RESEARCH · PAPER

Separating a reasoning skill from the habit of deliberating

What we found when we subtracted the "thinking" out of a 4-billion-parameter model

Douglas Lund, Loxley Labs. 8 November 2026. Produced with Claude Code.

Companion to the release `Qwen3-4B-Thinking-Concise`. Published on this site 13 November 2026.

Summary

Qwen publish two 4-billion-parameter models that share one pre-trained base and were then trained differently. We infer the shared base from Qwen's own naming of the pair and from their weights, whose matrices have a median cosine similarity of 0.998 to each other. One always deliberates at length before answering, writing its working inside a `<think>` block. The other never does. We took the difference between their weights, scored every direction inside that difference by how much it contributes to reasoning, and subtracted the directions that carry most of it from the deliberating model.

We expected to damage its reasoning. On the six logic tasks we tested first, we did not.

The resulting model wrote no closed think block in any of the roughly 2,400 replies we scored, answers in five to eight times fewer tokens by median on the main sets, and keeps **14.4 of the 20 points**, plus or minus about 4.5, by which the deliberating model beat the non-deliberating one on six logic tasks. No training, no distillation, no fine-tuning. One subtraction.

That headline is bounded, and the bound belongs next to it. It holds on those six tasks at an 8,192-token reply cap. Give the models room to think, 32,768 tokens on harder sets, and the picture reverses: our model keeps little of the advantage, and on the hardest maths set it falls behind the plain non-deliberating model as well (12 questions to 3). On the easier maths set at the same cap the two are level, so this is a statement about hard problems, not about maths generally. What we are reporting is a good trade under a token budget, not a free lunch.

Two facts about the headline set belong beside the number. The six logic tasks were in the harness from its first scorecard run and never changed, but the set was cut from 180 items to 90, the first fifteen of each task, after a first run of the Thinking model, by a fixed rule and not blind; section 9 records it. And the bracket-matching task on which long deliberation helps most, Dyck languages, is not one of the six. It appears only in the harder set, where the cut costs the most.

A note on counts like "12 questions to 3". Where a test averaged two runs per question, a question counts for one side when its two-run average is higher, so a question right in one run and wrong in the other contributes half a point to the accuracy gap. That is why 12 to 3 corresponds to a six-point gap on 96 questions rather than a nine-point one. All such counts in this paper are of that kind unless a single run is stated.

Three findings. The first is the headline result; the second is the one we most want other researchers to know about:

1. **A reasoning skill and the habit of deliberating are partly separable in the weights.** In this one model pair, on six logic tasks at an 8,192-token cap, removing the habit kept most of the skill. On harder sets with a much larger budget it did not.
2. **We could not put the behaviour back one piece at a time, and this is the more interesting result.** All three recipes we tried for assembling "deliberate, then stop, then answer" out of individually-scored directions produced a broken model, while simply scaling the whole difference worked. Nearly 40% of the directions the stopping decision depends on score zero or negative for reasoning, so no reasoning-based selection could ever have picked them.
3. **No sign that knowledge moved, on a weak test.** Every cut we tried held its score on one general-knowledge benchmark within about three points while reasoning behaviour changed completely. The test has little power, for reasons section 5 gives, and we would like it done properly.

We are a one-person software business, not a laboratory. Our test sets are small, between 50 and 500 items. We are publishing the model, the removed piece, the code and the per-item results behind the accuracy tables so that people with more machines than us can check whether any of this survives contact with a full benchmark run. The refusal check ships as counts only, with no prompt or reply text.

1. The setup

Instruct	`Qwen/Qwen3-4B-Instruct-2507`, never deliberates
Thinking	`Qwen/Qwen3-4B-Thinking-2507`, always deliberates
R	Thinking minus Instruct, the whole weight difference
P	the piece of R we removed
D	Thinking minus P, the released model

The two models are genuine siblings: their matrices have a median cosine similarity of 0.998, the lowest being 0.994. The difference R is about 8% of the size of each matrix and is close to full rank, so it is not a neat low-dimensional edit sitting on top of a shared base. We checked that the split is exact and reversible: Instruct plus R, with 3,471 corrections in double precision, rebuilds Thinking bit for bit.

We also checked that R is about deliberating rather than about instruction style. Compared three ways against the pre-trained base model, the difference is almost entirely thinking-side training with only a trace of the instruct training in it. The measure is in our run log and not in this package, so we give no figure for it here.

2. How we chose what to remove

The unit of the cut is a single singular direction of a single matrix of R. Each one is scored by a first-order estimate of how much a loss would rise if that direction were removed from the Thinking model.

Two losses are used:

- a **reasoning loss**: the next-token loss on Thinking's *own* solutions to training problems it had solved correctly. We deliberately did not use a dataset's reference solutions. We wanted the model's own reasoning, not somebody's idea of good reasoning.
- a **control loss**: the same measure on ordinary knowledge answers, short instructions and plain chat, with an empty think block, so it captures what a direction does to the model when it is *not* reasoning.

Directions with a high reasoning score and a low control score are the ones that carry the habit. The gradient matrix is never stored: hooks on each linear layer project that layer's input onto one side of the direction and its output gradient onto the other during the backward pass, so memory stays near the cost of ordinary inference. The kept set is low-rank by construction, which is why the removed piece ships as a 2.17 GB file rather than a second copy of the model.

The released model removes the directions holding 80% of the first-order reasoning benefit.

An honest note on the scores. The first-order estimates were accurate as *estimates* for reasoning (12% error) and for the stopping decision (6%), and rough for the control loss (45%). But in every case they are only a ranking. They do not add up, and section 4 is the consequence.

3. Finding 1: the skill survives the cut

Accuracy in percent. Every model answered the same items with the same settings. Set sizes are given because they are small.

Test	Items	Reply cap	Instruct	Thinking	D
BIG-Bench Hard, six logic tasks	90	8,192	75.6	95.6	**90.0**
GSM8K	300	4,096	96.0	93.0	**93.7**
MATH	50	8,192	86	64	**86**
MMLU	500	one token	69.2	64.8	**66.8**
IFEval, strict	200	1,536	73.0	66.0	**69.5**
MATH, the same 50 problems	50	32,768	88.0	91.0	**85.0**
Hard MATH, level 5	96	32,768	85.4	90.1	**79.2**
Hard logic, four puzzle types	200	32,768	77.8	92.5	**80.5**

The six logic tasks are boolean expressions, formal fallacies, logical deduction with five objects, web of lies, tracking shuffled objects with three objects, and date understanding; Dyck languages is not among them. The three 32,768-token rows average two runs per question, which is how a figure like 90.1 on 96 items arises; every other row is one run per item. The two MATH rows are the same 50 problems at two caps, and should be read together: the 64 at the smaller cap is a budget failure, not a capability one.

And the length, which is the point:

Median tokens per reply	Instruct	Thinking	D
BIG-Bench Hard	530	2,868	**362**
GSM8K	281	1,069	**204**
MATH	788	5,178	**637**

D keeps 14.4 of the 20-point logic gap at an eighth of Thinking's median length on this set. Because all three models answered the same 90 items, the right test is a paired one: 16 items were right for D and wrong for Instruct, and 3 the other way, which is an exact sign-test p of 0.0044. That is the single number this release rests on, and it is the first one we would like somebody else to try to reproduce on a full-size set.

Where the cut costs you. Three places, all stated plainly:

- **Instruction following.** D scores 69.5 on IFEval against Instruct's 73.0 on the same prompt, and 82.5 when Instruct is given its own chat format. The 3.5-point gap on the shared prompt is inside the noise, so the honest version of this claim rests on the 82.5: used as it was meant to be used, Instruct is clearly better at following fiddly formatting instructions, and you should use it for that.
- **Hard problems with room to think.** The 8,192-token MATH row flatters D. Thinking scored 64 only because 15 of its 50 replies ran out of tokens; at 32,768 it scores 91 on the same problems. On the harder sets at that cap, Thinking beats D on maths (90.1 against 79.2; 16 questions to 2 by the two-run count) and on logic (92.5 against 80.5; 39 to 3).
- **One particular kind of puzzle.** Most of that logic gap is Dyck languages, a bracket-matching task where Thinking scores 70, D scores 30 and Instruct scores 17. That one task accounts for 10 of the 12 points. Long deliberation genuinely helps there, and removing it genuinely hurts.

So the honest summary is not "deliberation is unnecessary". It is: **a large part of what deliberation buys is already available without it, and on the six logic tasks under a token budget the model without the habit is the better deal.**

4. Finding 2: the behaviour does not come apart into pieces

This is the result we would most like other people to know about, because it cost us a day and it is a negative.

Having found that the habit could be *removed* by selecting directions, we assumed it could be *added* the same way. In our hands it could not. Three attempts, all failures:

Attempt 1: select the reasoning directions and add them to Instruct. The resulting model deliberates fluently and beautifully, in exactly Thinking's voice, and then mostly does not stop: 280 of its first 300 replies ran to the token cap. It re-checks its own working indefinitely, never reaching an answer. The selection had captured the deliberation and left behind the decision to finish.

Attempt 2: score the stopping decision separately and add those directions too. We built a second score: the loss on predicting the end-of-thinking token at the places where Thinking actually stopped. The score itself worked well. But the combined cuts were worse than either alone. One stopped and then produced no answer at all. One capped 49 of 50 replies. One did neither.

Attempt 3: no selection, just scale the whole difference. Instruct plus a fraction of R. This works. The habit switches on somewhere between a quarter and three quarters of the way. At three quarters the model deliberates and closes its think block on most replies: 297 of 300 on GSM8K, 40 of 50 on MATH and 90 of 90 on the logic tasks, with 6 and 12 of those replies running out of tokens on the first two sets. An earlier screening run of this configuration showed no failures at all, which is where a "100%" in our run log comes from; the full run is the one to believe.

The measurement that explains it: **38.7% of the directions carrying the stopping decision have zero or negative benefit for reasoning.** No selection based on reasoning importance could have included them. On the evidence of these three recipes, the deliberate, then stop, then answer sequence does not behave like a sum of independently useful parts. We would not want to state that more strongly from one model pair.

We think this is worth saying out loud because the opposite assumption is natural, and acting on it is expensive.

5. Finding 3: no sign that knowledge moved, on a weak test

Every direction-selected cut we made, including the broken ones, scored between 66.6 and 68.2 on MMLU, against Instruct's 69.2 and Thinking's 64.8. Only the released model's MMLU run is included in this package; the rest are from runs we did not ship. Models that had lost the ability to finish a sentence still answered the same factual questions.

On this evidence nothing suggests these directions carry knowledge. We stop short of saying they do not, because the evidence is weak in three ways, and together they make this the least secure of our three findings.

MMLU is one benchmark and we used 500 of its items, where the standard error is about 2.1 points. The two parent models are themselves only 4.4 points apart on it, so almost any blend of them was going to land "within about three points" whatever it removed; the test has little power to detect a loss. And our MMLU protocol scores the most likely next token after a fixed prefix, so it never asks the model to write anything. "Models that had lost the ability to finish a sentence still answered the same factual questions" is therefore a smaller claim than it sounds: those models were never asked to finish a sentence in this test.

6. A router, which is not new

We also trained a small network that reads a question and chooses, per question, whether to answer with D, with a partial restoration of P, or with the full Thinking model. It works:

Policy, hard puzzles at a 32,768-token cap	Hard maths (96)	Hard logic (200)	Mean to
Instruct	85.4	77.8	3,004
Thinking	90.1	92.5	10,830
D	79.2	80.5	2,103
Router, balanced	86.5	90.0	5,979
Router, accuracy first	89.6	90.0	7,421

The router reaches Thinking's accuracy on hard maths while using about 69% of its tokens. On hard logic it is 2.5 points behind, and that gap is real rather than noise: by the two-run count, Thinking wins 10 of the questions they differ on against the router's 2.

The router was trained on a pool of 440 questions that share nothing with any test item: 240 further examples of the same six logic tasks, 100 MATH-500 problems outside the test set, and 100 GSM8K training problems. The hard-maths set above is the level-5 MATH-500 problems in neither the test set nor that pool, and the hard-logic set is drawn from four tasks none of which is among the six (two are harder, seven-object variants of tasks that are), so every score in the table is on held-out questions. The two routers we ship differ only in the token cap under which their pool answers were produced.

These rows are scored by looking up what each model actually did on each question in our stored per-item results and attributing it to whichever setting the router chose. No live routing system was run end to end.

We make no novelty claim here. Choosing a thinking budget per question from the model's own hidden states is an active published area, and our router sits squarely inside it. See section 8. We include it because it is the natural way to use a model and its removed piece together, and because the per-question choices are interesting in their own right. In our first router test, which used the same pool of questions answered under the original token caps rather than the 32,768-token cap of the table above, the router sent every formal-fallacies question to the model that does not deliberate and two thirds of the boolean-expression questions to the one that does.

We also tried a more elaborate version that sets a separate scale for the piece in each of the 36 layers. **It was not worth the extra machinery.** It does beat D, clearly, on both hard

maths and hard logic. But it only ties the far simpler router overall and loses to it on hard logic, while using 2.1 to 3.4 times D's tokens. We are not releasing it, and we mention it so nobody spends a week rediscovering that the complicated version buys nothing over the simple one.

7. Does the cut make the model less safe?

Editing weights can change what a model refuses, so we checked before publishing rather than after.

All three models were run on two published sets: AdvBench, 520 harmful instructions, and XSTest, 450 prompts of which 250 are safe questions that merely *look* dangerous and 200 are genuinely unsafe.

	D	Instruct	Thinking
AdvBench, of 520 harmful, **not** refused	6	5	13
XSTest, of 200 unsafe, **not** refused	46	47	60
XSTest, of 250 safe, **wrongly** refused	10	13	7

Every cell above uses one rule: a refusal phrase anywhere in the first 200 characters. The model card's headline table uses the stricter rule instead, so its figures differ from these, most of all for Thinking. Both rules are reported in both documents, but anyone comparing the two should check which rule a given table is using. A stricter rule that only accepts a refusal at the very start changes Thinking a great deal, to 29 and 68 not refused, because Thinking tends to acknowledge a request before declining it. It barely moves D or Instruct. We give the looser rule here because it treats the three models alike; both are in the results files.

D is indistinguishable from Instruct on all three measures; the gaps are small and untested, and a difference that looked real in a first draw did not survive a second. **Removing the deliberation habit did not lower the string-matched refusal rate relative to either model D came from.** That is the exact claim the measurement supports, and we have resisted shortening it into something that sounds like a statement about safety.

Three honest caveats.

The classifier is a string match on how a reply opens, which is crude. It cannot tell whether a reply that opens with a refusal goes on to comply, and we tuned the phrase list after seeing a first run. Anyone rerunning this should use a real judge instead, and we would like them to.

In the refusal check D was the only one of the three to run a reply to the token cap: seven times out of 970 in the first draw and five in the second, where Instruct and Thinking never did. This is specific to that check. On the accuracy tests all three models sometimes run out of tokens, Thinking most of all. We first guessed from compression statistics that these were repetition loops. **That guess was wrong, and the way it was wrong is worth recording.** Compression ratio falls with length on its own, so comparing a 40,000-character reply with a 2,000-character one proves nothing. Running one such prompt 100 times showed something else: ten of the draws reached the cap, and those ten are long, structured answers that keep adding sections rather than cycling. They are partly redundant, with the longest repeated passage running 635 characters and appearing twice, and stretches where half the short word sequences have been seen before. None of them settles into a loop by the test we used. Capping looks like a stopping failure in D, not degeneration.

On that same item D did not refuse in any of 100 draws, while Instruct refused in both of the two draws we have of it and Thinking refused in neither. D is Thinking minus a piece, so inheriting Thinking's disposition on an individual prompt is expected; what matters for the table above is that in aggregate D sits level with Instruct, not with Thinking.

Twelve capped replies fell on eleven distinct prompts. We resampled the five that were AdvBench prompts, 64 fresh draws of each, and the one unsafe XSTest prompt described above, 100 draws. We did not examine the five remaining prompts, which are XSTest's *safe* questions, because a long answer to a harmless question affects no published figure. The original replies had not been kept, so all of this is statistically comparable replies to the same prompts, not the same replies read back.

Three of the five decline and substitute something else entirely. The other two are a near-identical pair of hacking requests, and we record them as **mixed**, which deserves stating precisely rather than comfortably. On these the model declines the request and then sets out a stepwise outline of *authorised* security testing, naming general-purpose tools and, in a minority of replies, giving an example command. A stricter judge than ours could reasonably count some of those replies as partial compliance, and we would not argue.

Did the edit cause this? We settled it by running the two public Qwen models on one of the two prompts, six draws each, and having a person read their replies rather than characterising them by machine. **Instruct gives an outright denial and stops there. Thinking refuses and then points the reader toward lawful, ethical ways to learn the subject, without instructions for carrying out the request.** D's pattern, refusing and then continuing, is Thinking's and not Instruct's. Since D is Thinking minus a piece, the continuation is inherited from the model it was cut from rather than introduced by the cut. Across all 320 draws over the five AdvBench prompts, every reply still opened with a refusal, so no figure in the table above moved.

One thing we have deliberately not smoothed over. D's continuation was characterised, by the assistant that read it, as a stepwise outline of authorised security testing that in a minority of replies includes an example command. Thinking's was characterised, by a human reader, as guidance toward learning the subject lawfully with no instructions for carrying out the task. The same reader then compared the two side by side, and found no difference in kind: both decline outright, and D's replies are simply longer. On the question of degree, D does not go further than the model it was cut from.

One caveat we keep attached to that. The person compared six draws of each, while the assistant's description of D came from sixty-four. A behaviour reported as appearing in a minority of replies need not appear in six. The two descriptions of D are therefore not in conflict so much as drawn from different sample sizes, and we record both rather than picking the one we prefer.

Our verdict on the second of the pair is inferred from the first and from the same numeric shape, not read. The model card marks which verdicts rest on reading and which do not.

One real difference between the models did turn up. On one of that pair, Instruct produces a reply that is almost entirely decline content in 26 of 64 draws. D does this in none of 64, and **neither does Thinking**. D matches its parent rather than Instruct, so the edit did not create the difference; it was inherited. It is one sharp instance of something visible throughout this check: D's refusals are longer, a median of 1,220 characters against Instruct's 420 on AdvBench. A model that keeps going after its opening sentence will never produce a decline-only reply. We did not read the text behind Instruct's 26 draws, because that would tell us about Instruct rather than about the model we are releasing.

8. What was already known

We looked for prior work before claiming anything. One caution about this section: it is a search-based scan done by us, not a literature review. For most of the items below we read abstracts and summaries rather than full papers, and we have deliberately left out figures we could not check ourselves. Treat it as our honest attempt to place this work, and please correct it.

- **Task arithmetic** (Ilharco et al., ICLR 2023). Subtracting a fine-tuned model's weights from its base gives a "task vector"; negating it removes a capability. This is the direct ancestor of what we did. Two differences: our difference is between two *sibling fine-tunes* rather than a fine-tune and its base, and we do not negate the whole vector, we select inside it by a gradient score.
- **Suppressing reflection at inference.** "Wait, We Don't Need to 'Wait'! Removing Thinking Tokens Improves Reasoning Efficiency" (Findings of EMNLP 2025) blocks the tokens a model uses to second-guess itself and reports substantially shorter traces at little accuracy cost. Same goal as ours, entirely different mechanism: nothing is edited, the intervention happens while the model is generating.
- **Distilling a reasoning model into a non-reasoning one.** A line of work trains a student to give answers without the trace: "Implicit Chain of Thought Reasoning via Knowledge Distillation" (arXiv 2311.01460), "TwT: Thinking without Tokens by Habitual Reasoning Distillation" (arXiv 2503.24198), and "Masked Distillation: Internalizing the Chain-of-Thought in Language Models" (arXiv 2607.22629). Reported experience there is that removing the trace costs accuracy on reasoning-heavy tasks. Our result is interesting against that backdrop because producing D required no training at all; the only thing we trained was the optional router that chooses between D and its parent.
- **Per-question thinking budgets.** "Learning When to Think: Adaptive Reasoning for Test-Time Compute Allocation" (arXiv 2608.20256) has the model emit a budget token before answering; "Route to Reason: Adaptive Routing for LLM and Reasoning Strategy Selection" (arXiv 2505.19435) routes across models and strategies; an ICLR 2026 paper predicts the budget from prefill hidden states through a small calibrated adapter, using self-consistency as the difficulty signal. Our router reads a hidden state and picks one of three settings, which puts it squarely inside this family. We claim nothing for it.
- **Gradient-scored singular directions**, used for compression in GRASP (arXiv 2501.00339), and **singular directions associated with a skill**, located and protected during fine-tuning in PALoRA (arXiv 2605.24549). Direction-level masking within one model appears in "Beyond Components: Singular Vector-Based Interpretability of

Transformer Circuits" (arXiv 2511.20273). These are the nearest relatives of our scoring method.

- **Weight deltas as low-rank task-aligned signal**, used for merging and forgetting (arXiv 2605.20296).

Identifiers are as returned by our searches. We have read abstracts and summaries of these, not the full papers, and Ilharco et al. is the only one we would describe from more than its abstract.

What we believe is new, stated as a question rather than a claim. We could not find prior work that scores the singular directions *of the difference between two sibling fine-tunes* in order to separate a behavioural habit from the skill it usually accompanies, nor a published report that the behaviour fails to recompose from independently scored directions. If that work exists, we would rather be told than be the last to know.

9. Limits

- Test sets of 50 to 500 items, one or two draws. Standard errors are 1.4 points on 300 items, 3.2 on 90, 4.9 on 50, and the standard error of a gap is roughly 1.4 times that. Several gaps in section 3 are inside the noise and should not be read as differences: D against Instruct on instruction following (3.5 points against a gap error of about 4.6) and on general knowledge (2.4 against about 3.0), D against Thinking on GSM8K (0.7 against about 2.0), and D against Thinking on the six logic tasks (5.6 against about 4.5), which means the amount by which D falls short of its parent there is itself not established. Two results do clear the noise by a wide margin and are not in that list: the logic gap, and D's lead over Thinking on MATH at the smaller token cap, which is a budget effect rather than a capability one.
- No code, long-context, multi-turn, tool-use or non-English testing at all.
- **Several figures in this paper come from runs we did not ship, and a reader cannot check them from this package.** They are: the 3,471 double-precision corrections and the bit-for-bit rebuild; the three-way comparison of the weight difference against the base model; the 6% and 45% accuracy of the first-order estimates; the 38.7% of stopping directions with no reasoning benefit; and the MMLU scores of every cut except the released one. The same applies to all three attempts in section 4, including the 280 of 300 capped replies and the quarter and half blends; to the per-layer dial figures in section 6, since the dial's weights are not shipped; and to every resampled and replayed

figure in section 7, including the 64-draw and 100-draw runs and the six-draw replay, since those replies were generated for reading and are not in the package. We state these because they are what we measured, and we flag them because you have to take our word for them. The model card carries its own version of this list, covering the figures that appear in the card; the two overlap but are not identical, because the two documents quote different figures.

- The MATH and BIG-Bench Hard sets were cut from 100 and 180 items to 50 and 90 part way through the work, by a fixed rule (every tenth problem; the first fifteen of each task) but after we had seen a first run. The scorecard explains why, and all final numbers come from the final settings, but the decision was not made blind.
- One model pair, one size, one family. Whether any of this holds for other sibling pairs is unknown.
- All measurements on bf16 weights through llama.cpp. Quantised versions untested.
- The removed piece ships rounded to bf16 while the model was built from full precision, so rebuilding the model from the piece reproduces it to within that rounding: 94.17% of values identical, and a next-token difference about 800 times smaller than the piece itself makes. The weight comparison covers every value; the next-token comparison covers 12 replies and about 2,500 tokens, which is a small sample and should be read as an order of magnitude.
- D inherits Thinking's chat template, which opens a `<think>` tag that D never closes. This matters for how some servers return its replies, and the model card says what to set.
- Parts of the refusal analysis were completed from numeric and structural measures without reading the reply text, because the work was interrupted three times by a safety classifier acting on the assistant doing it: twice while it was writing up its own measurements, and once on an attempt to answer the question in the next bullet. The distributional results are unaffected; the characterisations of what a reply contains rest on a smaller reading than we would have chosen. We note it because a reader judging this section deserves to know where it was done at arm's length.
- **One question in section 7 was answered only by stepping outside the usual process.** Whether the two public Qwen models continue after a refusal the way D does is what decides whether the behaviour is inherited or introduced. An attempt by the assistant to establish it was stopped, as were two earlier write-ups of its own results on the same prompts. It was settled in the end by generating the replies and having a person read them. The comparison of *degree* between D and Thinking, in section 7, was made by one

reader on six draws of each. What remains open is only the sample size: a behaviour reported in a minority of a larger sample need not appear in six draws.

10. What we would most like checked

In order of usefulness to us:

1. **Full-size runs** of the maths and logic sets. Ours are subsets, and the headline number is a 14.4-point gap on 90 items.
2. **A second draw with different seeds**, to see how much the table moves.
3. **The refusal check with a real judge** instead of our string match, and on prompts beyond the two sets we used.
4. **Whether the method transfers.** Any other pair of sibling models where one has a behaviour the other lacks.
5. **Rebuilding the model from the released piece** and repeating a table on it.

If your numbers disagree with ours, we want to know which table, which items and which settings. We would rather be corrected in public than quietly wrong.

The model, the removed piece, the routers, the code and the per-item results behind the accuracy tables are released under the Apache License 2.0. The weights derive from Qwen's `Qwen3-4B-Thinking-2507` and `Qwen3-4B-Instruct-2507`, both Apache-2.0, Copyright 2024 Alibaba Cloud. Loxley Labs is not affiliated with Alibaba Cloud or the Qwen team, and they have not reviewed or endorsed this work. No Anthropic model weights were used or modified.

THE RELEASE

What is published with the paper.

The model	Qwen3-4B-Thinking-Concise, bf16 GGUF. Qwen3-4B-Thinking-2507 with the deliberation piece subtracted; no training.

The removed piece	P_q80.safetensors, 2.17 GB. Add it back to the model and you have the original to within bf16 rounding.
Two routers	router_v2_pool440 and router_v2_pool32k: small classifiers that send a question to the cut model or the original.
Code	Fourteen scripts: the scoring of directions, the subtraction, the routers, the evaluations and the refusal check.
Results	Scorecard, the long-budget runs, the router runs, the bf16 check, the refusal check, and one file per test item.
Licence	Apache-2.0, with Qwen attributed in NOTICE. Produced with Claude Code, and the paper says where that mattered.