参赛学生姓名：冯子上

中学：北京师范大学附属实验中学

省份：北京

国家/地区：中国

指导老师姓名：宋飞、何道

指导老师单位：清华大学物理系 /
北京师范大学附属实验中学

论文题目：Gewu: Building, Enhancing, and Evaluating a Tool-Augmented 9B Physics Tutoring System

2026 T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

Gewu: Building, Enhancing, and Evaluating a Tool-Augmented 9B Physics Tutoring System

Zishang Feng

The Experimental High School Attached to Beijing Normal University

2026 S. T. Yau High School Science Award
仅用于2026丘成桐中学科学奖公示

Abstract

Students who get stuck on a physics problem rarely have a tutor within reach, and the language models that do best on olympiad-level physics live behind commercial interfaces: they charge per question, the questions leave the school, and the school controls neither the model nor its availability. Many physics departments own a consumer graphics card that sits idle. This paper asks whether such a card can run a physics tutor a department could trust.

Small open models can be improved in three known ways: by training on a stronger model's worked solutions, by attaching a computer-algebra tool for the arithmetic they get wrong, and by answering several times and voting. Each is reported to help. Physics itself has had little attention at this scale: the sets that exist are built to rank the strongest models, and what a model small enough for one card can do in physics, and how to improve it, is rarely studied. Where gains are reported, the baseline was often not given the same tool or prompt as the improved model. Public physics benchmarks ship enough wrong answer keys to change a conclusion, and a small model's accuracy swings with the prompt it is given.

We therefore measured before we built. The rules of measurement were fixed first, written down before the runs they judge, and under them we built GewuBank-20K: 20,414 physics problems on a two-dimensional difficulty grid, answer keys checked in five stages, every difficulty measured on the base model rather than declared, with evaluation sets held out from training and the isolation checked against a positive control. Measured on that bank, fine-tuning the 9B model we build on, Qwen3.5-9B, on a stronger model's solutions changed how it works more than what it knows. The fine-tuned model derives as its teacher did, in Chinese; it reaches its answer in roughly a third to a half of the base model's tokens; trained on tool-free traces, with the tool declared only when it is served, it calls the computing tool in the declared format, and its calls run, although it never saw one in training; its accuracy depends less on the prompt it is given; and where its accuracy rises, it rises with that behaviour. What a small model can still gain on hard problems therefore comes at the moment of answering: from the tool when the model calls it, and from sampling the base model several times and voting.

Gewu is the tutor those findings shaped: one 8-bit weight file on a single 24 GB card, no outside service, the computing tool on both of its tiers. Everyday questions go to the fine-tuned model, which answers in one pass; harder problems go, on request, to eight tool-using samples of the base model and a vote. On the olympiad benchmark PhysElite, with text only, the fine-tuned model with its tool is level with models many times its size, and on the endorsed set, the part of that benchmark whose keys a second model reproduced, it is ahead of its base. GewuBank-20K and the rules behind every number we measured are built for reuse.

Keywords: physics tutoring; small language models; tool-integrated reasoning; test-time search; supervised fine-tuning; measurement protocol

Contents

1 Introduction 2
2 Related Work 4
3 The Measurement Protocol and GewuBank-20K 6
4 The Gewu System 10
5 Experiments 14
6 Discussion 25
7 Conclusion and Future Work 28
A Pre-Registered Criteria 29
B Prompt Contracts 29
C Alternatives Considered and Why They Were Not Adopted 31
D Reproducibility 32
References 35
Acknowledgements 37

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示



Figure 1. Gewu at a glance. One 8-bit weight file serves two tiers on a single 24 GB GPU with no external API at serving time. Top to bottom, the blocks are the paper's three contributions (Section 1.4): the measurement protocol and the bank; the served system, shown answering the magnetic scalar potential of a uniformly magnetised sphere on the default tier and Kepler's equation for the eccentric anomaly of a satellite, solved in the sandbox, on the enhancement tier; and what the traces taught and where the accuracy comes from. Every number in the figure is measured in Section 3, Section 4 or Section 5.

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖公众

1 Introduction

1.1 Background and Motivation

Frontier language models have begun to score on olympiad-level physics problems [1], behind an API. A university physics department is in a different situation. It has students whose questions it cannot send to an API, and an idle 24 GB consumer GPU, of the kind many teaching labs already own, that holds about 28 sessions at once at an 8k context. This paper asks whether that card can be turned into a physics tutor the department would trust.

Three requirements follow. Accuracy: a tutor that is wrong too often is worse than none, and a 9 B model's accuracy on hard problems swings with the instructions around the question, so it has to hold under one plain, fixed contract and be raised where the model alone cannot manage. Speed: an answer has to stream while a student waits, and one card has to serve many students at once, so tokens per answer and sessions per card are measured, not assumed. Locality: no external API on any serving path, so the questions stay in the department, and the model, its tool sandbox and any search over samples live on one 24 GB card that the 16-bit weights of a 9 B model alone would nearly fill.

1.2 Challenges

Three difficulties shaped this work.

The first is measurement. Public answer keys are wrong often enough to change a conclusion. One audited set [2] had wrong keys in about a fifth of its sample and was kept out of the rulers, and PhysElite [1], our external anchor, had content errors in roughly one item in eight. The baseline is not a fixed object. Changing the prompt moves the same base model on the same problems further than the fine-tuning this paper serves moves it (Section 5.4). And the comparison procedure can manufacture an effect. A naive comparison of a fine-tuned model against its base reports a large gain, and three fairness requirements—a verified grader, a matched budget, the same tool and prompt on both sides—leave a gain inside the noise floor (Section 5.3). A ruler had to be built before it could be used.

The second is distillation. A 9 B student does not simply absorb a strong teacher's traces. Every instruction in the teacher's prompt reached the traces verbatim and the student learned to recite it; traces from a tool-using teacher taught the student to think less before each call and to lose its way in the tool loop; distilling on problems matched to the hard set did not beat the base (Appendix C). What the traces reliably transferred was the teacher's way of carrying and closing a derivation.

The third is where accuracy on hard problems can come from once the weights are fixed. Most of the base model's errors on the bank's numeric problems carry no tool call at all; a tool helps only on the problems that use it, and a vote helps only over samples that are each good. The system therefore has to decide when to compute and which model to sample, the two decisions that give it its shape.

1.3 Our Approach

Gewu is a physics tutoring system with two serving tiers and one set of weights (Figure 3). The fine-tuned model is the base model plus a 173 MB LoRA adapter, trained on tool-free reasoning traces, so the tiers share one weight file and differ in the adapter, the contract and the sampler. The default tier attaches the adapter and answers in a single pass, with a local Python sandbox for the arithmetic. This is the product. The enhancement tier detaches the adapter, samples the base model eight times with the sandbox, and reconciles the samples by a vote that never sees the answer key. The tier is offered on request for harder problems, and on the internal hard set it clears the acceptance line fixed before the run.

Three findings organise the paper and give the system its shape. A naive comparison overstates what fine-tuning buys: three fairness requirements bring its gain on our own bank to the noise floor,

2026 Si-Tayuan High School Science Award
仅用于2026年成都市学科学论文公示

yet each share they remove is still a real difference between the two models, in answer format, in tokens, or in tool and contract (Section 5.3). Fine-tuning shapes behaviour more than it adds physics. Under balanced measurement the fine-tuned model reasons in Chinese, carries a derivation step by step in the teacher's form in roughly a third to a half of the base model's tokens, so an answer streams sooner and one card serves more questions, generalises tool syntax from a serve-time declaration with no tool data in training, and swings less than the base model when the prompt changes; where its accuracy rises it rises with that behaviour, and on our own bank it is level with the base at less than half the tokens.

The fine-tuned model's greedy answer already sits close to the best its samples can vote to, while the base model's vote climbs far above its single answer, so the enhancement tier samples the base and pays for it in time.

The accuracy gains are produced at answer time: a tool call is worth a large and statistically solid gain on the problems that use one, and on the hard set sampling the base model eight times and voting is worth a gain of the same order over its best single answer, measured against a no-op control. A reinforcement-learning pilot on the finished system is reported in Section 5.6. Hence the shape: the fine-tuned model in front, and the base model's samples behind it when a problem is hard.

1.4 Contributions

This work makes the following contributions:

- **A measurement protocol and GewuBank-20K.** The protocol came first, because the measurement was the problem. A naive comparison of a fine-tuned model against its base on our own bank reported a large gain; levelled, the gain falls to the noise floor at home and survives abroad. The rules are ordinary experimental hygiene, assembled here for this kind of claim: criteria fixed before the run, baselines levelled on grader, budget, tool and prompt, a positive control, and audited public keys. What the paper adds is the measurement of what they change, and an instrument on which they can be run. GewuBank-20K is that instrument: 20,414 physics problems on a two-axis grid (analytic or numeric answer; three tiers), built for control rather than for difficulty, its keys checked in five stages, every difficulty cell measured on the base model rather than declared, held-out rulers isolated from training and a training pool cleaned of its overlap with public test sets, so later work can train and evaluate on it.
- **Behaviour from training, accuracy from answer time.** Under a levelled comparison, fine-tuning a 9 B model on a strong teacher's traces changed its behaviour more than its physics: the fine-tuned model matches its base on our own bank and is ahead on an external olympiad set, a gain that arrives with the behaviour it learned. What the traces transfer is the teacher's procedure. The fine-tuned model carries a derivation to its close in a fraction of the base model's tokens. It reasons in Chinese. It swings less when the instructions change. It calls the tool in the declared format, and its calls run, with no tool data in training, under a recipe that three ablations support and that we have not found in prior work: a plainly prompted teacher, tool-free traces, and the tool declared only at serving time. For this base model, what is left to improve lies at answer time: in the behaviour that carries a derivation, which fine-tuning shapes, and in where computation and search are spent.
- **Gewu, a small, deployable tutor with an external anchor.** One 8-bit weight file serves both tiers on a single 24 GB GPU with no external API: the fine-tuned adapter answers everyday questions in one pass, in a fraction of the base model's tokens, and problems it cannot settle go to eight tool-using samples of the base model, reconciled by a vote that never sees the key. The accuracy the system adds on hard problems is produced at answer time, by the sandbox where the model calls it and by the vote, each measured against a control. Text-only on PhysElite, the fine-tuned model with its tool is level with models many times its size; on the endorsed set it is ahead of its base and far ahead of a same-size model on the same harness. The quantised

2026 S.T. Yau High School Science Award
仅用于2026年丘成桐中学科学奖论文公示

weights passed a gate fixed before the run.

1.5 Organization

Section 2 places the work among physics benchmarks, distillation, tool use, reinforcement learning with verifiable rewards (RLVR), test-time search and small-model deployment. Section 3 describes the protocol and the dataset, Section 4 the system, Section 5 the experiments; Section 6 interprets the central finding and Section 7 concludes. Alternatives considered and not adopted are in Appendix C.

2 Related Work

2.1 Physics Benchmarks for Language Models

Physics problems differ from the mathematics problems on which most reasoning benchmarks are built: an answer carries a magnitude and a unit, the modelling step comes before any algebra, and in olympiad-level sets many problems are only meaningful with a figure. Benchmarks for this space have arrived quickly. UGPhysics [3] covers undergraduate physics across thirteen subjects in Chinese and English; PHYSICS [4] is a larger textbook-derived set that reaches graduate level; PHYBench [5] evaluates physical reasoning with an expression-level metric; OlympiadBench [6] and PhysElite [1] target olympiad problems, the latter reporting frontier models under a three-judge protocol; ABench-Physics [2] emphasises high-difficulty, dynamically varied problems; SciBench [7] spans college-level science more broadly. Different from these approaches, we take no external answer key as given: every set's error rate is audited before use, and a row that still rests on shipped keys says so (Section 3.6); our own bank is built as a controllable coordinate system rather than as a harder benchmark.

2.2 Long-CoT Distillation and Supervised Fine-Tuning

Reasoning models produce long chains of thought—exploration, verification and backtracking before an answer—and a small model can be taught the form of that behaviour by supervised fine-tuning on traces from a stronger one. DeepSeek-R1 [8] released models distilled this way; s1 [9] showed that about a thousand selected traces suffice; Loong [10] synthesises long chains at scale and keeps those a verifier accepts; OpenThoughts [11] scales open reasoning data across mathematics, code and science. Rejection sampling against a verifier is the common filter [12], with roots in self-taught reasoning [13].

Whether a small student can use such traces is a separate question. The capacity gap between teacher and student is an old observation in distillation: a larger, more accurate teacher does not necessarily make a better student [14]; specialising a small model towards multi-step reasoning trades away some of its general ability [15]; and Li et al. [16] find that models of a few billion parameters learn less from long, exploratory chains and from stronger teachers than from shorter chains, and propose mixing the two. Orca 2 [17] trains a small model to choose its own solution strategy rather than imitate the teacher's, in part by withholding from the student the detailed instructions that guided the teacher.

Our training pool is drawn from UGPhysics [3], the physics slice of WebInstruct-verified [18] and the advanced-physics split of Loong [10]. Different from the distillation work above, we compare the fine-tuned model with its base under a levelled serving stack—same budget, same tool declaration and contract, same audited grader—and report what that comparison shows (Section 5.3); and we measure three of the hazards above in controlled arms on a physics student: the teacher's instructions echoed into its traces, tool calls inside the traces, and distillation on problems matched to the hard set (Appendix C).

2026 S.T. Yau High School Science Award
仅用于2026丘成桐中学科学奖公示

2.3 Tool-Integrated Reasoning

Numerical physics is unforgiving of arithmetic, and a language model's arithmetic is unreliable. PAL [19] and Program of Thoughts [20] have the model write the calculation as code; ToRA [21] and MathCoder [22] train on interleaved language-and-code trajectories, and MAmmoTH [23] on a hybrid of language and program rationales, so that tool use becomes learned behaviour. Toolformer [24] lets the model decide where a call helps and keeps in its training data only the calls that improve its predictions; ReTool [25] learns the placement of code execution inside long reasoning by reinforcement learning rather than by imitation alone. Our sandbox is Python with SymPy [26], run on the CPU beside the model. Different from these approaches, which put the tool call into the training data, and from agent distillation [27], which argues that a small student needs the tool inside its traces, the model we deploy was not trained on tool trajectories at all: fine-tuned on tool-free traces from a plainly prompted teacher, it uses the sandbox once the tool is declared at serving time. We are not aware of prior work that adopts this as the recipe. It is not free. A long-chain reasoning model can suppress a latent tool ability [28]; a declared but unused tool can cost accuracy [29]; and what training must avoid, we find, is teaching the model *not* to call, or to think less before each call and lose its way in the tool loop: an adapter trained on a teacher's multi-turn tool trajectories learned the second habit, and one trained with the tool declared over traces that never call it learned the first (Section 5.4).

2.4 Reinforcement Learning with Verifiable Rewards

When a task has a checkable answer, the check can be the reward. DeepSeekMath [30] introduced GRPO, which normalises rewards within a group of samples for one prompt; DeepSeek-R1 [8] showed that such training can produce long-chain reasoning from a base model; General-Reasoner [18] extends verifiable-reward training beyond mathematics. Hochlehnert et al. [31] reran a set of recently reported reasoning gains under controlled seeds and decoding settings and found many of them inside the run-to-run variance of small test sets, and dependent on evaluation settings that differed between the model and its baseline. Different from these approaches, we treat reinforcement learning as a pilot rather than a result: a few updates on a small verified training set, under a verdict fixed before the run and read against a noise floor and a levelled baseline (Section 5.6).

2.5 Test-Time Search and the pass@k Envelope

Sampling a model several times and voting is the oldest form of test-time compute. Self-Consistency [32] takes the majority answer over sampled chains of thought; Large Language Monkeys [33] showed that coverage, the chance that at least one of k samples is correct, keeps rising with k past the point where voting saturates, and the gap between the two is the space a verifier could fill. Two studies read the pass@k envelope as a diagnostic: Yue et al. [34] find that reinforcement learning raises pass@1 but narrows the envelope; Chu et al. [35] contrast supervised fine-tuning, which tends to memorise, with reinforcement learning, which tends to generalise. Different from these approaches, we use the envelope as a diagnostic on physics at 9 B scale and judge the enhancement tier against a threshold fixed before the run (Section 5.5).

2.6 Deploying Small Models

A 9 B model in 16-bit weights needs about 19 GB before the first token of context, so single-GPU deployment is a question of weights first and cache second. LoRA [36] trains low-rank adapters instead of full weights; post-training quantisation such as GPTQ [37] and AWQ [38] is reported to compress weights to four bits with small accuracy loss; vLLM [39] pages the KV cache so that many sessions share one card; the base model we use, a Qwen3.5 model succeeding the Qwen3 line [40], keeps a KV cache on only a fraction of its layers. Different from these approaches, each of which optimises one layer of the stack, we cost the whole system on one 24 GB card and put the quantised deployment behind a pre-registered quality gate rather than reporting it as free (Section 5.7).

2026 S. T. Yau High School Science Award 项目2026区域赛中学科学论文公开

3 The Measurement Protocol and GewuBank-20K

The system is evaluated under a protocol fixed before the verdicts it governs were read, and against a bank we built. What the rest of the paper's numbers are worth depends on it. The bank is a contribution in its own right, built to give later work higher-quality data to train and evaluate on than the public sets we audited.

3.1 Measurement Protocol

Seven rules govern every measurement. Six are set out here; the seventh, an audited error rate before any use of a public benchmark, is stated in Section 3.6.

Pre-registration. Acceptance criteria are written before the result exists: for the distillation verdict, the enhancement tier, the reinforcement-learning pilot and its second run from the base, and the quantisation gate (Appendix A).

Levelled baselines. No gain is claimed until the grader, the generation budget, the prompt protocol and the tool availability of both arms are equal; the protocol difference alone can manufacture about 35 percentage points (pp) (+37.5 naive against +2.4 levelled, Section 5.3).

Mechanism bounds. Every result is screened against a back-of-envelope bound before it enters a table; a call rate and a per-call gain bound what a tool arm can add over a no-tool arm.

Positive controls. A filter, grader, cross-machine transfer or audit panel must first catch a known positive before its zero is believed.

One implementation per criterion. A decision rule is implemented once and imported everywhere.

One denominator. Arms are compared only on the same denominator, since dropping a grader's abstentions on one side and not the other shifts an arm's accuracy.

3.2 Two-Dimensional Difficulty Design

Every problem sits on two axes. The first, which we define for this bank, is the *computation type*: an A-type problem has an answer that closes in elementary functions; an N-type problem is one whose answer must be obtained numerically, a transcendental root or an integral with no closed form. The second is a *difficulty tier*, T1 to T3, set by the generation specification through structural complexity: steps, coupled quantities, and conditions implied rather than stated.

The A/N axis is the one a tool cares about. The distinction cannot be read off the surface. A 9 B model discovers that a problem is N-type only by deriving far enough to find that the integral will not simplify, which is exactly the moment a tool becomes useful, so N-type problems are the natural test of tool use. They are also what public training sets rarely contain. Of 30 problems drawn at random from the three public sources of our training pool, none needed a numerical method (24 close in elementary functions, 6 are conceptual or defective), a rule-of-three bound of about 10% on their N-type share, against about half of the bank.

Figure 2 shows what the axes measure. Across the bank the base model scores 74.6%; between the two types the gap is 25.8 pp, from the unrounded row means (87.3 against 61.6 as printed), and within a type the three tiers move accuracy by 14 to 17 pp. What makes a physics problem hard for this model is less how many steps it has than whether it can be finished in closed form.

3.3 Generation and Structural Gates

Problems were generated by GPT sol5.6 Ultra, used through its desktop application, from per-subtopic specifications over a grid of thirty physics domains. The raw archives are recorded in Appendix D. Generation is the cheap part.

Every item then passes a structural gate before it is graded: an answer-type tag, and one gradable answer per item, the key that the deterministic derivation of Section 3.4 reduces to a single value. An item that fails the gate is removed, not repaired (Appendix C records why the gate is mandatory). A duplicate screen runs beside the gate. A text screen on each problem's skeleton, numbers stripped,

2026 S.T. Yau High School Science Award 项目2026年度研究论文公开

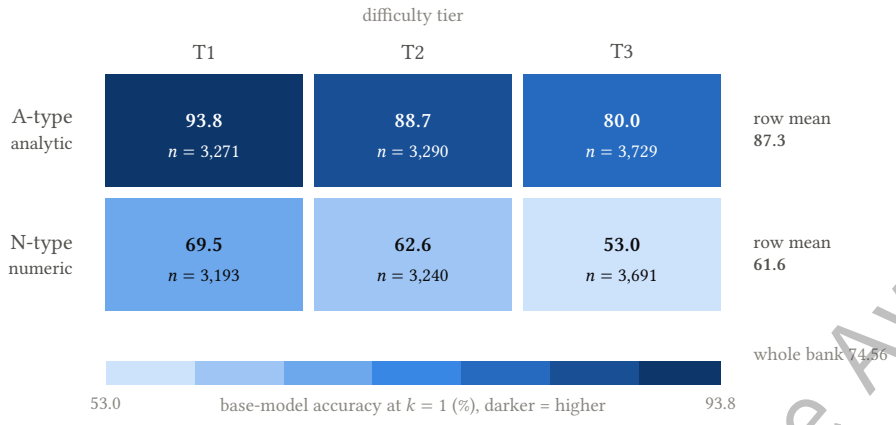


Figure 2. Difficulty runs along the A/N axis more than along the tier. Base-model accuracy, one answer per problem, on the bank (20,414 items); every cell shows its value and its n over all its items, unsettled items counted wrong; row means are over the items the grader could settle. The gap between the two row means is 25.8 pp on graded items (25.9 over all items); the span across the six cells is 40.8 pp.

with 10-gram overlap at intake found no repeats. A semantic screen over the whole bank with BGE-M3 embeddings put near-duplicates (cosine at least 0.95) at 0.8% of items, parameter variants of one template rather than the same problem reworded, so they stay as distinct problems. An item counts as part of the bank only once it has also received the base-model measurement behind Figure 2 and its key has survived the verification of Section 3.4, so that its difficulty is measured rather than declared and its key is checked rather than asserted. Generated rows that never received that measurement served only as distillation inputs and are not counted. The bank holds 20,414 problems.

3.4 Gold Verification

A model-generated key is a hypothesis about the answer, so every key goes through four stages of verification, and a fifth stage samples what they leave behind.

Structure and normalisation. Stage one is the structural gate of Section 3.3, which every item passes before it is graded. Stage two is key normalisation. The deterministic derivation, one implementation shared with every ruler, reduces each key to a single gradable value, and an audit removed what it could not (45 keys by the gold audit, and 464 that could not be reduced; Appendix D gives the intake counts).

Blind solve. Stage three is an independent blind solve by a reasoning teacher: DeepSeek-V4-pro with the sandbox during the distillation runs and DeepSeek-V4-flash in the re-check runs, each answer compared with the key by the symbolic grader and, where that grader abstained, by the judge (Appendix D records which release checked which items, and the paired comparison behind the change of release). The two are releases of one teacher family; each run's disagreement rate is read against controls checked by the same release, so a change of release cannot pass for key noise, and what no release of one family can catch, its shared blind spots, is left to the base model's reproduction and the independent re-solve of stages four and five.

Verdicts. Every key in the bank was put to the teacher. Over the 20,845 items that reached this stage the teacher reproduced 20,072 keys, declared 153 problems ill-posed, produced nothing gradable on 189 (a truncated solve or no boxed answer) and disagreed with the key on 431. The 431 contested keys are not in the bank. The 153 ill-posed verdicts and the 189 items without a gradable answer stay in the bank and are flagged per item, and the bank is the 20,414 items that remain. The re-check runs carried 246 control items whose key the base model had reproduced, stratified over the six grid cells; the teacher agreed with 92.3% of them, disagreed with 1.6% and returned no verdict on the rest, which bounds its false-alarm rate. Some of the removed keys were good, and the review under stage five measures how many.

2026 S. T. Yau High School Science Award 仅用于2026丘成桐中学科学奖论文公示

Second solver. Stage four is a second solver. The calibration measurement of Figure 2 is itself a blind solve by a different model, and the base model reproduces the key on 15,178 of the 20,414 items (74.4% with the 58 items the grader could not settle counted wrong; 74.6% over graded items, as in Figure 2), an endorsement independent of the teacher.

Residual error. Stage five calibrates the residual error. An independent blind re-solve of 60 items by a reasoning model found no key errors, which by the rule of three bounds the residual rate near 5%. A stratified sample of 120 of the bank's own contested and ill-posed verdicts, drawn in two halves of 60, was reviewed item by item by an independent reasoning model of a different family from the teacher (named in Appendix D), which recomputed with code the step on which the teacher and the key part: it found a real defect in 68% of the sample (a wrong key in 38 items, a statement that is ill-posed or depends on an unstated convention in 44), in 76% of the teacher's wrong verdicts, which is why those items are removed, and in 47% of its ill-posed verdicts, which is why those items stay, flagged; the two halves agree (65 and 72%). The author then reviewed the same 120 items independently and marked 101 as defective; the two reviews agree on 95 of the 120 items and differ almost only on whether an unstated convention makes a statement ill-posed. The count above is the lower of the two. No key was altered on the strength of a single verdict.

The verification also located the key noise. Contested keys concentrate in the problems the base model gets wrong. At intake the teacher contested 0.6% of the keys the base model had reproduced and 6.2% of those it had not.

3.5 Held-Out Rulers and Noise Floors

Isolation. The medium set is isolated by sub-topic, not by item. Items within one sub-topic share a physical setup, and an item-level split lets the setup leak from training into test. 87 of the bank's sub-topics are frozen and enter no training use of the bank, the reinforcement-learning training set included; the hard set is held out by item. The isolation function is applied by import wherever the bank is read, with a positive control confirming zero overlap with the hard set.

The medium set. 84 of the frozen sub-topics hold 1,495 items (the other three hold none), the medium held-out, on which the base model scores 74.8% under the calibration protocol. The medium set is a 253-item draw from it, balanced across the six cells of the grid (42 or 43 items per cell, all 30 subjects). It is the ruler for the fine-tuning comparisons, where a balanced grid matters more than size.

The hard set. The hard set comes from 60 further sub-topics, held out by item; they were specified almost entirely at the top tier. Of their 564 items, 51 were spent on calibration and the remaining 513 form the hard set: 288 analytic and 225 numeric items, 502 of them at tier 3, all 30 subjects. Under the calibration protocol the base model scores 51.8% on the 564 items of these sub-topics, 23 pp below the medium held-out.

The random set. This is a 500-item sample drawn from the bank outside every ruler, for the one comparison that places the two models on the bank at large (Section 5.2).

Noise floors. Each ruler carries a noise floor that travels with every number. For an accuracy p on n items,

$$\sigma(p,n) = \sqrt{\frac{p\,(1-p)}{n}}, \tag{1}$$

with the value for each of the four rulers in Table 2. Differences inside twice the floor (2σ) are reported as differences, not findings, and differences on fewer than thirty items are not interpreted.

3.6 External Benchmark Audit and Double Endorsement

No external key is used before it has been audited, under a rule fixed in advance: a content-error rate at or below 10% admits a set as it is; above 10% and at or below 20% admits it only through the double-endorsement construction below; above 20% excludes it. In a 40-item audit drawn across four public sources, 3 of the 7 ABench [2] items had wrong keys and a fourth was ill-posed, and

2026 S. T. Yau High School Science Award 仅用于2026年成师中学科学奖论文公示

the set was excluded from the rulers; its items reappear only in the paired set of Section 5.4, which measures a within-item difference rather than a level. PHYBench [5] showed no errors in the seven items audited, too few to place it in a band (the bound is about 35%); nothing rests on it. A later audit of 100 items per source, solved blind by a teacher with the sandbox and counting the final answer only, put the share of correct keys at an estimated 78% for ABench and 84, 86 and 91% for the three sources of the training traces (UGPhysics, WebInstruct, Loong); traces enter training only where the teacher reproduced the dataset's key (Section 4.2). PhysElite [1], the olympiad-level set that serves as the external anchor, sits in the middle band. A 60-item stratified audit found content errors in 13.3% (±4.4) of items and form defects in another 13.3%; a census of the whole set found LaTeX escape damage in 49.1% of items (repairable deterministically), duplicates in 20.3% and missing figures in 4.2%.

PhysElite is therefore used in two ways that keep the audit visible.

The endorsed set. Of 11,586 items, 1,860 are self-contained as text, and 1,307 of those pose a single gradable question. Four of these share eight or more character 20-grams with a problem in one of our training pools and were set aside before any sampling (34 hits across the whole benchmark; 20,797 training-pool statements screened, a planted positive control caught). A near-duplicate screen and the removal of keys the audit had found wrong or that bundled several sub-answers left a pool of 1,032. The production teacher, DeepSeek-V4-flash, attempted every item of the pool blind, together with the 60 already audited, 1,092 attempts in all. The 708 on which its answer independently reproduced the key form the endorsed set (374 Chinese, 334 English). The 357 on which its answer was gradable and wrong are kept as the endorsed set's complement and reported beside it, so that the selection can be removed (Section 5.2). The remaining 27 attempts ended without a gradable answer and are in neither.

What the endorsed set is not. The benchmark's duplication still reaches the endorsed set. Four of the 708 are copies of one experimental problem, so three of them are redundant, and three more carry keys that refer to text outside the problem. The endorsed set holds 702 distinct usable items. Five of the six were answered wrong by every arm and the sixth was credited to one arm by the judge. Removing all six leaves the paired tests of Section 5.2 unchanged. By the benchmark's own labels the endorsed set is easier than the whole set (mean 3.5 against 4.1 on its 1–7 scale, top three labels on 21% of labelled items against 41%), so it is not a sample of the whole set and is not used as one. It carries an honesty clause wherever it is reported. Results are conditioned on the problem being solvable by the endorsing teacher. Neither the endorsing teacher nor the judge appears as a comparison row. Verdicts come from one judge, DeepSeek-V4-pro, on nearly every item of each arm, since the deterministic checker settles few symbolic answers, where the original protocol used three judges. And the problem statements are the original text, with only the keys escape-repaired for grading.

The ranking set. The second use is an unfiltered sample: 500 items drawn at random from the whole set (matching its difficulty distribution; 28 of them are also in the endorsed set), answered as text only from the original statements, with the handicap the original paper imposed on its text-only row—73% of items carry a figure the model cannot see. That row is for ranking; the endorsed set is for matched-scale comparison.

One set is conspicuously absent. UGPhysics [3], the largest source of the fine-tuning traces, shares its source pool with PHYSICS [4]: a three-criterion check found 622 overlapping items, 11.3% of the pool, across all thirteen subjects. Neither could serve as an external anchor, and the overlap was removed from the training side; no ruler in this paper is drawn from PHYSICS-test, which enters only through that same paired set.

2026 Shanghai High School Science Award
仅用于2026上海高中科学奖论文公示

4 The Gewu System

4.1 Overview

Gewu is one 8-bit weight file, a 173 MB adapter attached or detached per request, a Python sandbox on the CPU beside the model, and two serving paths over them (Figure 3).

A question takes one of two paths. On the default path it is wrapped in the generic contract, the adapter is enabled, and the model answers in one pass, calling the sandbox when it needs to. On the enhancement path the adapter is disabled, eight samples are drawn from the base model with the sandbox available, and a blind symbolic vote reconciles them. The default path's answer is shown beside the result and is not counted. The paths share every component except the adapter flag, the contract and the sampler. The enhancement tier is offered on request, for numeric problems of the bank's kind; its contract was chosen on the hard set (Section 5.5).

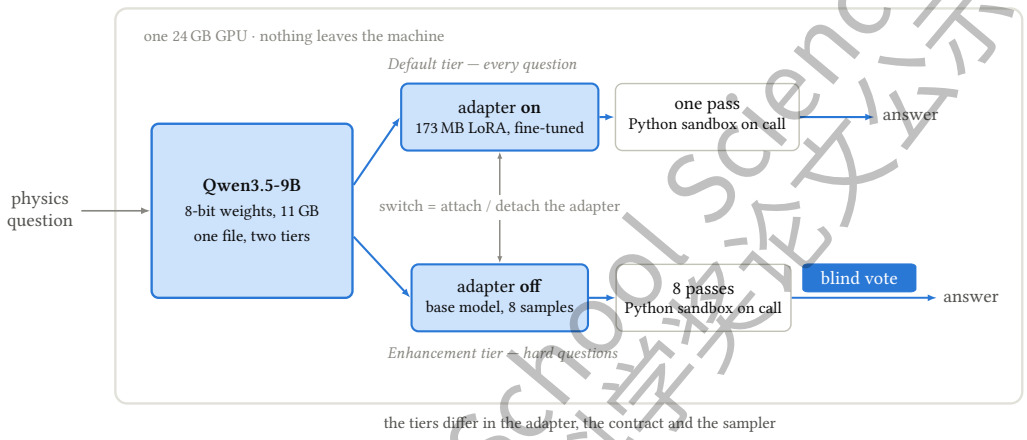


Figure 3. The Gewu system: two serving tiers, one set of weights. The fine-tuned model is the base model plus a 173 MB LoRA adapter, so the tiers share one weight file and differ in the adapter, the contract and the sampler. Default tier: adapter on, a single pass—reasoning in Chinese, ahead of the base on an external olympiad set and level with it on ordinary questions. Enhancement tier: adapter off, eight samples of the base model reconciled by a blind symbolic vote (Section 5.5). Runs on a single 24 GB GPU (measured on an RTX 3090) with no external API on any serving path; both weight profiles the tiers use, the base and the fine-tuned model in 8 bits, passed the quantisation gate; the vote was measured in 16 and in 8 bits (Section 5.7).

The three requirements of Section 1.1 fall to the two paths and the stack beneath them. Speed: the default path answers everyday questions in one pass in a third to a half of the base model's tokens. Accuracy: the enhancement path spends eight base-model samples and a vote on the problems the default path cannot settle, and pays for them in time. Locality: everything, sandbox and vote included, runs on the one 24 GB card (Section 4.5).

4.2 Base Model and Adapter

The base model is Qwen3.5-9B [41], a successor of the Qwen3 line [40]. Three of its properties matter on one card: its hybrid attention keeps a KV cache on eight of thirty-two layers; its 8-bit weights fit in 11 GB; and it is natively multimodal—the same weights read images and video. The third property is exercised only in the paired measurements on the external ranking set; both tiers and every other number in the paper are text-only. It is served as one W8A16 weight file that both tiers share.

The fine-tuned model, Gewu-SFT, is that base plus a LoRA adapter [36] of rank 32 ($\alpha = 64$) over twelve target modules, the linear-attention projections included, 173 MB on disk. Each adapted matrix W is served as

$$W' = W + \frac{\alpha}{r} BA, \qquad B \in \mathbb{R}^{d_{\text{out}} \times r},\ A \in \mathbb{R}^{r \times d_{\text{in}}},\ r = 32, \tag{2}$$

2026 S.-T. Yau High School Science Award 仅用于2026丘成桐科学奖论文公示

so the adapter file holds only the two low-rank factors. The adapter is enabled per request: on, the model is the default tier; off, the same weights are the base model of the enhancement tier. The adapter that ships is Gewu-SFT. Why a reinforcement-learning update does not replace it is set out in Section 5.7.

It was trained for three epochs, 936 steps, on 4,983 reasoning traces from one recipe. DeepSeek-V4-pro, in a preview release since withdrawn, solved each problem blind, without the key and without tools, and a trace was kept only if its final answer agreed with the dataset's key under the grader (one retry on failure; first-pass hit 93.5%). The problems came from UGPhysics [3] (2,620), the physics slice of WebInstruct-verified [18] (2,365) and the advanced-physics split of Loong [10] (263); training used 4,983 of these 5,248. The preview release has been withdrawn, so the recipe can be described but not re-run. Later teacher releases differ from it in language, trace length and backtracking rate, a caveat that applies to distillation in general (Appendix D records which release produced each dataset).

Two properties of the recipe are deliberate. No trace contains a tool call, and later measurements support the choice. Adapters trained on multi-turn tool trajectories, the teacher solving with the sandbox in the loop, scored below this recipe on the medium set and learned to think less before each call and to lose their way in the tool loop (Table 7 in Section 5.3), so the tool is declared at serving time instead, and whatever tool behaviour the fine-tuned model shows there was not taught (Section 5.4). And the traces carry the teacher's own instructions and no others, the invariant Section 4.3 states with its gate.

4.3 Tool Loop and Serving Contract

The tool is a Python sandbox (SymPy [26], NumPy, SciPy) on the CPU beside the model. When the model emits a code block in the declared format, the loop executes it in a sandboxed process, returns what it printed into the context, and lets the model continue until it produces a final answer. The loop allows at most twelve tool turns and then asks for a final answer. The sandbox covers what makes an N-type problem N-type: symbolic simplification, equation solving and numeric evaluation.

The contract is the part of the prompt that declares the tool and says how it is to be used. It is a serving-time object. The adapter's training data contains no tool, and the contract is what switches tool use on. Three contracts appear in the results, each a different text, and every number in the paper names the one it was measured under (Table 1; full texts in Appendix B).

The minimal contract is an empty system prompt with a 30-character suffix asking for a boxed answer. The generic contract, the default tier's, is a 1,853-character system prompt, the tool declaration in XML and a physicist's working instructions (settle the physics before computing, derive in full, one numerical answer in one box, declare an ill-posed problem rather than guess), with a 208-character suffix. The domain contract keeps the declaration but replaces the working instructions with a stopping rule (after the number, one dimensional and order-of-magnitude check, then stop; no recomputation, no second method) and moves the answer rules into a 353-character suffix (three to four significant figures, several quantities in one box, an ugly result is not a defect). It was written for the bank's own problems and is not served to open-ended questions.

One invariant governs how the contract meets training: the student sees at training time every instruction the teacher saw, with the contract as the one permitted exception, and a batch whose traces contain no call carries no tool declaration. A machine gate with positive controls in both directions fails the build on a violation, because a declaration over demonstrations that ignore it teaches the model not to call. Figure 4 shows the loop once, end to end.

2026 S. Liu High School Science Award
仅用于2026年度中学生科学论文公示

Table 1. The three serving contracts. Each is a different text; the two longer ones share the tool declaration. Every accuracy in Section 5 names the contract it was measured under.

Contract	Text
minimal	empty system prompt; 30-character suffix (answer in a box)
generic	1,853-character system prompt (tool declaration in XML; physicist's working instructions; one boxed number; ill-posed exit); 208-character suffix; served by the default tier
domain	1,656-character system prompt (same tool declaration; check once, then stop, no recomputation); 353-character suffix (significant figures, one box, ill-posed exit); used by the enhancement tier

ONE ITEM, END TO END: `hg3f:geometrical_optics:f17:T3:A:005` (HARD SET, TIER 3, ANALYTIC ANSWER)

Problem (verbatim; gold `35.2646614885`, in μm):

空气中两枚薄透镜沿轴相隔 $d_0 = 34.0\,\text{mm}$，初温焦距分别为 $f_{10} = 82.0\,\text{mm}$、$f_{20} = 135.0\,\text{mm}$。它们的焦距温度系数满足 $(1/f_j)df_j/dT = \kappa_j$，其中 $\kappa_1 = 24.0\times10^{-6}\,\text{K}^{-1}$、$\kappa_2 = -11.0\times10^{-6}\,\text{K}^{-1}$；间隔框架线膨胀系数 $\alpha_b = 19.0\times10^{-6}\,\text{K}^{-1}$。升温 $42.0\,\text{K}$ 后求系统有效焦距变化，正号表示增长，单位为 μm。

① **Reasoning** — base model under the domain contract; 25,438 characters, excerpt:

> I need to calculate the change in effective focal length (Δf) when the temperature increases by $\Delta T = 42.0\,\text{K}$. [...] Wait, the relation is $\frac{1}{f}\frac{df}{dT} = \kappa$. Integrating gives $\ln(f) = \kappa T + C$, so $f(T) = f_0 e^{\kappa \Delta T}$. [...] Let's write a python script to perform these calculations precisely.

② **Tool call** — one call to the sandbox; the code, excerpt (protocol wrapper omitted):

```
import numpy as np
d0, f10, f20 = 34.0, 82.0, 135.0          # mm
kappa1, kappa2, alpha_b = 24.0e-6, -11.0e-6, 19.0e-6   # K^-1
delta_T = 42.0                            # K
[...]
inv_f_eff0 = inv_f10 + inv_f20 - d0 / (f10 * f20)
f_eff0 = 1.0 / inv_f_eff0
[...]
f1_new = f10 * np.exp(kappa1 * delta_T)
f2_new = f20 * np.exp(kappa2 * delta_T)
d_new = d0 * np.exp(alpha_b * delta_T)
[...]
delta_f_um = delta_f_mm * 1000.0
print(f"delta_f_um = {delta_f_um:.6f} um")
```

③ **Return and self-check** — the model re-derives the returned number by hand, excerpt:

> The calculation yields $\Delta f \approx 35.29\,\mu$m. Let's double check the logic. [...] $F_0 \approx 1/0.016531 \approx 60.49$ mm. Matches code output.

Final answer, as graded: $35.3\,\mu$m — gold $35.26\,\mu$m — **correct**.

Control — the same item answered by the echo arm, the distillation arm trained with the domain contract inside its training prompt: an empty reasoning field, tool calls 0, **wrong**. Its answer text begins, verbatim, by reciting the contract's suffix:

```
final answer in μm, 3-4 sig figs. Let's compute accurately.
Let's do precise arithmetic manually.
u=1/82 = 0.012195121951219512195121951219512195121951
[...]
```

2026 S.T. Yau High School Science Award 仅用于2026丘成桐中学科学奖公示

Figure 4. One run of the tool loop, verbatim. One hard-set item as an illustration of the mechanism: the model derives the formula, delegates the arithmetic to the sandbox, checks the returned number by hand, and answers correctly; the control arm, trained with the contract inside its prompt, produces no reasoning, no call and a wrong answer. Observational and illustrative only; the causal evidence is the ablation table, Table 7. Text reproduced from the run logs; [. . .] marks omitted passages.

4.4 Test-Time Search

The enhancement tier samples the base model eight times at temperature 0.6, top-p 0.95, with the adapter disabled and the sandbox available to every sample, and reconciles the answers by a vote that never sees the key. Final answers are clustered by the symbolic part of the grader's test—symbolic simplification of their difference, or numeric agreement within tolerance, falling back to string match—so the vote needs no key. With $a_1, \dots, a_k$ the final answers and $\equiv$ that test, the served answer is the representative of the heaviest cluster,

$$\hat{a} \;=\; \arg\max_{c \in \mathscr{C}} \sum_{i \le k} \mathbf{1}[a_i \equiv c], \qquad k = 8, \tag{3}$$

where $\mathscr{C}$ is the set of clusters $\equiv$ induces on the eight base samples; a sample without a final answer abstains, and a tie goes to the cluster whose first sample was drawn earlier. The contract is the one the request was served under.

Two choices follow from measurements. The sampler is the base model, not the adapter. At the same temperature the adapter's samples stay close to its greedy answer while the base model's climb above its own (Section 5.5), and the reinforcement-learning update from the base model is not a better sampler either (Section 5.6).

The fine-tuned model's answer is shown beside the vote and not counted in it. Counted as one vote, or as four samples at half weight, it moves the result by less than a point; at weight two or more it costs accuracy (Section 5.5).

Every SC@k number (a vote over k samples) of the enhancement tier states its vote composition and is reported beside a no-op control, one sample chosen at random from the same eight, so that the vote's gain is measured against sampling itself.

4.5 Serving Stack on One 24 GB GPU

Why this model fits. Qwen3.5-9B's thirty-two layers are twenty-four linear-attention layers and eight full-attention layers, and only the eight keep a KV cache. With L_{full} such layers, n_{kv} key-value heads of width d_{head} and 16-bit cache entries, a session costs

$$\text{bytes per token} \;=\; \underbrace{2}_{K,V} \times L_{\text{full}} \times n_{\text{kv}} \times d_{\text{head}} \times 2\,\text{B}, \tag{4}$$

32 KB per token for this model: 0.25 GB at 8k context and 1.0 GB at 32k, a quarter of a dense model of the same size. So the memory of a deployment is dominated by the weights, the lever is weight quantisation, and concurrency is nearly free, which is what makes the enhancement tier affordable on the same card.

Weight profiles and the gate. The 16-bit file is 19.3 GB and leaves a 24 GB card under 5 GB before the runtime's own workspace and any KV cache, too little to serve. The 8-bit profile, GPTQ [37] int8 at group size 128, calibrated on 384 of the bank's own problems, measures 11 GB; the 4-bit profile measures 7.6 GB. Every profile has to pass the quantisation gate of Appendix A before it is served (Section 5.7). One trap is specific to this model: its vocabulary has 248,320 entries and the embedding and output matrices are untied, 3.8 GB in 16 bits, and the 4-bit profile keeps both in 16 bits.

2026 S. T. Yau High School Science Award 应用于2026丘成桐中学科学奖论文公示

Stack and budget. vLLM [39] serves the quantised weights with the adapter enabled per request;[1] the tool-loop middleware runs the sandbox on the CPU at zero GPU memory; the voter clusters answers by symbolic equivalence with the judge switchable off. Nothing on the path makes a network call. With the 11 GB weights and the adapter loaded, the KV pool left on a 24 GB card holds tens of sessions at 8k context, and the enhancement tier's eight samples run concurrently inside it (measured in Section 5.7). Configurations below 24 GB have not passed a gate and are not claimed.

5 Experiments

The experiments follow the contributions of Section 1.4. Section 5.2 reports the two served tiers on the rulers and on PhysElite. Sections 5.3 and 5.4 carry the central finding: what fine-tuning changed, and where accuracy is added at answer time. Section 5.5 measures the vote the enhancement tier serves. Sections 5.6 and 5.7 test the finished system from two further sides, a reinforcement-learning pilot and the deployment gate, and change nothing in it.

5.1 Experimental Setup

Names. Arms and evaluation sets are named in this paper by what they are. The names differ from those in our logs and code; nothing else, no number, item or protocol, was changed.
Evaluation sets. Four rulers (Table 2): two internal, drawn from GewuBank-20K under the isolation of Section 3.5, a 253-item medium set and a 513-item hard set (288 analytic, 225 numeric); two external, drawn from PhysElite under the audit of Section 3.6, the 708-item endorsed set and an unfiltered 500-item ranking set. Further sets appear where noted: the endorsed set's complement, which with the endorsed set makes the whole clean pool of 1,065 items; 200 figure-bearing items of the ranking set; a 1,436-item paired set of public problems, ABench 400, PHYSICS 764, OlympiadBench 172 and PHYBench 100, assembled before the audit rule of Section 3.6 was adopted; and a 500-item random set drawn from the bank outside every ruler.

Table 2. The four rulers. The medium set is held out from the bank by sub-topic, the hard set by item (Section 3.5). Every accuracy is reported with the 1σ binomial floor of its ruler; differences inside 2σ are reported as differences, not findings.

Ruler	n	Source	Used for	1σ (pp)
medium set	253	GewuBank, held out by sub-topic	fine-tuning comparisons	2.7
hard set	513	GewuBank, held out by item	behaviour, search, RL, quantisation	2.2
endorsed set	708	PhysElite, double-endorsed	external comparison at matched scale	1.9
ranking set	500	PhysElite, unfiltered random	ranking against quoted rows	1.8

Models and arms. The base model is Qwen3.5-9B; the fine-tuned model, Gewu-SFT, is the base plus the LoRA adapter of Section 4.2. Every arm except the minimal-contract control has the same Python sandbox, declared through one of the three contracts of Section 4.3 and stated with every number. Search arms sample at temperature 0.6, top-p 0.95 (the model's recommended sampling setting for thinking mode), and vote by blind symbolic clustering (Section 4.4).
Grader. A deterministic numeric-equivalence checker; what it does not accept goes to a model judge.

[1] Two rules of operation came out of the measurements. The adapter's parameter names must follow the layout the server resolves for the weight file, which differs between the 16-bit and the quantised file; a mismatch attaches nothing and raises no error, so every server start is followed by a probe that runs one prompt with the adapter on and off and refuses to serve if the outputs are identical. And every accuracy of an adapter-bearing model in this paper was measured on merged weights; the mounted form serves only the default tier, where it was measured level with merged weights (-1.4 pp, paired $p = 0.60$) for Gewu-SFT.

2026 S.T. Yau High School Science Award 仅用于2026丘成桐中学科学奖公示

Metrics. Accuracy at $k = 1$; the pass@k envelope; tool-call rate, termination and truncation. For k sampled answers $a_1, \dots, a_k$,

$$\text{pass@}k \;=\; \Pr\big[\exists\, i \le k \;:\; a_i \text{ is correct}\big], \tag{5}$$

averaged over the ruler. Paired comparisons use McNemar's test on the discordant items: with b_{10} the items only the first arm gets right and b_{01} the items only the second does,

$$p \;=\; 2\,\Pr\big[X \le \min(b_{10}, b_{01})\big], \qquad X \sim \text{Binomial}\big(b_{10} + b_{01},\, \tfrac{1}{2}\big), \tag{6}$$

exact when the discordant count is small and by the continuity-corrected χ^2 approximation otherwise.

Pre-registered criteria. Five verdicts were fixed before their results existed (Appendix A). Distillation line: keep distilling only if the corrected arm beats the base by at least 3 pp on the medium set. Search: at least +5 pp over the strongest single arm on the hard set; +2 to 5 mild; under +2 a negative result. Reinforcement learning: ±3 pp around the starting policy's hard-set score under the generic contract, and the same rule for the second run from the base under the minimal contract. Quantisation: at most 3 pp lost, paired $p > 0.05$, no drift in call or termination rate. Arms are compared only on the same denominator; items on which the grader abstains count as wrong on every arm.

5.2 Results on the Rulers and on PhysElite

Endorsed set. On the endorsed set (Table 3) the fine-tuned model scores 62.1% against 54.5% for its base under the same contract, +7.6 pp (paired $p < 0.001$, Eq. (6); 133 items it alone gets right against 79), under a protocol that favours the base: the better for the base of the two contracts that declare the tool, the same sandbox, and a budget the base never exhausts (protocol termination 100%). Two same-size models on the same harness, a vision-language model answering as text and a reasoning distillation of the previous generation, score 27.4 and 17.2%. The enhancement tier was not run on PhysElite, so the external numbers contain no test-time search. The gap between the endorsed set (54.5) and the ranking set (19.8) measures what the double endorsement selects: problems the teacher can solve, text-self-contained by construction, and mostly below the set's top difficulty labels; the endorsed set's honesty clause applies to every number reported on it.

Whole clean pool. The endorsed set's keys are trustworthy, but its items were selected by a model of the same family as the fine-tuned model's teacher, a selection that could favour the fine-tuned model. Removing the selection removes little of the margin: on the whole clean pool of 1,065 items, the 708 plus the 357 on which the teacher's answer was gradable and wrong, keys as shipped, the fine-tuned model is ahead by 6.7 pp (47.0 against 40.4%; 167 items against 96, $p < 10^{-4}$; Table 3).

2026 S. T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

Table 3. PhysElite: the endorsed set (708 double-endorsed items, $1\sigma = 1.9$ pp) and the whole clean pool (1,065 items, $1\sigma = 1.5$ pp). All rows run by us on the same harness and grader, greedy decoding; every row has the Python sandbox declared except the minimal-contract rows, whose prompt declares no tool. The honesty clause applies to the endorsed set (Section 3.6). The domain-contract pair is given in Section 5.4. Differences are computed from item counts and can differ from the difference of the rounded percentages by 0.1.

Model	Contract	Set	n	Acc.	Note
Gewu-SFT	minimal	endorsed set	708	**70.2**	+6.2 vs base, $p = 0.001$
base	minimal	endorsed set	708	64.0	
Gewu-SFT	generic	endorsed set	708	**62.1**	+7.6 vs base, $p < 0.001$; calls on 3%
base	generic	endorsed set	708	54.5	calls on 9%
Gewu-SFT	generic	whole clean pool	1,065	**47.0**	+6.7 vs base, $p < 10^{-4}$
base	generic	whole clean pool	1,065	40.4	
Qwen3-VL-8B-Instruct	generic	endorsed set	708	27.4	−27.1 vs base
DeepSeek-R1-Distill-Qwen-7B	generic	endorsed set	708	17.2	−37.3 vs base; never calls

Internal rulers. Table 4 collects the two served tiers against their baselines on the internal rulers. On the medium set the fine-tuned model and the base with the same tools and contract are level in accuracy and far apart in cost: 75.5 against 73.1% on the 253 items (+2.4, not significant) at less than half the tokens (Table 6); on the random set, 500 bank problems drawn outside every ruler, they are level again (77.8 against 78.2; 53 items against 55, paired $p = 0.92$). On the hard set the base with tools scores 58.5% under the domain contract; the enhancement tier, eight base samples reconciled by the vote of Eq. (3), scores 74.1%, +15.6 pp, against a no-op control of 59.5 (Section 5.5).

Table 4. The two served tiers on the internal rulers, each against its own baseline, and three open models of a similar size (7 to 12 B) on the hard set. Every row has the Python sandbox declared; contracts as in Table 1; the vote is eight base samples at temperature 0.6 (Eq. (3)). The last block shares one harness and the generic contract, and its base row is the reference for the three models (a separate run of the same configuration gives 54.4, Section 5.4). 1σ: 2.7 pp on the medium set, 2.2 on the hard set. p: exact McNemar test against the block's base row on the same items; in the last block Fisher's exact test on the counts.

Model	Contract	Decoding	Acc.	p vs base
Medium set (253 items)				
base	generic	greedy	73.1	
Gewu-SFT (**default tier**)	generic	greedy	**75.5**	0.51
Random set (500 bank items outside every ruler)				
base	generic	greedy	78.2	
Gewu-SFT (**default tier**)	generic	greedy	77.8	0.92
Hard set (513 items)				
base	domain	greedy	58.5	
base (**enhancement tier**)	domain	SC@8 (blind vote)	**74.1**	3×10^{-13}
Hard set, generic contract: the base and three models of a similar size				
base (9 B)	generic	greedy	57.1	
Gemma-4-12B-it (12 B)	generic	greedy	30.2	$< 10^{-15}$
Qwen3-VL-8B-Instruct (8 B)	generic	greedy	16.0	$< 10^{-15}$
DeepSeek-R1-Distill-Qwen-7B (7 B)	generic	greedy	9.6	$< 10^{-15}$

2026 S.T. Yau High School Science Award
仅用于2026丘成桐中学科学奖公示

Models of a similar size. The hard set is not hard only for our model: three open models of a similar size, run on it under the same harness and contract as a base reference of 57.1 (the last block of Table 4), score between 9.6 and 30.2%, and 4 to 8% on its numeric half. Two of them barely call the sandbox; Gemma calls it on 31% of problems and is right on a fifth of those, so the tool alone does not close the gap. For the two models also run on the endorsed set the distance from the base is of the same order there.

Ranking set. On the unfiltered ranking set (Figure 5), 500 items at random answered as text only, 73% of them referring to a figure the model cannot see, the base model with the tool scores 19.8% (±1.8), with five items the grader could not settle and three whose shipped key is not gradable (an essay answer, an empty key, a proof) counted wrong. That is level with the benchmark's two 20.0 rows and level with or slightly above its 671 B text-only row (+2.8 pp, 1.6σ). The fine-tuned model on the same 500 items, same contract and grader, scores 23.6% (118 of 500), level with the benchmark's 24.0 row and a difference of 3.6 pp from its 20.0 rows (1.9σ). The frontier reference we ran ourselves on the same 500 items under the same grader, DeepSeek-V4-pro with a code tool, scores 41.8% (209 of 500; the judge, itself V4-pro, decided 432 of the 500 verdicts); the 9 B base model with the tool reaches 47% of that with no network call. The ranking set selects nothing, but its keys are the shipped ones, repaired for formatting and not verified, and it is reported for the ranking of Figure 5 and for nothing else.

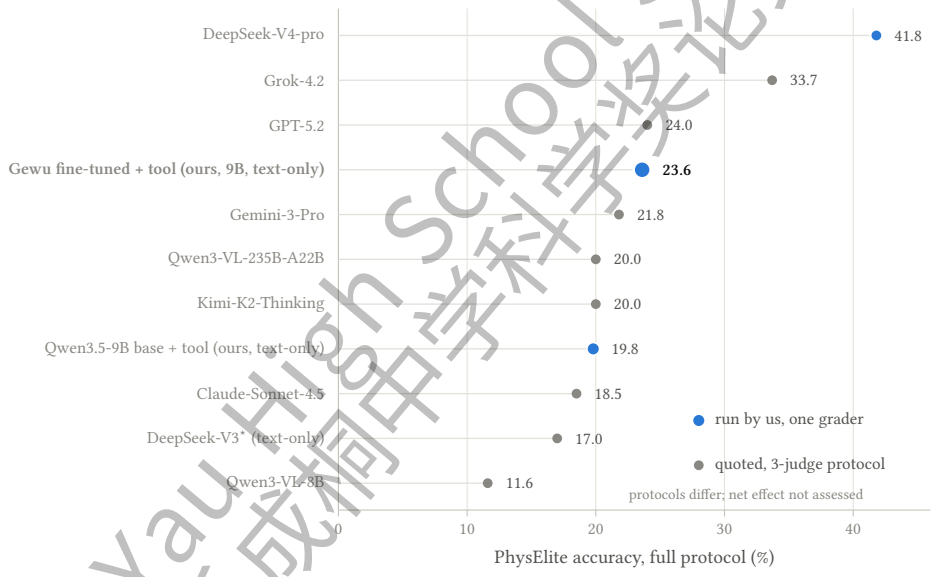


Figure 5. A 9 B model run locally, text only: its base level with the benchmark's 20.0 rows, its fine-tuned model with the 24.0 row. Full protocol, $n = 500$, $1\sigma = 1.8$ pp. Blue rows were run by us under one grader, the two Gewu rows under one harness; grey rows are quoted from the original paper, which scored with three judges against the unrepaired keys where we score with one judge against repaired keys, two differences whose net direction we do not claim. The double-endorsement conditions do not apply to the ranking set. An asterisk on a quoted row is the benchmark's own mark for a text-only entry.

With the figure. The base model's weights read images, so we measured whether showing the figure helps, on 200 of the ranking set's figure-bearing items drawn at random, each answered with and without its figure under the same per-item seed at temperature 0.6, a tool-free prompt and a 16k-token budget, undecided items counted wrong (Table 5). For the base model the figure changes nothing that clears noise, and its figure arm runs out of budget more often. For the fine-tuned model, one sample per arm under the same seeds, the figure is worth +6.5 pp. At this budget the figure's gain clears noise for the model fine-tuned on text alone and not for the base's vote; the ranking set is reported text-only, and the conclusion returns to what closing the gap would take.

2026 Shi-Tsung Yau High School Science Award
仅用于2026石成金中学科学论文公示

Table 5. Showing the figure: 200 figure-bearing ranking-set items, same per-item seeds in both arms, undecided items counted wrong; an item on which none of the eight base samples finishes counts wrong for the vote.

Model	Measure	no figure	figure	Δ (items; p)
base, 8 samples	eight-sample vote	20.0	24.5	+4.5 (12 vs 21; $p = 0.16$)
	none of its eight finished	9.5%	30.5%	
fine-tuned, 1 sample	accuracy	14.5	21.0	+6.5 (8 vs 21; $p = 0.024$)

5.3 What Fine-Tuning Buys

What fine-tuning bought. Under the levelled protocol of Section 3.1, the same grader, budget, tool and contract on both sides, fine-tuning bought four things, in the order the evidence supports them.
A disciplined procedure. On the hard set the fine-tuned model reasons in Chinese on 92.8% of problems and the base on none (the traces are 84% Chinese). With no tool in its training data, once the tool is declared at serving time, it calls the sandbox on 54.2% of medium-set problems, with at least one successful call on 99.3% of them; it calls selectively, on 2.7% of the endorsed set against 8.9% for the base.
Token economy. It reaches its answer in roughly a third to a half of the tokens: the median sequence is 4,949 tokens against 10,856 on the medium set, 2,772 against 7,913 on the endorsed set, 4,926 against 9,623 on the ranking set and 4,556 against 10,774 on the random set, at level or higher accuracy on each; for a tutor that streams its answer this is the difference between about a minute and about three.

Table 6. Tokens per answer. Median sequence length (prompt included) of the base and the fine-tuned model, each pair under the same contract, tool, budget and greedy decoding; the accuracies of these pairs are in Table 4, Table 3 and Figure 5.

Ruler	base, tokens	fine-tuned, tokens
medium set (253)	10,856	4,949
endorsed set (708)	7,913	2,772
ranking set (500)	9,623	4,926
random set (500)	10,774	4,556

Robustness to the prompt. Between its serving contract and the minimal one the fine-tuned model moves less than the base, which moves 22.6 pp (58.5 to 35.9), and across the three contracts of the endorsed set it spans 33.9 pp against the base's 36.8 (the domain end of both on the 397 items that contract was run on); under the minimal contract the two are level (38.0 against 35.9, $p = 0.32$) and the fine-tuned model still reasons in Chinese on 94.7% of problems, which we read as the structure the prompt used to supply travelling in the weights.
Accuracy: level at home, ahead abroad. On the medium set the fine-tuned model is +2.4 pp over the base with the same tools and contract, inside the noise floor ($p = 0.51$), and on the random set −0.4 ($p = 0.92$; Table 4); on the endorsed set it is +7.6 pp ahead, and on the whole clean pool, without the teacher's selection, +6.7 (Table 3). Section 6.1 reads the gain and the behaviour together.
What a naive comparison reports. Without that levelling the same comparison, the grader as shipped, unmatched budgets (the base capped at 4,096/8,192 tokens), the base without the tools and contract the fine-tuned model was given, reports +37.5 pp on the medium set. Figure 6 imposes the three fairness requirements in turn, each of which moves the protocol towards the base: verifying the grader on both sides, which lifts the base more than the fine-tuned model, leaves +34.8; giving the base the generation budget it needs, under the same sampling, leaves +16.6 ($p = 1.4 \times 10^{-6}$); giving the base the same tools and contract leaves the +2.4 above. Every correction went the base

2026 S. T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

model's way, and the fine-tuned model remains ahead on the medium set, inside the noise floor, at less than half the tokens (Table 6). Each share removed is still a difference between the two models, of behaviour (answer format, budget) or of system (tool and contract), and the last bar measures physical reasoning itself, with format, budget, tool and contract held equal.

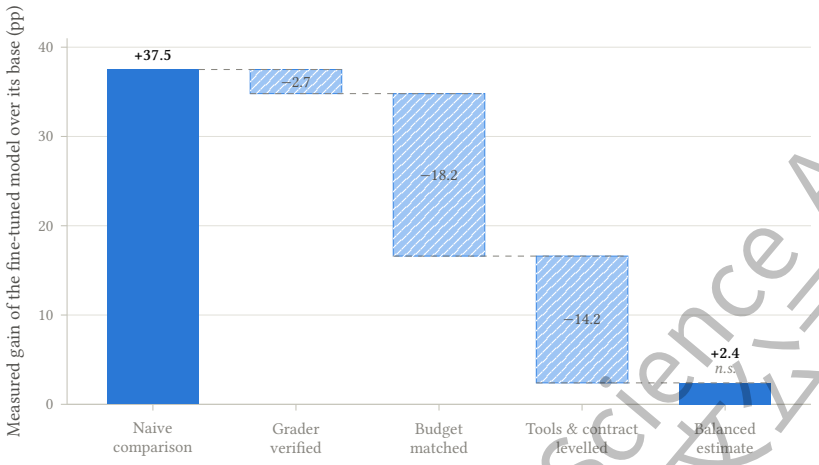


Figure 6. What a naive comparison reports, and what a balanced one reports. Hatched blocks are what each fairness requirement removes: a grader verified on both sides, a matched generation budget, and a base model given the same tools and contract. The surviving +2.4 pp is not significant ($p = 0.51$); on the endorsed set the same model is +7.6 pp above the base (Table 3). Nothing about the model differs between the first bar and the last.

Why abroad and not at home. The pattern is not a property of one set's keys. The endorsed set and the whole clean pool, with and without the teacher's selection, agree (+7.6, +6.7); the medium set and the random set agree with one another (+2.4, −0.4); and the base model's level is anchored on both sides, at 19.8 against the benchmark's 20.0 rows and at 74.6% on a bank whose contested keys were removed. Nor is it the tool: with no tool call in its training data the fine-tuned model calls the sandbox and completes the protocol as reported above. Our reading is that it is a property of distributions. Nineteen in twenty of the adapter's traces come from UGPhysics and WebInstruct, problems written for students and collected from textbooks, courses and instructional pages; PhysElite was collected the same way, from the textbooks, past competition papers and training handouts of contest students, transcribed and corrected by hand [1]. The two are similar in distribution. A cross-lingual semantic screen finds about 2% of the endorsed set shared with the training pool, and removing those items leaves the gain at +7.5 pp. The bank is a different distribution, problems generated to a specification by a language model and calibrated on the base model, and no training source is drawn that way. The gain appears where the problem distribution is close to the traces', and rulers that separate +15.6 and −25.2 report level when it is level.

Ablations. Gewu-SFT was trained first, on the plain recipe of Section 4.2; the six arms of Table 7 came after it, each changing one of the recipe's three decisions (the table names what else moves with each), and each ended below it or below the base. Three of those changes rule out the reading that the traces were bad, and each closes one road. Multi-turn trajectories from a teacher solving with the sandbox in the loop taught the student to think less before each call and to lose its way in the tool loop. A tool declared in the student's training prompt over demonstrations that never call it taught the student never to call, and removing the declaration alone recovered most of the loss. Distilling on problems matched to the hard set, with both corrections applied, added no accuracy the pre-registered criterion could see, and the distillation line was closed (Appendix A). What the traces transfer is the teacher's procedure more than its physics. On the teacher's side the sandbox is a plain positive for data production: it raised the yield of usable traces from 63.3 to 98.3% and removed truncation (35% to 0, paired $p < 10^{-5}$). That the shipped adapter is trained on tool-free

2026 S. T. Yau High School Science Award 仅用于2026丘成桐中学科学奖公示

traces is a matter of trace form, not of the tool.

Sampling. Under temperature sampling the fine-tuned model's samples sit below its greedy answer, so the enhancement tier samples the base model (Section 5.5). What remains to improve lies at answer time, where the next two subsections measure it.

Table 7. Ablations of the training recipe. Every arm was trained after Gewu-SFT and differs from its row in the cells shown. Base references: 73.1 on the medium set under the generic contract, 58.5 on the hard set under the domain contract; 1σ 2.7 and 2.2 pp. Gewu-SFT's row repeats Table 4. Teachers: DeepSeek-V4-pro (preview) without a tool for Gewu-SFT, V4-pro with the sandbox for the tool trajectories, V4-flash for the other arms, with the sandbox on their numeric items; each arm's provenance is in Appendix C.

Arm	Teacher prompt	Traces	Student's training prompt	Ruler	Acc.	Behaviour
Gewu-SFT	plain	4,983 tool-free, V4-pro	generic suffix only	medium	75.5	calls the sandbox once it is declared at serving time
Change 1: tool trajectories in the training traces (medium set; this comparison also changes the problem source and the teacher's version)						
mixed arm	half and half	half of the tool arm's trajectories, half tool-free	generic, tool declared	medium	67.6	
tool arm	with the sandbox	7,000 tool trajectories, V4-pro	generic, tool declared	medium	63.2	825 thinking tokens per turn against 2,880; calls on 78%; 30% unfinished
Change 2: the tool declared in the student's training prompt (hard set; the same 2,526 traces)						
analytic-declared arm	serving contract	2,526 analytic, tool-free, V4-flash	domain, tool declared	hard	29.2	calls on 0.6%
analytic-undeclared arm	serving contract	the same 2,526	domain, declaration removed	hard	45.0	calling restored, +15.8 pp
Change 3: the serving contract in the teacher's prompt (hard set; this pair also changes the student's prompt)						
echo arm	serving contract	4,058 with the sandbox on numeric items, V4-flash	domain, tool declared	hard	43.9	echoes the teacher's instructions in 55% of responses; 9% unfinished
corrected arm	plain	5,031 with the sandbox on numeric items, V4-flash	minimal suffix only	hard / medium	43.1 / 71.9	calls on 46% / 97%; a fifth unfinished; the pre-registered line not met

5.4 The Tool and the Contract

Where the tool is needed, and what it is worth. Of the N-type problems the base model gets wrong, 70.2% show no tool call; on N-type problems it scores 80.6% when it called and 33.2% when it did not. Both statements are observational and the second compares self-selected groups; the positive control is the same split on A-type problems, which differs by only 8.2 pp. The causal measurement is the paired set, 1,436 problems from four public collections, each answered by the base model with the sandbox declared and without it. The set predates the audit rule of Section 3.6 and is used only for this within-item contrast, in which both arms answer the same problem against the same key, so a wrong key costs both arms the same item and cancels in the difference instead of propagating into it. On the 312 problems (21.7%) on which the model chose to call, having the tool is worth +14.7 pp from item counts (73.1 against 58.3, paired $p = 1.2 \times 10^{-6}$). On the 1,124 it did not call, the declaration alone costs −6.3 pp (35.1 against 41.5, $p = 2.3 \times 10^{-5}$): a declared tool changes how

2026 S.-T. Yau High School Science Award 仅用于2026丘成桐中学科学奖论文公示

the model reasons even when it never calls it. Over the whole set the two arms differ by −1.7 pp ($p = 0.21$), and routing the tool to exactly those 312 problems would be worth +3.2. How often a tool is needed depends on the distribution: the base calls on about 60% of the bank's N-type problems, on about 9% of olympiad problems (Table 3) and on 1% of PHYBench, so a tool gain is reported with its call rate rather than as one number.

Tool use without tool data. The fine-tuned model saw no tool in training and, once the sandbox is declared at serving time, calls it on 54.2% of medium-set problems, with at least one successful call on 99.3% of them. This is not a syntax the adapter picked up by accident. The base model's line is post-trained for tool calling [40, 41]; the declaration switches that ability on, and an adapter trained on tool-free traces leaves it intact, so the model applies an ability it already had to a tool it had never seen. Training can also switch it off: a training prompt that declares the tool over demonstrations that never call it produces a model that almost never calls, and removing the declaration alone recovers most of the loss (Table 7).

Prompts overfit; weights do not. The domain contract was written for the bank's problems. At home it pays: on the hard set the base model scores 35.9 under the minimal contract, 54.4 under the generic and 58.5 under the domain contract (the last two 4.1 pp apart, $p = 0.08$, inside the floor). Abroad the order reverses: on the endorsed set the base model scores 64.0 under the minimal contract, 54.5 under the generic and 27.2 under the domain contract (the domain rows on the 397 items on which that contract was run; the minimal and generic pairs are in Table 3), with protocol termination at 100% in every case, so the loss is in the instructions being followed, not in formatting. The domain contract is thus worth +22.6 pp at home, and abroad it costs 23.7 pp against the generic contract on the 397 items both were run on (50.9 to 27.2). The weights do not behave like this: the fine-tuned model, trained on the training pool and served under the generic contract, is inside the noise floor of the base at home and ahead of it abroad under every contract, by 6.2, 7.6 and 9.1 pp (70.2, 62.1 and 36.3 on the same items). A prompt fitted to one distribution is a capability there and a handicap elsewhere; the weights carry no such handicap. Hence the deployment rule: open-ended questions are served under the generic contract, and the domain contract only where the distribution is known to match the bank (Section 6.2).

5.5 Test-Time Search

The enhancement tier was judged against a line drawn before the run: at least +5 pp over the strongest single arm on the hard set, the base model's 58.5% under the domain contract, with a no-op control (one of the eight samples chosen at random) fixed at the same time so that a gain from voting could be told apart from a gain from sampling.

Eight samples of the base model, voted, score 74.1% (Figure 8): +15.6 pp over the strongest single arm and +14.6 over the no-op control of 59.5, which is itself level with greedy decoding; the line at 63.5 is cleared by 10.6 pp. Four samples reach 67.6 on average, and the 70 four-sample subsets of the same eight span 64.9 to 70.4, which is why every SC@k number in the paper states its vote composition. The gain is larger on numeric problems (48.3 to 66.2 on 225 items) than on analytic ones (68.2 to 80.2 on 288), both measured from the no-op control. The oracle, at least one of eight correct, is 84.6%, so the verifier headroom is 10.5 pp. The vote count is a usable confidence signal: where five or more of the eight agree, 48.7% of problems, the served answer is correct 97.6% of the time. The contract matters as well: the same eight-sample vote under the generic contract scores 68.6% (352 of 513); the domain contract alone is right on 49 problems, the generic alone on 21 ($p = 0.001$); on numeric problems the domain contract calls the sandbox on 72.9% of samples against 60.7%.

Which model to sample. Under temperature sampling the fine-tuned model's samples sit below its greedy answer: the path the traces taught is its best, and samples that leave that path score below it, 44.2% per sample at temperature 0.6 under its serving contract, where a base sample scores 59.5 against a greedy 58.5. Its vote of eight therefore stays close to its greedy answer (55.8) where the base model's climbs to 74.1. The width of the sampled distribution is unchanged (the envelopes open by +28.7 and +25.1 pp above pass@1, and all eight samples fall in one cluster on 7.8 and 8.2%

2026 年 Yau High School Science Award 复旦中学科学奖入选

of problems); the oracle of eight reaches 72.9 against 84.6, and sixty-five problems are solved within eight samples by the base only against five by the fine-tuned model only (McNemar $p < 10^{-13}$). The enhancement tier therefore samples the base model and pays for it in time: eight generations, each two to three times the length of the fine-tuned model's one. Figure 7 plots the base model's curves.

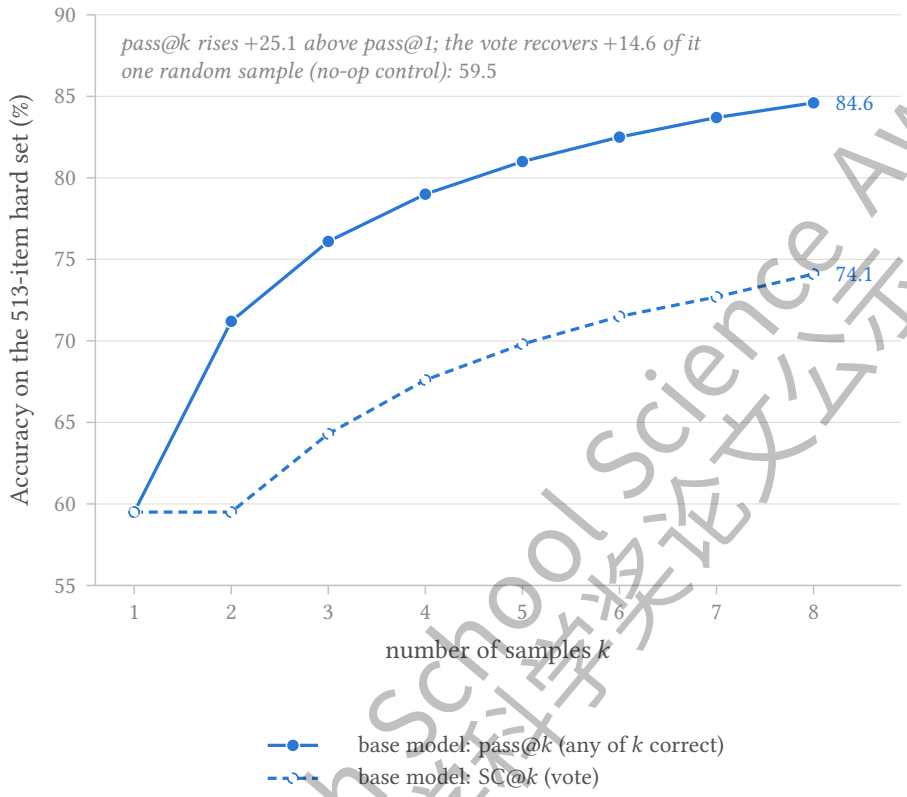


Figure 7. The base model's pass@k envelope, hard set, domain contract, temperature 0.6, $k = 1$ to 8. Solid: pass@k, the oracle for any selector; dashed: SC@k, the vote of Eq. (3); means over all k-subsets of the eight. The gap at $k = 8$ is the verifier headroom.

The fine-tuned model's answer as a vote. Counting the fine-tuned model's samples or its greedy answer in the vote was tested and not adopted: at their best weight they move the vote by +0.6 pp (74.7 against 74.1, McNemar $p = 0.45$), at weight two or more they cost accuracy, and the room was small to begin with, the base vote being wrong and the fine-tuned model right on only 13 problems (Appendix C). The served tier counts the base model's eight samples only and shows the default tier's answer beside the result.

Headroom and cost. Agreement is a weak verifier, and the 10.5 pp between the vote and the oracle is that gap. The search makes no network call: a session at 8k context holds about 0.3 GB of cache and linear-attention state, so eight samples cost about 2.5 GB on the card that already holds the weights (Section 5.7), and the latency is paid in the open, since the tier is reserved for problems the default tier cannot settle.

5.6 Further Studies on the Built System: the RLVR Pilot

The system is complete without reinforcement learning. Two of its design decisions rest on measurement: the fine-tuned adapter in front (Section 5.3), and the base model, not the adapter, as the sampler behind it (Section 5.5). A pre-registered reinforcement-learning pilot run on top of the finished system tests both from an independent direction.

Design. The pilot asks one question: can the accuracy that eight samples and a vote produce at inference time be trained into the weights, so that a single pass gives it? It is a group-relative policy

2026 S.T. Yau High School Science Award 仅用于2026丘成桐中学科学奖论文公示

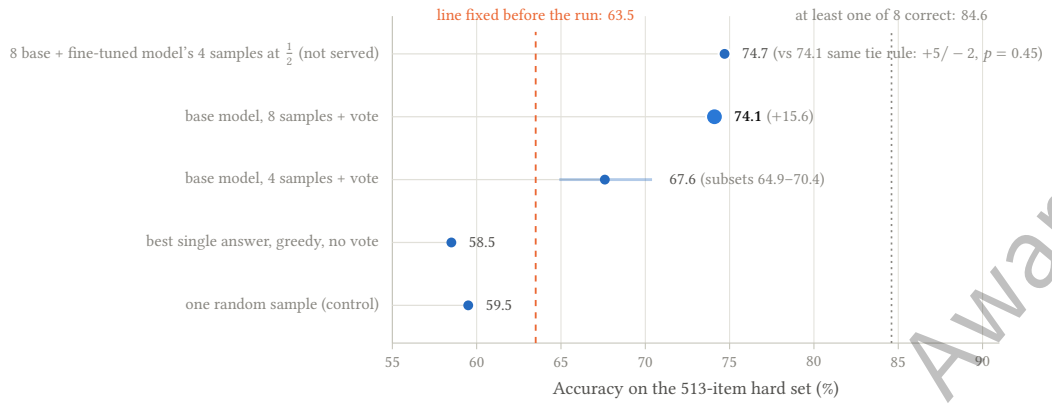


Figure 8. Test-time search on the hard set, domain contract, samples at temperature 0.6. From the bottom: the no-op control (59.5); the best single answer, greedy (58.5); four base samples voted (67.6, bar = the 70 four-sample subsets, 64.9 to 70.4); eight base samples voted (74.1, the served composition; the acceptance line was fixed before the run at 63.5); the same eight plus the fine-tuned model's four samples at half weight, tested and not served (74.7; $+5/-2$, $p = 0.45$). Dotted: the oracle, at least one of eight correct (84.6).

update [30] implemented as advantage-weighted cross-entropy over groups of eight samples per problem, assembled from components verified elsewhere in the paper.

Training set. The pass@8 distribution of the hard set defines the learnable band: of 513 problems, 79 are never solved in eight samples of the base model, 126 in all eight, and 308 in some but not all, and the last group is where a reward can move a policy. The training set is 365 numeric-answer problems from the training-side pool, isolated from the medium set by sub-topic and from the hard set by item, with a positive control confirming zero leakage into the 513; problems the base model gets wrong are admitted only with an independent endorsement of their key (Section 3.4), since a wrong key is a reversed reward, and 35 unendorsed items were removed.

Reward and loss. The reward is the deterministic grader with the judge out of the loop (of 293 responses the grader marked correct, the judge would have added one); $r \in \{0, 1\}$. For the $G \le 8$ samples of one problem that reached a gradable answer,

$$A_i = \operatorname{clip}\Big(r_i - \frac{1}{G}\sum_{j=1}^{G} r_j,\ -0.5,\ +1\Big), \tag{7}$$

and the update minimises the advantage-weighted token-mean cross-entropy of each sample y_i under the policy π_θ,

$$\mathcal{L}(\theta) = -\sum_{i:\,A_i \neq 0} A_i \frac{1}{|y_i|}\sum_{t=1}^{|y_i|} \log \pi_\theta\big(y_{i,t} \mid x_i,\ y_{i,<t}\big), \tag{8}$$

so a sample is pushed up or down by how much better or worse it did than its own group, and a group without variance drops out.

Gates and loop. Each outer iteration passes three behavioural gates before the next: tool-call-rate drift within ±15 pp, mean length within 1.5×, and at least 25% of responses reaching a valid conclusion. A tripped gate stops the run for review. Iterations are offline: roll out eight samples per problem, train at learning rate 10^{-5}, merge the adapter into the weights and reload them, evaluate on the 513 with the same grader. Two iterations were run, under a verdict fixed before the first: ±3 pp around the starting policy's 57.1% on the hard set under the generic contract, the starting policy being the base model.[2]

[2]The 57.1% is the greedy run the updates were evaluated against. The base model under the same contract scored 54.4% in the separate greedy run of the contract comparison of Section 5.4; the two differ by 2.7 pp, inside the 2σ floor, which is why every paired comparison names the reference run it was made against.

2026 S.-T. Yau High School Science Award 仅用于2026丘成桐中学科学奖论文公示

The pilot moved behaviour by amounts no noise floor could absorb and accuracy by an amount inside the floor, and the updated policy, sampled, sits below the base model it was trained from. The first result is consistent with the paper's central finding, that the accuracy this system has on the hard set is produced at inference time and not by further updates to its weights under a verifiable reward of this budget, so the adapter stays the served weights and reinforcement learning stays a pilot. The second repeats for the updated policy what Section 5.5 found for the adapter: a policy trained towards one way of finishing samples below the base model—not because its sampled distribution is narrower, the pass@k envelopes open as far, but because more of its samples never finish.

Two updates from the base model. With undecided items counted wrong on every arm, as everywhere in the paper (the updated policy produced 18 to 24 of them where the start produced none), the dose curve runs 57.1, 54.4, 56.7%: −0.4 pp from the start, paired $p \approx 0.93$, neutral. Behaviour during training moved by amounts no noise floor could absorb: truncation fell from 36 to 19%, response length by 22%, the tool-call rate rose from 54.5 to 71.8% (far enough to trip the ±15 pp drift gate; released after review), and the share of responses reaching a valid conclusion rose from 59 to 70%. The reward's steepest descent was along finishing and delegating to the tool, not along physics: a truncated response carries no gradient in a group where every response is truncated, and once the policy finishes, the group contains negatives (in-group negatives 0 to 488, training rows 598 to 1,136); the second update trained on that signal, recovered the first update's drop, and did not cross the start.

Why neutral. Not the starting point: starting from the same base model under the minimal contract instead, at 35.9%, scored 31.8% after the same two updates, −4.1 pp. The gradient budget was small, two updates, and the reward was dense in the positive direction, with group means above 0.75, leaving little advantage (Eq. (7)) to learn from.

As a sampler. The second update from the base model, sampled four times under the domain contract at temperature 0.6, scores 57.1% per sample against 59.5 for the base, 63.9 voted against 67.6 (below the lowest of the seventy base four-sample subsets, 64.9), and 77.0 at pass@4 against 79.0, with 11 to 14% of its samples unfinished against under 2% of the base model's at a call rate of 81 to 84% against 62%: the updates moved the policy towards calling more often, and under sampling that policy runs out of turns more often, while its envelope opens as far as the base model's (+19.9 against +19.5 from one sample to four). The enhancement tier therefore keeps the base model as its sampler, for the reason Section 5.5 gave for the adapter: the trained policies do not sample a narrower distribution; their samples sit below the base model's.

5.7 Deployment Measurements

Table 8. The quantisation gate on the hard set ($n = 513$, $1\sigma = 2.2$ pp): a profile passes with a loss of at most 3 pp, paired McNemar $p > 0.05$, and no drift in call rate or termination rate, both reported in the text. Weight sizes are measured files; the adapter adds 173 MB.

Profile	Weights	Δ against its 16-bit weights (pp)	Gate
bf16 base, domain (reference)	19.3 GB	—	does not fit
W8A16 base, adapter off, domain	11 GB	+1.9 (n.s.)	pass
W8A16 fine-tuned model (merged), generic	11 GB	+2.5 (n.s.)	pass
W4A16 base, domain	7.6 GB	−25.2	fail

The gate. The 8-bit profile passed (Table 8): 60.4 against 58.5%, +1.9 pp; its call rate moved from 60.4 to 54.2% and its termination rate from 99.2 to 98.6%, drifts for which the gate set no threshold. The gate covers single-pass decoding. Run again on the 8-bit weights with eight fresh samples of the same items under the same decoding parameters, the vote of Section 5.5 scores 71.3% (366 of 513) against 74.1 with the 16-bit weights; votes the judge did not reach are counted wrong, so 71.3 is a lower bound. The gap is 2.7 pp, inside twice the floor of Eq. (1) and not significant (paired $p = 0.10$).

2026 S. T. Yau High School Science Award
仅用于2026年成都中学科学奖公示

The fine-tuned model in 8 bits moves by +2.5 pp against its own 16-bit weights (paired $p = 0.30$), call rate 59.8 against 58.3% and termination 90.4 against 91.2% within that pair, and passes; the gate was measured on merged 8-bit weights, and the served form, the adapter on the 8-bit base, is level on the same items (paired $p = 0.60$ against the merged 8-bit weights and $p = 0.65$ against the 16-bit weights). The 4-bit profile failed informatively: 33.3%, −25.2 pp, responses of twice the length and no format collapse, a degradation of reasoning rather than of syntax, which we attribute, as a hypothesis, to the linear-attention layers' state and gating projections; it is reported as a controlled negative and not shipped.

Concurrency, measured on a 24 GB card. An RTX 3090 (24 GB, rented for the purpose) served the 8-bit weights with the adapter attached under vLLM at 0.92 memory utilisation. The runtime's KV pool held 28 concurrent sessions at 8k context and 8 at 32k. A single default-tier answer streamed at 58 tokens per second (1,024 tokens in 17.8 s); the enhancement tier's eight base samples, 2,048 tokens each, ran concurrently in 29.4 s at 557 tokens per second aggregate, inside the same pool. A dense 9 B model would serve a quarter as many sessions. The concurrency, throughput and adapter probe were measured on the 3090 (Appendix D).

What is served. The default tier serves the fine-tuned adapter on the 8-bit weights under the generic contract, in one pass; the vote costs eight generations per question and is reserved for problems the default tier cannot settle. The domain contract is an option for channels whose distribution is known to match the bank. The adapter served is Gewu-SFT; the other candidate, the second reinforcement-learning update, matches its starting policy under greedy decoding, samples below it, and was not run on the endorsed set, so it stays a pilot.

6 Discussion

6.1 Why Fine-Tuning Shapes Behaviour Rather Than Accuracy

The traces transferred procedure and language more than physics: under the levelled protocol the two models are level on the bank, and the fine-tuned model is ahead where the problem distribution is close to the traces' (Section 5.3). Two mechanisms account for what did change.

One sharpened path, not a narrower distribution. The traces carry two things. The first is a distribution, and a narrower one would be the natural reading of three epochs on a single teacher's traces. The samples show no narrowing (Section 5.5): agreement rates match the base model's and the pass@8 envelopes open by the same order; what sampling does is move the fine-tuned model below its own greedy answer while leaving the base level with its. The reading we can defend is that the traces sharpened one good path—the greedy one, with the Chinese, contract-following behaviour it carries—without lifting the distribution around it, which is why the fine-tuned model gives a single answer level with or above the base's and the base model is the better sampler. A vote can only average what the samples contain: the fine-tuned model's vote stays close to its greedy answer while the base model's climbs well above its, which is why the enhancement tier samples the base and pays for it in eight generations (Section 5.5). The narrowing reported after reinforcement learning [34] does not appear here: the envelope did not narrow, and the samples are spread as widely as the base model's.

Structure moved from the prompt into the weights. The second thing the traces carry is structure: how a solution is laid out, when to stop, what a final answer looks like. Once that is in the weights the prompt that used to supply it is redundant, which is why the base model swings between contracts and the fine-tuned model less (Section 5.4), and why with no system prompt at all the fine-tuned model keeps its language and its margin abroad: level with the base on the hard set, ahead of it on the endorsed set by about the margin it holds under the generic contract. Robustness to wording is the prompt's job having moved into the weights.

Why the tokens fell. The traces are the teacher's finished derivations, one layout and one stop, and the fine-tuned model reproduces their economy: roughly a third to a half of the base model's

2026 年 Lyau High School Science Award 入围作品 2026 届科学奖论文公示

tokens on all four sets of Table 6 under the served contracts, at the same score within the floor or a higher one. The base model has to be told from outside to check once and stop, the rule the domain contract imposes. The economy is relative to the contract: with no system prompt at all both models answer the endorsed set in about two thousand tokens (medians 2,078 and 1,912 in the minimal-contract runs of Section 5.2), so what fine-tuning removed is the base model's long response to the contract's instructions, not a cost the base model carries on its own. What the traces changed is how a derivation is carried and closed: the fine-tuned model writes it the teacher's way, in one pass, where the base looks back and forth and closes unevenly, at two to three times the length for the same or a lower score. A fixed layout, a stop rule, one boxed answer and one language, kept under every contract we tested, are what the external sets reward; on the bank's numeric problems, what the traces taught the model to do the base model could already do once told how.

Lessons for distilling a small physics solver. Four rules came out of the pilots (Table 7, Section 5.6).

1. **Keep the teacher's prompt plain.** Every instruction placed in it reached the traces verbatim and, from there, the student, which learned to recite the instruction sooner than the physics: the hazard Orca 2 guards against by withholding the teacher's instructions from the student [17], here measured.
2. **Treat tool-augmented traces with care.** A teacher that can call a sandbox reasons less between calls, and a 9 B student imitating it calls more and finishes less. Tool-free traces, with the tool declared only at serving time and the sandbox kept on the teacher's side, where it raised the yield of usable traces, gave a student that calls the sandbox in the declared format without ever having seen one: the separation between reasoning and computation that PAL [19] draws at inference time, drawn here at distillation time.
3. **Prefer short, plain traces.** A large teacher can afford to explore before it converges; a 9 B student cannot follow every jump, the learnability gap reported for small students of long chains [16]. That is our reading of why the plainer traces trained better.
4. **Expect behaviour, not physics.** Fine-tuning at this scale shapes behaviour, the language, the protocol, when to stop, and buys accuracy through those behaviours more than through new physics; two verifiable-reward updates from the base model likewise moved the behaviours it rewarded and left accuracy where it started (Section 5.6).

6.2 Prompts Overfit, Weights Do Not

The domain contract was written for the bank's problems, which end in a number: one dimensional check after the number, then stop; three to four significant figures; several quantities in one box. On the hard set those are the right instructions, and the contract scores 22.6 pp above the minimal one (Section 5.4). On olympiad problems, mostly symbolic, on which the base model calls the tool on about 9% of items, the same instructions are the wrong ones, and the model follows them just as faithfully (termination stayed at 100%): there the contract scores 23.7 pp below the generic one. The adapter crosses the same boundary at no measurable cost, level with the base at home and ahead of it abroad. The name is borrowed from prompt optimisation, where a prompt tuned on one distribution is known to lose accuracy under a distribution shift [42]; here the prompt was written by hand and the effect is the same. Our reading is that a prompt fixes a procedure from the outside, where the model cannot decide not to follow it, while weights trained on varied problems also learn when the procedure applies; that the swing is some 20 pp either way we read as a small model's prompt sensitivity, a size dependence we did not test. Two consequences follow.

- A bespoke prompt should be reported with an off-distribution control, or its gain is a property of the test set.
- A product that does not control its question distribution should ship the generic contract and put structure into the weights, which is what the default tier does.

6.3 Where the Accuracy Comes From

Both gains the system adds on hard problems are produced at answer time, and each has a shape.

2026 S-[You High School Science Award
仅用于2026日本高中生科学奖论文公示

The tool's gain is a routing problem. It is conditional: large where the model calls, negative where the declaration is present and it does not, and small over a whole set (Section 5.4). It is also a gain the public sets barely show: the three sets our training pool draws on hold almost no N-type problems, and the base model calls the tool on about 9% of the olympiad items and 1% of PHYBench's, so a model's weakness on the numerical step, and the sandbox that answers it, are visible most clearly on a bank that asks for that step.

The vote's gain is a verifier problem. It is unconditional but bounded: the oracle of Section 5.5 marks 10.5 pp that the samples already contain and the vote fails to pick.

Diversity is an agreement signal, not a vote. Diversity between models exists: each model solves endorsed-set items the other does not (Section 5.2), yet adding the adapter's samples to the vote moves it by less than a point (Section 5.5), because within eight samples the adapter rarely solves what the base cannot; its samples add weight to answers the base already holds, not new candidates. What agreement can still do is mark confidence: where five or more of the base model's eight samples agree, the answer is usually right (Section 5.5).

6.4 Implications for Small-Model Tutoring

What a department gets. What it can expect from a system of this size is now a measured quantity (Section 5): one card, no network call, tens of sessions at once, and an answer to a bank problem streamed in about a minute on the default tier and about three on the base. On olympiad-level problems the base model with the tool, one pass, answers at the level of the benchmark's larger open models and at about half a frontier model's accuracy under the same grader; on the hard set of its own bank the enhancement tier answers correctly about three times in four, and the default tier reasons in Chinese.

What it is not. It is not a teaching assistant: on hard problems it is wrong about one time in four, and the product has to show that. The streamed derivation lets a student check the reasoning rather than accept the answer, and the vote count of Section 5.5 marks the answers most likely to be right, on about half of problems.

The checklist. The protocol of Section 3.1 is the checklist for anyone about to trust a fine-tune. The steps are cheap to run, and on this project several of them changed a conclusion: the levelled baseline of Section 5.3 and the audit of Section 3.6 among them.

6.5 Limitations

The limitations are those of a study of a few months.

- **One base model, one teacher family.** Every result about fine-tuning is for one base model and one teacher family, and the teacher whose traces made the adapter has since been retired, so the recipe can be described but not re-run.
- **Noise floors.** The floors of Table 2 mean that several neutral verdicts are statements about a floor, not about zero.
- **The endorsed set.** It is conditioned on the endorsing teacher, a condition stated with every number reported on it; the whole clean pool of Section 5.2 removes that selection and keeps the margin.
- **The judge.** The model that settles what the rule-based grader cannot is of the teacher's family (Section 3.6); the teacher's contested verdicts were reviewed; the judge's agreement with a judge of another family was not measured.
- **Language.** The traces are mostly in Chinese, so the behaviour that transferred is the behaviour a Chinese-reading student sees; the same recipe in another language was not tested.
- **Scope of the pilots.** The reinforcement-learning pilot is two updates, the system serves text only, and the deployment was measured in one session, not under a term of student load.
- **Effect on understanding.** Whether the answers help a student understand was not measured, and there is no classroom use yet; the system has been tried only by the author and people around them.

2026 S. Tiyau High School Science Award
仅用于2026日成桐中学科学奖公示

7 Conclusion and Future Work

This paper asked whether an idle 24 GB card can run a physics tutor a department could trust, and found that the harder part of the question was the word *trust*. Three things came out of it. A measurement protocol, ordinary experimental hygiene assembled for claims about small models, under which a naive fine-tuning gain on our own bank fell to the noise floor while a gain on an external olympiad set survived, and GewuBank-20K, the instrument built under it, its keys checked and its difficulty measured, released for later work to train and evaluate on. A finding about training at this scale: fine-tuning on a stronger teacher's solutions transferred the teacher's way of working, in Chinese, in a fraction of the tokens, with the tool called in the declared format and the prompt mattering less, while the accuracy on hard problems was added at answer time, by the tool where the model calls it and by a vote over the base model's samples, and not by two further updates under a verifiable reward. And Gewu, a small, deployable tutor: one 8-bit weight file serving both tiers on one card with no external API, ahead of its base on the endorsed set under every contract we tested and, on PhysElite, level with models many times its size. Against the three requirements of Section 1.1: on the hard set the enhancement tier answers 74.1% of problems in 16 bits (at least 71.3% on the 8-bit file) and marks the answers most likely to be right; the default tier answers in a third to a half of the base model's tokens; and both run from one 11 GB file on one card with no network call.

Three directions follow. The gap between the vote and its oracle is a verifier problem: a rule stronger than answer agreement is needed to pick the right sample. The next adapter should be selected on its samples, not only on its greedy answer, and the next reinforcement-learning attempt should change the budget and the reward rather than the starting point, whose variation landed below the pre-registered band (Appendix A). And a student's problems often come with a diagram; a bank that carries its figures, and training that uses them, is what the ranking set, most of whose items refer to a figure the model cannot see, asks for. All three run through the bank, which is why it is released. Beyond them, the tutor has to be used in a classroom, where its effect on a student's understanding can be measured.

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖公文公开

A Pre-Registered Criteria

The noise-floor rule was fixed in July 2026, before the comparisons of Section 5, and has not changed since; the grader has been fixed since 2026-08-11 (Appendix D). The practice of also writing the acceptance line before a run dates from 2026-09-03; the verdicts below are the decisions taken from then on, each written down before its result existed. The wording is condensed from the project logs, which carry the full texts with their dates. The measurement protocol of Section 3.1 is what these rules have in common.

Verdict	Fixed	Criterion	Outcome
Distillation line	2026-09-03, before the runs	Keep distilling only if the corrected arm beats the base by at least 3 pp on the medium set with a healthy tool probe (A-type call rate 30–90%, protocol pass $\geq$ 85%, output at most twice the no-tool arm's); otherwise freeze fine-tuning and carry Gewu-SFT forward.	71.9 against 73.1: frozen; Gewu-SFT carried forward (Section 5.3).
Test-time search	2026-09-03, before the samples	Eight samples reconciled by blind symbolic clustering, judged against the strongest single greedy arm on the hard set (58.5): at least +5 pp accepted, +2 to +5 reported as mild, under +2 a negative result; a no-op control (one random sample) reported beside the vote. The composition written down was four base and four fine-tuned samples; after the fine-tuned model's samples measured below the base model's (Section 5.5) the vote is eight base samples, the criterion and its reference unchanged.	74.1, +15.6 pp; line at 63.5 cleared (Section 5.5).
RL pilot	2026-09-05, before the first update	±3 pp around the starting policy's greedy score on the hard set under the generic contract (57.1): above, effective; inside, neutral; below, a negative result reported as such. Behavioural gates each outer iteration; a hard stop on the evening of 2026-09-08. Rollout and serving budgets were raised once, before the first gradient step.	56.7, −0.4 pp: neutral (Section 5.6).
RL from the base, minimal contract	2026-09-07, before the run	Single-variable A/B against the pilot: only the contract the policy is trained and evaluated under changes; verdict ±3 pp around the base's 35.9 under the same contract.	31.8, −4.1 pp: below the band (Section 5.6).
Quantisation gate	2026-09-08, before the first quantised run	A profile ships only if it loses at most 3 pp on the hard set, paired $p > 0.05$, with no drift in call rate or termination rate.	W8A16 passed (+1.9 base, +2.5 fine-tuned); W4A16 failed (−25.2) (Section 5.7).

B Prompt Contracts

The three contracts of Section 4.3 differ only in the text below. Both non-minimal contracts open with the same tool declaration, an XML description of the sandbox function and the required call format, reproduced first. The declaration is in English, in the form of the tool prompt the base

2026 S. T. Yau High School Science Award 仅用于2026年成绩科学论文公示

model's chat template uses for function calling; the working instructions are in Chinese, the students' language. What follows is the working instruction each contract appends after that block, and the suffix appended to every problem, reproduced verbatim from the prompt files.

Tool declaration — shared by the generic and domain contracts; verbatim

```
# Tools

You have access to the following functions:

<tools>
{"name": "python", "description": "Run Python in a stateful sandboxed interpreter (like a Jupyter kernel) to
 do any nontrivial symbolic or numerical computation. Available: sympy, numpy, scipy (incl. scipy.constants)
, math, cmath, fractions, decimal, mpmath. No filesystem, no network. Variables and imports PERSIST across
calls within a problem — no need to re-import each time. The value of the last bare expression is auto-
displayed (like Jupyter); print() also works.", "parameters": {"type": "object", "properties": {"code": {"
type": "string", "description": "Python code. End with a bare expression to see its value, or print()."}}, "
required": ["code"]}}
</tools>

If you choose to call a function ONLY reply in the following format with NO suffix:

<tool_call>
<function=example_function_name>
<parameter=example_parameter_1>
value_1
</parameter>
<parameter=example_parameter_2>
This is the value for the second parameter
that can span
multiple lines
</parameter>
</function>
</tool_call>

<IMPORTANT>
Reminder:
- Function calls MUST follow the specified format: an inner <function=...></function> block must be nested
within <tool_call></tool_call> XML tags
- Required parameters MUST be specified
- You may provide optional reasoning for your function call in natural language BEFORE the function call,
but NOT after
- If there is no function call available, answer the question like normal with your current knowledge and do
 not tell the user about function calls
</IMPORTANT>
```

Minimal contract — no system prompt; suffix only

请解答上述问题，并把最终答案放在 \boxed{} 中。

Generic contract (the default tier's) — working instruction, then suffix

你是一位物理学家。请完整、严谨地解决下面的物理问题。要求：先把物理想清楚：判断适用的定律、对称性、近似的适用范围，再动手算。推导要完整，但不要啰嗦重复。最后把最终数值答案放进 \boxed{} 里，只放数值（可带单位），不要放推导过程。全文只允许出现一个 \boxed{}，就是所求的那个数值。题目附带要求的定性说明（分类、自由度、物理意义等）写在正文里，不要放进 \boxed{}。如果这道题本身有毛病——条件缺失、自相矛盾、或者所求量没有定义清楚——不要硬凑一个数，直接输出 \boxed{ILL_POSED} 并简要说明缺什么。

Suffix. 请解答上述问题，并把最终答案放在 \boxed{} 中。如果题目本身有缺陷——缺少求解所必需的物理量或边界条件、记号有歧义（同一符号有多种合理解释）、或给定的数据在物理上自相矛盾——不要猜测题意，也不要臆造条件。请直接输出 \boxed{ILL_POSED}，并在其后用一行说明缺的是什么。注意：题目「难」不算缺陷。只有当你不臆造条件就无法确定题目要问什么时，才使用这个出口。

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

Domain contract — working instruction, then suffix

你是一位物理学家。请完整、严谨地解决下面的物理问题。要求：先判断适用的定律、对称性与近似范围，再动手算。推导完整，不重复。数值算出后做一次量纲与量级检查，通过即收尾；不重算，不换方法复核。

Suffix. 请解答上述问题，并把最终答案放在 `\boxed{}` 中，只放数值（可带单位），不放推导过程。最终数值给 3–4 位有效数字。全文只允许出现一个 `\boxed{}`。题目要求多个数值量时，按题面出现顺序写进同一个框，逗号分隔。定性说明（分类、自由度、物理意义、稳定性判断）写在正文，不要放进框里。如果题目本身有缺陷——缺少求解所必需的物理量或边界条件、记号有歧义（同一符号有多种合理解释）、或给定的数据在物理上自相矛盾——不要猜测题意，也不要臆造条件。请直接输出 `\boxed{ILL_POSED}`，并在其后用一行说明缺的是什么。注意：题目「难」不算缺陷。需要数值迭代、结果不好看，也都不算。只有当你不臆造条件就无法确定题目要问什么时，才使用这个出口。

C Alternatives Considered and Why They Were Not Adopted

Each entry records what was tried, what was measured, and why it is not in the system. Controlled negatives that bear on a claim stay in the body; this appendix holds the directions.

1. **A first stack with a same-model verifier (April to May 2026).** The first system was a LoRA fine-tuned on a public physics set, a verifier that was the same 9 B model under a second system prompt scoring the solution from 0 to 10 for up to three rounds, a symbolic-algebra skill and a retrieval store. On 30 electrodynamics items the bare base model scored 30.0%, the fine-tune 26.7, the fine-tune with retrieval 20.0 (33.3 with a relevance threshold) and the full stack 20.0: at that size none of the other four is distinguishable from the bare model, and a verifier that is the same model has no more physics than the solver. A comparison of three 4-bit models under the retrieval stack (a 2 B distillation, the 9 B fine-tune, DeepSeek-R1-Distill-Qwen-7B) kept the 9 B. The stack was replaced by long-trace distillation (Section 4.2); the blind vote of Section 4.4 took the verifier's place.
2. **Retrieval-augmented generation.** A corpus of 5,169 chunks over fourteen areas of physics (3,213 worked examples, 1,136 formulas, 820 concept notes), embedded with BGE-M3 and re-ranked, was measured in July 2026 on the 251-item internal hold-out of the time, under the grader of the time. It lifted the base model from 57.4 to 64.9% (19 items) and left the first fine-tuned adapter at 79.7% either way; on 100 PHYBench items no single component was distinguishable from noise. A later check found why: the corpus is Chinese while 50.8% of the paired set is English, the median top-1 relevance is 0.202 for English items against 0.648 for Chinese, and querying with the answer key itself does not raise it (0.309 against 0.335), so the corpus does not hold what the hard problems need. Not adopted: in a fine-tuning setting the knowledge goes into the weights, and a retrieval round adds latency for a measured gain of zero on the model that is served.
3. **Distilling with the serving contract in the teacher's prompt, and with the tool in the loop (the arms of Table 7).** Distilling with the serving contract in the teacher's system prompt produced traces that open by reciting the instruction they were given (100% of traces; 0% when the instruction is removed); the student recited it too, and where it had seen the bank's own numeric traces it reproduced the teacher's private vocabulary (echo arm). A tool declared in the student's training prompt over demonstrations that never call it taught the student never to call (analytic-declared arm), and removing the declaration alone recovered 15.8 pp (analytic-undeclared arm). Adapters trained on multi-turn tool trajectories reasoned less between calls and finished less (tool and mixed arms). A further arm applied both corrections on problems matched to the hard set and did not meet the criterion fixed on 2026-09-03 (corrected arm; Appendix A). Not adopted: the shipped adapter is trained on tool-free traces from a plainly prompted teacher, the tool contract is carried outside the traces by the machine gate of Sec-

2026 S. T. Yau High School Science Award
仅用于2026年丘成桐中学科学奖论文公示

tion 4.3, and the sandbox went to the teacher instead, where it raised the yield of usable traces from 63 to 98% (Section 5.4).

4. **Generating without the structural gate.** Letting the generator decide how many quantities an item asks and in what form its key is written. Not adopted: without the gate of Section 3.3 an item's tier is the generator's own claim and its key need not reduce to one gradable value, so neither the difficulty grid nor the symbolic grader can be trusted on it. The gate is applied to every item at intake, and an item that fails it is removed rather than repaired (Appendix D gives the counts).
5. **Serving the prompt that is best on the bank.** The domain contract lifts the base model by 22.6 pp on the hard set and costs it 23.7 pp on the endorsed set (Section 5.4). The default tier serves the generic contract; the domain contract is kept as an option for a channel whose distribution is known.
6. **The fine-tuned model as the sampler, and its samples in the vote.** At temperature 0.6 its samples are 44.2% per sample against 59.5 for the base model, 55.8 against 74.1 voted, and 72.9 against 84.6 as an oracle; as two votes in ten beside the eight base samples they move the vote by +0.6 pp (74.7 against 74.1, five problems gained, two lost, $p = 0.45$). On the odd-numbered half of the hard set the grid over per-sample weights 0.25, 0.5, 1 and 2 moves the vote by +0.0, +0.4, +0.0 and −3.5 pp, and the even half gives +0.8 at 0.5; the fine-tuned model's greedy answer as one or two votes moves it by +0.2 and −0.4, and at three or more costs −4.9 pp ($p < 0.001$) (Sections 5.3 and 5.5). The base model samples; the fine-tuned model's answer is shown beside the vote and not counted.
7. **Reinforcement learning as a product step.** Two updates from the base model moved behaviour and left accuracy at −0.4 pp; the same training set from the base model under the minimal contract gave −4.1; the second update from the base model is a worse sampler than the base (Section 5.6). It is reported as a pilot and is not in the served system.
8. **Public benchmark keys as shipped.** The audit of Section 3.6 found every public set it could place in a band unfit as a ruler as shipped, PHYBench's sample being too small to place; the endorsed set and the unfiltered ranking set followed. The first audit's 40 items also drew 14 from OlympiadBench [6] (2 wrong keys, 1 ill-posed) and 12 from PHYSICS [4] (3 wrong keys); neither set was used as a ruler.

D Reproducibility

Models in the data pipeline. Problem generation: GPT sol5.6 Ultra through its desktop application (August 2026; raw exports archived in the repository). Distillation teacher for Gewu-SFT: DeepSeek-V4-pro in the preview release served in July 2026, blind and tool-free, since replaced. Key verification of the bank: DeepSeek-V4-pro for the blind solve during the distillation runs, DeepSeek-V4-flash for the re-check runs of 2026-08-28, 09-12 and 09-13 (Section 3.4); V4-flash replaced V4-pro as the production teacher after a paired comparison on 200 N-type items under the same protocol put it within 4 pp (92.0 against 96.0%, McNemar $p \approx 0.06$) at half the output tokens. PhysElite: DeepSeek-V4-flash made the 1,092 blind attempts that define the endorsed set; DeepSeek-V4-pro is the reference row on the unfiltered 500-item ranking set and the fallback judge wherever a symbolic verdict is unavailable. No DeepSeek model appears as a comparison row on the endorsed set. Blind re-solve of 60 items (stage five): Claude (Anthropic), August 2026. Contested-verdict review (Section 3.4, stage five): Claude Opus 5 (Anthropic), 120 items in six batches of 20, every verdict recomputed with code.

Bank intake. The generation runs produced more rows than the bank holds. The gold audit of 2026-08-22 removed 45 keys; the deterministic derivation, one implementation shared with every ruler, could not reduce 464 keys to a single gradable value (275 flagged for review, 189 without a single value) and those items were removed; rows that never received the base-model measurement

2026 Shi-I Yau High School Science Award 仅用于2026石锐桐中学科学奖公示

of Figure 2 are not in the bank and served only as distillation inputs; after the verification of Section 3.4, the 431 items whose key the teacher contested were removed (their identifiers are kept with the bank). The bank is the 20,414 items that passed the gate, received the measurement and kept their key.

Key re-checks. DeepSeek-V4-flash under the distillation contract, A-type items without the tool and N-type items with it, the symbolic grader first and the judge where it abstains. 2026-08-28: 1,414 items, among them the 246 controls. 2026-09-12: 2,761 items, 46.6 M input and 20.9 M output tokens, 39 s per item. 2026-09-13: 4,398 items, 51.4 M input and 26.9 M output tokens, 30 s per item, after which every key had a verdict (V4-pro 11,912 items, V4-flash 8,933). Verdicts are recorded per item; no key was altered. An adversarial arbitration panel (an attacker, a defender and a recomputing judge, calibrated with hidden controls), run on 105 items of the external benchmark, measured the rate at which "the teacher says the key is wrong" converts to "the key is wrong" at 48.1%, a narrower outcome on a different population than the defect rate of stage five.

Calibration scan. Every problem in the bank carries a difficulty that was measured rather than declared (Figure 2); this is the run that measured it.

- **What was run.** The base model with the sandbox, under the generic contract, greedy, one pass per problem, 16,384 tokens per turn and at most twelve tool turns.
- **How much.** The bank was cut into 40 slices of about 500 items: slices 1–30 on the RTX 5090 under WSL2, slices 31–40 on a rented A100. That covers 19,867 items and 223.7 M generated tokens. The remaining 978 items were measured later in two runs under the same protocol and grader (500 and 513 items, 35 of them shared).
- **Coverage.** 20,356 of the bank's 20,414 items received a verdict. The 58 the grader could not settle count as wrong in Figure 2; over the items it could settle the bank scores 74.56%.
- **Machine change control.** Fixed before the A100 slices ran, 126 items of slice 30 were re-run on the A100: 75.4% there against 77.0 on the 5090, the same verdict on 81.0% of items, 13 flips one way against 11 the other. The difference is inside the floor, so slices from the two machines are read together.

Software and hardware. What was held fixed, and which machine did what.

- **Fixed software.** vLLM 0.27.1 on every machine; the base weights are 19,306,216,416 bytes in four shards; the grader has been fixed since 2026-08-11 and the evaluation harness since the calibration scan, both checksummed in the repository.
- **Sampling.** Greedy runs use a repetition penalty of 1.1. Sampled votes use temperature 0.6, top-p 0.95 and top-k 20, and each request's seed is recorded in its run's metadata.
- **Machines.** An RTX 5090 (32 GB) under WSL2: quantisation, calibration slices 1–30, part of the votes, the default-tier measurements. A rented cloud A100 (80 GB): slices 31–40, the reinforcement-learning pilot, most votes. A second rented A100 container: two votes and the first version of the demonstration; the screenshots in this paper come from the 16-bit demonstration service on the RTX 5090 (2026-09-13). A rented H20: five of the eight 8-bit vote samples. A rented RTX 3090 (24 GB, sm_86, 2026-09-12): the concurrency and throughput of Section 5.7. Each paired comparison names the reference run it was made against.
- **Archives.** `data/eval/` holds the rulers and the slices, `logs/` the per-run JSON with metadata and transcripts, `logs/samp_votes_0910/` the twenty votes with their machine of origin, and `figures_data/` the CSV behind every figure.

Adapter training. LoRA rank 32, $\alpha = 64$, dropout 0.05, twelve target modules, 86,556,672 trainable parameters; three epochs, 936 optimiser steps at batch size 1 with 16-step gradient accumulation, learning rate 5×10^{-5} on a cosine schedule with warm-up ratio 0.1, sequence cut-off 16,384 tokens;

2026 Siyuan High School Science Award
仅用于2026年度中学生科学奖公示

PEFT 0.18.1, Transformers 5.2.0, PyTorch 2.13.0; 21,308 s on one A100. The shipped adapter is the bf16 export of these weights.

Quantisation. W8A16 with llm-compressor GPTQ: int8 weights, 16-bit activations, group size 128, symmetric, calibrated on 384 bank problems rendered through the chat template (seed 0, up to 2,048 tokens each); the projections of the linear-attention layers are quantised as well, while their convolutions, normalisations and state parameters stay in bfloat16; the recipe file ships with the weights.

Availability. The bank, the traces, the adapter and the served weights are public on the Hugging Face Hub under the names used in this paper. Each repository carries a full SHA-256 checksum file, repeated in the supplementary package.

- `https://huggingface.co/datasets/Alumin-Hydro/GewuBank-20K`: the 20,414 bank items with the frozen medium and hard sets, the 431 contested items flagged rather than deleted, and the base-model measurement of every item (CC BY 4.0).
- `https://huggingface.co/datasets/Alumin-Hydro/gewu-sft-traces`: the 5,248 traces, licensed per source split (UGPhysics CC BY-NC-SA 4.0, WebInstruct-verified Apache-2.0, Loong per item).
- `https://huggingface.co/Alumin-Hydro/Gewu-SFT`: the adapter in bfloat16, 173 MB, with its serving-time key layout in the `vllm/` folder (CC BY-NC-SA 4.0, because part of its training data carries that license).
- `https://huggingface.co/Alumin-Hydro/Gewu-Base-W8A16`: the base weights quantised as in Section 5.7 (Apache-2.0).
- `https://huggingface.co/Alumin-Hydro/Gewu-SFT-W8A16`: the fine-tuned model merged and quantised by the same recipe (CC BY-NC-SA 4.0).

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公开

References

[1] R. Xu, W. Gao, L. Chen, et al. *PhysElite: How Far Are LLMs from Solving Olympiad-Level Physics Problems?* 2026. arXiv: 2608.25097.

[2] Y. Zhang, Y. Ma, Y. Gu, et al. *ABench-Physics: Benchmarking Physical Reasoning in LLMs via High-Difficulty and Dynamic Physics Problems.* 2025. arXiv: 2507.04766.

[3] X. Xu, Q. Xu, T. Xiao, et al. *UGPhysics: A Comprehensive Benchmark for Undergraduate Physics Reasoning with Large Language Models.* 2025. arXiv: 2502.00334.

[4] S. Zheng, Q. Cheng, J. Yao, et al. *Scaling Physical Reasoning with the PHYSICS Dataset*. 2025. arXiv: 2506.00022.

[5] S. Qiu, S. Guo, Z.-Y. Song, et al. *PHYBench: Holistic Evaluation of Physical Perception and Reasoning in Large Language Models.* 2025. arXiv: 2504.16074.

[6] C. He, R. Luo, Y. Bai, et al. *OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems.* 2024. arXiv: 2402.14008.

[7] X. Wang, Z. Hu, P. Lu, et al. *SciBench: Evaluating College-Level Scientific Problem-Solving Abilities of Large Language Models.* 2023. arXiv: 2307.10635.

[8] DeepSeek-AI, D. Guo, D. Yang, et al. *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. 2025. arXiv: 2501.12948.

[9] N. Muennighoff, Z. Yang, W. Shi, et al. *s1: Simple test-time scaling*. 2025. arXiv: 2501.19393.

[10] X. Huang, Rishabh, G. Franke, et al. *Loong: Synthesize Long Chain-of-Thoughts at Scale through Verifiers.* 2025. arXiv: 2509.03059.

[11] E. Guha, R. Marten, S. Keh, et al. *OpenThoughts: Data Recipes for Reasoning Models.* 2025. arXiv: 2506.04178.

[12] Z. Yuan, H. Yuan, C. Li, et al. *Scaling Relationship on Learning Mathematical Reasoning with Large Language Models.* 2023. arXiv: 2308.01825.

[13] E. Zelikman, Y. Wu, J. Mu, et al. *STaR: Bootstrapping Reasoning With Reasoning*. 2022. arXiv: 2203.14465.

[14] J. H. Cho and B. Hariharan. *On the Efficacy of Knowledge Distillation.* 2019. arXiv: 1910.01348.

[15] Y. Fu, H. Peng, L. Ou, et al. *Specializing Smaller Language Models towards Multi-Step Reasoning*. 2023. arXiv: 2301.12726.

[16] Y. Li, X. Yue, Z. Xu, et al. *Small Models Struggle to Learn from Strong Reasoners.* 2025. arXiv: 2502.12143.

[17] A. Mitra, L. Del Corro, S. Mahajan, et al. *Orca 2: Teaching Small Language Models How to Reason.* 2023. arXiv: 2311.11045.

[18] X. Ma, Q. Liu, D. Jiang, et al. *General-Reasoner: Advancing LLM Reasoning Across All Domains.* 2025. arXiv: 2505.14652.

[19] L. Gao, A. Madaan, S. Zhou, et al. *PAL: Program-aided Language Models.* 2022. arXiv: 2211.10435.

[20] W. Chen, X. Ma, X. Wang, et al. *Program of Thoughts Prompting: Disentangling Computation from Reasoning for Numerical Reasoning Tasks.* 2022. arXiv: 2211.12588.

[21] Z. Gou, Z. Shao, Y. Gong, et al. *ToRA: A Tool-Integrated Reasoning Agent for Mathematical Problem Solving*. 2023. arXiv: 2309.17452.

[22] K. Wang, H. Ren, A. Zhou, et al. *MathCoder: Seamless Code Integration in LLMs for Enhanced Mathematical Reasoning*. 2023. arXiv: 2310.03731.

[23] X. Yue, X. Qu, G. Zhang, et al. *MAmmoTH: Building Math Generalist Models through Hybrid Instruction Tuning*. 2023. arXiv: 2309.05653.

[24] T. Schick, J. Dwivedi-Yu, R. Dessì, et al. *Toolformer: Language Models Can Teach Themselves to Use Tools.* 2023. arXiv: 2302.04761.

2026 Shing Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

[25] J. Feng, S. Huang, X. Qu, et al. *ReTool: Reinforcement Learning for Strategic Tool Use in LLMs*. 2025. arXiv: 2504.11536.

[26] A. Meurer, C. P. Smith, M. Paprocki, et al. "SymPy: symbolic computing in Python". In: *PeerJ Computer Science* 3 (2017), e103.

[27] M. Kang, J. Jeong, S. Lee, et al. *Distilling LLM Agent into Small Models with Retrieval and Code Tools*. 2025. arXiv: 2505.17612.

[28] C. Li, M. Xue, Z. Zhang, et al. *START: Self-taught Reasoner with Tools*. 2025. arXiv: 2503.04625.

[29] Q. Cheng, Y. Zhang, Z. Wang, et al. *Teaching Thinking Models to Reason with Tools: A Full-Pipeline Recipe for Tool-Integrated Reasoning*. 2026. arXiv: 2605.06326.

[30] Z. Shao, P. Wang, Q. Zhu, et al. *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models*. 2024. arXiv: 2402.03300.

[31] A. Hochlehnert, H. Bhatnagar, V. Udandarao, et al. *A Sober Look at Progress in Language Model Reasoning: Pitfalls and Paths to Reproducibility*. 2025. arXiv: 2504.07086.

[32] X. Wang, J. Wei, D. Schuurmans, et al. *Self-Consistency Improves Chain of Thought Reasoning in Language Models*. 2022. arXiv: 2203.11171.

[33] B. Brown, J. Juravsky, R. Ehrlich, et al. *Large Language Monkeys: Scaling Inference Compute with Repeated Sampling*. 2024. arXiv: 2407.21787.

[34] Y. Yue, Z. Chen, R. Lu, et al. *Does Reinforcement Learning Really Incentivize Reasoning Capacity in LLMs Beyond the Base Model?* 2025. arXiv: 2504.13837.

[35] T. Chu, Y. Zhai, J. Yang, et al. *SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training*. 2025. arXiv: 2501.17161.

[36] E. J. Hu, Y. Shen, P. Wallis, et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685.

[37] E. Frantar, S. Ashkboos, T. Hoefler, et al. *GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers*. 2022. arXiv: 2210.17323.

[38] J. Lin, J. Tang, H. Tang, et al. *AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration*. 2023. arXiv: 2306.00978.

[39] W. Kwon, Z. Li, S. Zhuang, et al. *Efficient Memory Management for Large Language Model Serving with PagedAttention*. 2023. arXiv: 2309.06180.

[40] A. Yang, A. Li, B. Yang, et al. *Qwen3 Technical Report*. 2025. arXiv: 2505.09388.

[41] Qwen Team. *Qwen3.5-9B*. Hugging Face model card. 2026.

[42] M. Li, W. Wang, F. Feng, et al. *Robust Prompt Optimization for Large Language Models Against Distribution Shifts*. 2023. arXiv: 2305.13954.

2026 S. T. Yau High School Science Award
仅用于2026丘成桐中学科学奖论文公示

Acknowledgements

Background. In the spring of 2026, with the help of the Beijing Youth Science and Technology Club (北京青少年科技俱乐部), I joined a research placement in the Department of Physics at Tsinghua University, where I met Mr. Fei Song. Seeing the laboratory's consumer graphics cards sit idle, he suggested a starting direction: fine-tune a small model that fits on one such card and put it to work on physics question answering. My first two months were self-directed exploration and learning. From June onwards the design, execution and writing of the research were carried out by me.
Advisor relationship and role. Mr. Fei Song (Department of Physics, Tsinghua University) proposed the fine-tuning direction and helped me settle the research question. After the project started he followed its progress through the research log I kept on Yuque, which he read on several occasions. Mr. Dao He (physics teacher, the Experimental High School Attached to Beijing Normal University) gave me advice and help at the final stage of the writing. Both are the supervising teachers of this entry, and both supervised without payment.
Division of labour. This research was completed by one student. *Topic:* the fine-tuning direction came from Mr. Song. After a literature survey I fixed the specific topic myself, a tool-augmented 9B physics tutoring system together with a protocol for measuring it trustworthily. *Data:* I designed the two-dimensional difficulty frame (numeric and analytic answer types over three difficulty tiers), the generation specifications for its thirty physics sub-domains and the structural gates. The problems were generated in bulk by an AI tool under those specifications and entered the bank only after the five-stage verification I designed and a full-bank calibration. *Analysis and experiments:* I wrote the measurement protocol and the pre-registered criteria, and designed every controlled comparison. The scans and evaluations ran with AI assistance on my own GPU (an RTX 5090) and on rented cloud machines (A100s, an H20, and an RTX 3090 for the deployment measurements). I verified every result, and the key numbers were re-run across machines. *Writing:* I set up the writing rules and the ledger that every number must trace to, supplied the structure and content of every section, and decided every formulation. Mr. He commented on the finished draft.
Difficulties and how they were resolved. First, measurement: public answer keys were wrong often enough to change conclusions. I designed the five-stage key verification, and I use the external set only through an audit protocol and a double-endorsement construction, disclosed wherever its numbers appear. A naive comparison reported +37.5 percentage points, which shrank to +2.4 (not significant) once grader, budget, tool and prompt were levelled, a rule that now governs every comparison in this paper. Second, distillation: a 9 B student does not simply absorb a strong teacher's traces. Controlled comparisons tested the recipe: a plain teacher prompt, tool-free traces, the tool declared at serving time. Third, accuracy once the weights are fixed: the base model fails mostly by not carrying out the computation. The sandbox and the vote are the two decisions that answer it.
Declaration of AI use. Both supervising teachers knew throughout the project which AI tools I was using, and agreed to their use.

- *GPT5.4 Extended Thinking and Qwen3.6-Plus* (April to May 2026, occasional sessions in the exploration phase): knowledge overviews and trial data for early fine-tuning experiments, none of which entered the final work.
- *OpenAI GPT sol5.6* (the high and Ultra editions and the Codex agent, July to September 2026). In July, over about two weeks and not adopted: a fourteen-area retrieval corpus of 5,169 chunks reported in the appendix, and tool-call trajectories for an early tool-training attempt together with a quality review of them. Nothing from that line reached the final model or its training data. In August, in batched generation sessions over three weeks with per-chunk task sheets and progress logs: bulk generation of the physics problems, used as the generator under study. Every item was audited and calibrated before entering the bank. In September, in a few sessions: answers to physics questions that came up while I reviewed contested bank items by hand.
- *Claude* (Anthropic, June to November 2026, the Claude Code command-line agent, running Opus 5, Opus 5.5, Fable 5 and Fable 5.1 at different points in the period, in recorded working sessions

2026 全国青少年科技创新大赛 高中组 科学论文 Award
仅用于2026国际青少年科学论文大赛 High School Science

several times a week, nearly daily during the August experiments and the September writing): it wrote and debugged part of the code to specifications set by my experimental design and planning, organised data, executed experiments under criteria I had pre-registered, kept the research log, and assembled the submission's AI-usage records. In July it also drafted, with parallel sub-agents, an earlier version of the fourteen-area retrieval corpus named above, which was not adopted. It supplied literature-search leads, and every citation was verified by hand against the arXiv abstract page. It supplied physics knowledge and derivation checks while I manually reviewed samples of the bank's problems, keys and verdicts. In the writing of this paper it handled the Chinese-to-English translation, English polishing and the expansion of early drafts for which I had written the outline and the core narrative. The scientific content, the structure, the direction of the experiments, the choice of every number and the final wording are my own.

- *DeepSeek V4-pro / V4-flash* (July to November 2026, called programmatically in per-experiment batches, with platform usage records submitted with the materials): the solving teachers and the grader, objects of research rather than writing tools.

The research log is on Yuque and kept by me. Entries I wrote and entries written with AI assistance are interleaved, and I reviewed every entry. Every AI output was verified by me before use. For the committee's inspection, the submitted materials contain a representative selection of the records rather than every session. They include passages from working sessions across the project period, each reproduced as a complete two-way exchange, log excerpts and process snapshots for the other stages, platform usage records, and repository commit history. Together they cover almost every tool and every stage declared above.

2026 S.-T. Yau High School Science Award
仅用于2026丘成桐中学科学奖公示