More Fitting, Limited Transfer

Abraham September 2026 · Exploratory research report · Version 1.0

Abstract

Can additional answer-only fine-tuning rescue weak learning on a small arithmetic task? We trained three adapter initializations of Qwen2.5-1.5B-Instruct at cumulative token budgets of 1×, 4× and 16×. Mean training accuracy rose from 16.67% to 52.34% to 79.56%. Development accuracy on new starting values for trained functions rose more modestly: 6.77%, 18.23% and 21.35%. A separately reserved final set improved from 3.125% at base to 21.35% at 16×. The result establishes substantial fitting under this recipe and some within-function transfer, but not general algorithm learning or a benefit from social teaching. Increasing exposure produced diminishing transfer returns and worsening development likelihood relative to lower trained doses.

1. Question and design

The task asks a model to execute arithmetic instructions involving addition, subtraction and multiplication. Each program implements an affine function, a × start + b. Training used 256 computations and answer-only supervision: no explanation targets, teacher interaction or reinforcement learning.

Three independently initialized adapters, with seeds 11, 23 and 47, followed the same cumulative dose schedule. The 1×, 4× and 16× labels denote processed-token budgets, not epochs. Each training example appeared approximately 4-5 times at 1×, 17-18 times at 4× and 71-72 times at 16×. Checkpoints are dependent stages of the same three trajectories, not independent training experiments.

Development and final sets each contained 64 computations: four new starting values for each of 16 trained functions. The two sets were disjoint and excluded computations allocated to the preceding pilot and diagnostic. Both used familiar functions and training-style operation patterns. This is not a test of unseen-function generalization or a claim that the base model could never have encountered similar arithmetic during pretraining.

The final set was evaluated only at base and 16×, as fixed before observing the curve. Generated answers were graded deterministically. Malformed or length-capped answers remained failures in the denominator. The registered primary descriptive contrast was 16× minus 1× training accuracy; no confirmatory significance test was specified.

2. Accuracy

Values are means over three adapter seeds, except the single base evaluation.

Exposure	Training accuracy	Development accuracy
Base	5.86%	6.25%
1×	16.67%	6.77%
4×	52.34%	18.23%
16×	79.56%	21.35%

At 16×, training scores were 199, 211 and 201 out of 256. The primary 16× minus 1× contrast averaged +62.89 percentage points, with sample standard deviation 1.70 points across seeds. That standard deviation is not a confidence interval.

On the separately reserved final set, base answered 2/64 correctly. The 16× adapters answered 12/64, 15/64 and 14/64 correctly, averaging 21.35%: a descriptive improvement of 18.23 percentage points. All three seeds improved over base, but absolute performance remained low. Development and final means happen to match; their items and predictions are distinct.

Only one final output was malformed and length-capped. Most final failures were wrong answers rather than formatting failures. The cap and malformed output were the same case, not two separate failures.

3. Transfer and answer turnover

The training-development accuracy gap widened from 9.90 points at 1× to 34.11 at 4× and 58.20 at 16×. A post-hoc paired check found that increased exposure did not uniformly preserve earlier correct development answers:

Seed	Newly correct at 16× versus 4×	Previously correct, now wrong	Net gain
11	6	6	0
23	9	8	1
47	8	3	5

These counts describe the same 64 tasks within each seed; they must not be pooled as independent trials. At 16×, no adapter answered all four final starting values correctly for any of the 16 functions. This argues against claiming reliable acquisition of those arithmetic rules. It does not prove pure memorization: final computations were absent from training and some improvement transferred.

4. Efficiency and contrary evidence

Cumulative processed training tokens across seeds were 299,844 at 1×, 1,199,973 at 4× and 4,799,924 at 16×. From 1× to 4×, about 0.90 million additional tokens accompanied an 11.46-point development gain. From 4× to 16×, about 3.60 million additional tokens accompanied only a 3.125-point gain. The latter interval yielded approximately fifteen times less marginal development improvement per processed token. This is a descriptive ratio, not a scaling law or a monetary cost estimate.

Development numeric negative log likelihood, where lower is better, averaged 1.258, 1.319 and 1.966 at the three trained doses. It worsened from 4× to 16× in every seed, despite the small mean accuracy gain. The 16× value remained better than untrained base, 2.427. This likelihood measure conditions on correct preceding answer tokens; it is not an independent confidence-calibration score or the probability of generating an entire correct answer unaided.

The completed run used 54,888 training updates, 435,029 supervised tokens, 3,456 answer generations and 3,456 teacher-forced scoring operations. Cumulative checkpoint costs must not be summed as independent runs. A previous interrupted attempt is preserved separately and was not pooled; it consumed additional resources not included in these completed-run totals.

5. Interpretation and limitations

The pipeline can substantially improve fitting with additional answer-only exposure. A weak low-dose result should therefore not be interpreted as an inability to learn these training tasks. However, higher fitting accuracy is not equivalent to broad transfer. Widening train-development gaps, answer turnover and worsening development likelihood are consistent with specialization or overfitting, without isolating a unique mechanism.

This is an exploratory calibration with three training seeds, one model and one synthetic task family. Repeated checkpoints and tasks are dependent. No p-value, optimal-dose claim or general significance claim follows. In particular, final 4× accuracy was never measured, so 4× cannot be declared the best final-test dose.

For the broader social-learning program, this supplies an answer-only baseline. A future comparison could ask whether verified teacher explanations improve fresh-task transfer at matched training budgets, while separately accounting for teacher costs. No social treatment was tested here.

Evidence and reproducibility

A read-only integrity audit verified the frozen candidate, operation ledger, saved adapters, accounting and deterministic replay. All 27 fitting tests passed. A separate post-hoc analysis regraded recorded generations and matched all 24 original metric groups. These checks establish consistency of recorded evidence, not independent regeneration of every model output.

Model: Qwen/Qwen2.5-1.5B-Instruct, revision 989aa7980e4cf806f80c7fef2b1adb7bc71aa306. The immutable original analysis has SHA-256 de32e9cfe0e75da88ab85cf360a536f7ba520a9600c828e10e5bbacd1181d698. Public replication materials are not attached to this report. No claim of novel methodology is made.