmark (Zhong et al., 2024) consists of highly challenging, standardized physics questions derived from the Chinese National College Entrance Examination (Gaokao). It assesses the model's ability to tackle complex, multi-step physical problem-solving that requires rigorous mathematical reasoning. This is evaluated in a **zero-shot** setting, with performance measured by length-normalized accuracy (Acc_norm).
- **MMLU[High School Physics]:** Derived from the Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2021), the task covers standard curriculum topics ranging from basic kinematics to advanced theoretical concepts. We evaluate in a **zero-shot** setting, using standard accuracy (Acc) as the metric.
- **MMLU-Pro[Physics]:** As the physics-specific subset of MMLU-Pro (Wang et al., 2024), this benchmark significantly increases the evaluation difficulty by expanding the number of distractor options to ten and systematically filtering out trivial questions. We evaluate this subset in a **5-shot** setting and report the Exact Match (EM) metric.

D.3 Low-Resource Multilingualism

D.3.1 Language Selection

We focus on all 8 low-resource languages Tab. 10 among the 17 languages supported by `BenchMAX` (Huang et al., 2025). Our language selection is motivated by the observation that the Qwen3 series are relatively undertrained in low-resource languages, which makes them a suitable test bed to evaluate the effectiveness of our pipeline in challenging LRM settings. To determine whether a language is categorized as high- or low-resource, we follow the classification logic introduced in Tu et al. (2025). The selected 8 languages span 5 writing systems and 6 language families, providing substantial typological and orthographic diversity. Thus we consider this set sufficiently representative for assessing the performance of our pipeline on low-resource multilingualism.

Table 10: **Properties of** 8 **low-resource languages** we experiment for LRM domain and English language, in ascending order of ISO 639-1 codes (for Standardization, 2023).

ISO Code	ISO Language Name	Endonym	Writing System	Language Family
bn	Bengali	Bāŋlā	Bengali–Assamese	Indo–European
cs	Czech	Čeština	Latin	Indo–European
en	English	English	Latin	Indo–European
hu	Hungarian	Magyar nyelv	Latin	Uralic
sr	Serbian	Srpski	Cyrillic	Indo–European
sw	Swahili	Kiswahili	Latin	Niger–Congo
te	Telugu	Telugu	Telugu	Dravidian
th	Thai	Phasa Thai	Thai	Kra–Dai
vi	Vietnamese	tiéng Viêt	Latin	Austroasiatic

D.3.2 LRM TRAINING DATA— OPUS

Our LRM training data are directly sourced from Lego-MT (Yuan et al., 2023), who curated and cleaned large-scale parallel corpora from OPUS (Tiedemann, 2012), the largest publicly available collection of translated texts aggregating data from diverse domains including legislative proceedings, subtitles, localization files, and web-crawled content. We use their preprocessed English↔target parallel sentence pairs for each of our 8 low-resource target languages. The resulting training set sizes vary considerably across languages, reflecting the inherent data availability imbalance in the low-resource regime. For the rationale filtration, we apply a quality-based filtration strategy: for each translation direction (i.e., `en` $\rightarrow$ `xx` and `xx` $\rightarrow$ `en`), we retain only the top 20% of parallel pairs ranked by `spBLEU` (Papineni et al., 2002; Goyal et al., 2022) scores, ensuring that the curated training signal is of sufficiently high quality for distilled students.

D.3.3 LRM IT BENCHMARK — FLORES-101

We evaluate IT LRM performance on Flores-101 (Goyal et al., 2022), a standardized multilingual translation benchmark covering 101 languages with approximately 1K professionally translated sentences in the `devtest` split. We report results for both English→target and target→English directions.

Although our training data come from OPUS, we use Flores-101 for evaluation because it provides a clean, high-quality, and widely adopted test set for multilingual machine translation. We treat it as an IT benchmark in the sense that it matches our training setup at the task level: both involve sentence-level translation in the same language directions, even though the test set is not drawn from the same corpus collection as OPUS.

D.4 LRM ID BENCHMARKS

IFEval. To evaluate the models in a shifted-in-domain (ID) setting, we first assess their capability in **Rule-based Instruction Following** using the IFEval dataset (Zhou et al., 2023). IFEval is designed to test whether LLMs can strictly adhere to specific formatting requirements and verifiable constraints (e.g., word count limits and specific output formats). Rather than using the original English dataset, we adopt the high-quality multilingual translated version curated by `BenchMAX` (Huang et al., 2025). This allows us to measure how well instruction-following capabilities transfer to low-resource scenarios. We evaluate the models on the 8 selected low-resource languages detailed in Tab. 10 and report the average score across these languages. Following Llama-Team (2024), the reported IFEval accuracy is computed as the average of four metrics: *prompt-strict*, *prompt-loose*, *inst-strict*, and *inst-loose* accuracies.

GPQA. Furthermore, we evaluate the models' proficiency in general **Science Reasoning** using the GPQA benchmark (Rein et al., 2023). GPQA consists of challenging, graduate-level questions spanning various scientific domains such as physics, biology, and chemistry, requiring deep logical

reasoning and domain knowledge. Similar to our IFEval setup, we do not evaluate on the original English version of GPQA. Instead, we utilize the translated datasets provided by `BenchMAX` (Huang et al., 2025) to investigate whether complex reasoning skills can be effectively elicited in undertrained languages. Consistent with our methodology, the reported results represent the average performance across the same 8 low-resource languages (Tab. 10).

D.5 OOD BENCHMARKS

To evaluate the generalization capabilities of our model beyond the *target* training domain and to verify whether our domain-specific reasoning enhancements translate to general contexts, we assess the model on three highly rigorous out-of-Domain (OOD) benchmarks covering complex reasoning, mathematics, and coding. All experiments in this section are conducted in a **zero-shot** setting using the `OpenCompass` framework (Contributors, 2023). The specific metrics and configurations for each benchmark are detailed below.

Aggregated OOD Score. The three OOD domains adopt different official evaluation protocols (harmonic mean for BBEH, average accuracy for AIME, and $\mathrm{Pass@1}$ for LCB) and contain an unequal number of constituent benchmarks (1 for complex reasoning, 2 for mathematics, and 2 for coding). To provide a single, balanced indicator that faithfully reflects cross-domain generalization, we instead report a **macro-average** across the three OOD domains in Tab. 2, giving each domain equal weight regardless of the number of datasets it contains. Formally, let S_{reason}, S_{math}, and S_{code} denote the per-domain scores, defined as

$$\begin{aligned} s_{\text{reason}} &= \mathrm{HarmonicMean}_{\mathsf{BBEH}}, \\ s_{\text{math}} &= \frac{1}{2}\left(\mathrm{Acc}_{\mathsf{AIME2025}} + \mathrm{Acc}_{\mathsf{AIME2026}}\right), \\ s_{\text{code}} &= \frac{1}{2}\left(\mathrm{Pass@1}_{\mathsf{LCB\text{-}V5}} + \mathrm{Pass@1}_{\mathsf{LCB\text{-}V6}}\right), \\ s_{\text{OOD}} &= \frac{1}{3}\left(s_{\text{reason}} + s_{\text{math}} + s_{\text{code}}\right). \end{aligned} \tag{15}$$

The OOD column reported in Tab. 2 is s_{OOD}, which is computed according to Eq. 15, rather than as a flat arithmetic mean over the 5 individual benchmark scores.

D.5.1 COMPLEX REASONING — BIG-BENCH EXTRA HARD

BIG-Bench Extra Hard (BBEH, Kazemi et al., 2025) is explicitly designed to push the frontier of general-purpose reasoning evaluation for frontier LLMs. It is constructed as a successor to BIG-Bench Hard (BBH, Suzgun et al., 2023), where each of the original BBH tasks is replaced by a substantially more challenging variant that probes a similar underlying reasoning skill but at a markedly increased level of difficulty. The resulting suite spans a diverse spectrum of reasoning competencies, including formal and informal deduction, multi-hop and counterfactual reasoning, long-context understanding, causal analysis, error identification, and the manipulation of complex symbolic, linguistic, and commonsense structures. Owing to the extreme difficulty and broad coverage of BBEH, even the strongest contemporary reasoning models attain only modest scores, making it a particularly stringent testbed for assessing whether reasoning capabilities acquired in our target domain transfer to genuinely out-of-distribution and cognitively demanding scenarios. We report the *(adjusted) harmonic mean*[3] of per-task accuracies across all BBEH subtasks as the final metric, following the official evaluation protocol.

D.5.2 MATHEMATICS — AIME 2025 AND 2026

The American Invitational Mathematics Examination (AIME, Zhang, 2025; 2026) is a prestigious, highly competitive mathematical Olympiad benchmark that demands exceptionally deep multi-step logical reasoning and advanced problem-solving skills. To rigorously test the model's mathematical reasoning capabilities on the most up-to-date problems, we utilize the latest AIME 2025 and AIME 2026 test suite. To account for the variance in generating complex mathematical CoT derivations,

[3]To deal with zero values, Kazemi et al. (2025) add a value of 1 to all accuracy numbers.

we evaluate the model across 32 independent runs and report the *average accuracy* over these 32 runs.

D.5.3 CODING — LIVECODEBENCH

LiveCodeBench (LCB, Jain et al., 2025) is a dynamic, continuously updated evaluation framework designed to rigorously assess the code generation capabilities of LLMs while inherently preventing data contamination. It collects newly published algorithmic problems from competitive programming platforms (e.g., LeetCode, Codeforces, and AtCoder). In our evaluation, we utilize two recent temporal splits to ensure absolute OOD rigorousness: **V5** (from August 2024 to February 2025) and **V6** (from February 2025 to May 2025). The performance on both splits is measured using the standard $\mathrm{Pass@1}$ metric.

Forward Synthesis (FS): You are an expert chemist. Given the SMILES representation of reactants and reagents, your task is to predict the potential product using your chemical reaction knowledge. The input contains both reactants and reagents, and different reactants and reagents are separated by ".". Your reply should contain the SMILES representation of the predicted product wrapped in <SMILES> and </SMILES> tags. Your reply must be valid and chemically reasonable.

Retrosynthesis (RS): You are an expert chemist. Given the SMILES representation of the product, your task is to predict the potential reactants and reagents using your chemical reaction knowledge. The input contains the SMILES representation of the product. Your reply should contain the SMILES representation of both reactants and reagents, and all reactants and reagents should be enclosed **together** within a single pair of <SMILES> and </SMILES> tags, separated by ".". Your reply must be valid and chemically reasonable.

Molecule Captioning (MC): You are an expert chemist. Given the SMILES representation of a molecule, your task is to describe the molecule in natural language. The input contains the SMILES representation of the molecule. Your reply should contain a natural language description of the molecule. Your reply must be valid and chemically reasonable.

Molecule Generation (MG): You are an expert chemist. Given the description of a molecule, your task is to generate the potential SMILES representation of the molecule. The input contains the description of the molecule. Your reply should contain the potential SMILES representation of the molecule wrapped in <SMILES> and </SMILES> tags. Your reply must be valid and chemically reasonable.

Name Conversion (I2F): You are an expert chemist. Given the IUPAC representation of compounds, your task is to predict the molecular formula of the compound. The input contains the IUPAC representation of the compound. Your reply should contain only the molecular formula of the compound wrapped in <MOLFORMULA> and </MOLFORMULA> tags and no other text. Your reply must be valid and chemically reasonable.

Name Conversion (I2S): You are an expert chemist. Given the IUPAC representation of compounds, your task is to predict the SMILES representation of the compound. The input contains the IUPAC representation of the compound. Your reply should contain only the SMILES representation of the compound wrapped in <SMILES> and </SMILES> tags and no other text. Your reply must be valid and chemically reasonable.

Name Conversion (S2F): You are an expert chemist. Given the SMILES representation of compounds, your task is to predict the molecular formula of the compound. The input contains the SMILES representation of the compound. Your reply should contain only the molecular formula of the compound wrapped in <MOLFORMULA> and </MOLFORMULA> tags and no other text. Your reply must be valid and chemically reasonable.

Name Conversion (S2I): You are an expert chemist. Given the SMILES representation of compounds, your task is to predict the IUPAC representation of the compound. The input contains the SMILES representation of the compound. Your reply should contain only the IUPAC representation of the compound wrapped in <IUPAC> and </IUPAC> tags and no other text. Your reply must be valid and chemically reasonable.

Property Prediction (ESOL): You are an expert chemist. Given the SMILES representation of compounds, your task is to predict the log solubility of the compound. The input contains the SMILES representation of the compound. Your reply should contain the log solubility of the compound wrapped in \boxed{}. Your reply must be valid and chemically reasonable.

Property Prediction (Lipo): You are an expert chemist. Given the SMILES representation of compounds, your task is to predict the octanol/water partition coefficient of the compound. The input contains the SMILES representation of the compound. Your reply should contain the octanol/water partition coefficient of the compound wrapped in \boxed{}. Your reply must be valid and chemically reasonable.

Property Prediction (BBBP): You are an expert chemist. Given the smiles representation of the compound, your task is to predict whether blood-brain barrier permeability (BBBP) is a property of the compound. The input contains the compound. Your reply should only contain Yes or No. Your reply must be valid and chemically reasonable.

Property Prediction (ClinTox): You are an expert chemist. Given the smiles representation of the compound, your task is to predict whether the compound is toxic. The input contains the compound. Your reply should contain only Yes or No. Your reply must be valid and chemically reasonable.

Property Prediction (HIV): You are an expert chemist. Given the smiles representation of the compound, your task is to predict whether the compound serve as an inhibitor of HIV replication. The input contains the compound. Your reply should contain only Yes or No. Your reply must be valid and chemically reasonable.

Property Prediction (SIDER): You are an expert chemist. Given the smiles representation of the compound, your task is to predict whether the compound has any side effects. The input contains the compound. Your reply should contain only Yes or No. Your reply must be valid and chemically reasonable.

Figure 12: **Task-specific instructions for the** 14 **subtasks of SMolInstruct.** Each instruction defines the chemical context, specifies the input format, and constrains the output format.