When Fine-Tuning Is Not Enough

EXA Automotive AI Research — Engineering Research Edition v1.3

Date: 16 November 2026 **Status:** Publication edition. Not human peer reviewed or production approved.

This edition reports scientific methods, configurations, results and limitations. It omits proprietary data contracts, prompts, code, product design and raw research records. Case descriptions are scientific paraphrases of observations; the explicitly labeled illustration is newly written fiction. No claim of exact public reproducibility is made.

Abstract

Question. Does LoRA adaptation improve a small model's interpretation of changing automotive requirements under a shared evaluation format?

Method. A base Qwen3.5-4B model and a fresh adapter were compared on 24 held-out and 12 known regression cases. Training used 172 AI-generated examples and 129 optimizer updates, taking 1,433.76 seconds with 8.87 GiB peak allocated GPU memory. Separate four-bit Qwen3-14B teacher probes reused four selected, teacher-forced turns.

Results. Held-out checker-gated record agreement changed from 19/24 to 20/24; excluding one disputed reference, both arms scored 19/23. Across 36 cases, exact next-step-label agreement changed from 11/36 to 27/36. Budget agreement changed from 8/8 to 6/8, including two observed tenfold errors. Teacher run 2 followed revisions informed by run-1 failures on the same turns and still failed clarification checks.

Interpretation. Training completed and fixed-sample outcomes were mixed. Label agreement is not independently validated conversational quality; small positive and negative differences do not establish population effects. Training-data rights and human reference adjudication have not been independently verified.

Keywords: language-model adaptation; automotive language; LoRA; synthetic evaluation; numerical errors; structured output; teacher–student learning.

1. Scientific scope

The **base model** is Qwen3.5-4B 2 without the experimental adapter; the **adapted model** is the same foundation with that adapter. The experiment studied interpretation of changing requirements expressed in language. It examined whether additional training improved agreement with AI-generated reference answers, and whether syntactically valid outputs could still contain consequential mistakes.

This is a task-level research account limited to experimental methods, observations and scientific interpretation.

The experiments used fictional vehicle evidence rather than establishing real vehicle specifications. Outputs and scores concern controlled AI-generated cases. The public methodology below is deliberately abstracted; internal representations and decision rules are not disclosed.

1.1 Explanation without an internal schema

Illustration — fictional and not an experiment record: A shopper first expresses a color preference and later changes it. A correct interpretation should recognize the new preference. This natural-language example illustrates context dependence; it does not reproduce a data field, serialized output, algorithm or product rule.

The scientific distinction is between an output's format and its meaning. An output may satisfy a requested format while disagreeing with a reference answer or making an unsupported assertion.

2. Data and evaluation design

2.1 Dataset sizes and provenance limitations

The training set comprised **172 AI-generated examples, including their customer messages and target reference answers**, according to the inspected provenance records. No human authorship of these examples is claimed. Their use in a completed historical experiment is described for research transparency. Evaluation comprised 24 held-out cases and 12 known regression cases.

Split	Examples	Scientific role
Training	172	Adaptation targets
Held-out evaluation	24	Excluded from training
Regression evaluation	12	Previously examined behavior

Automated checks screened for duplicate identifiers, overlapping scenario families, duplicate inputs and near-matching dialogue across training and evaluation. Passing those checks does not exclude paraphrase leakage, shared annotation assumptions or other semantic dependence.

The inspected provenance records also identify **all 36 evaluation cases (24 held-out and 12 regression), including their customer messages and reference answers, as AI-generated**. Independent human adjudication was not established. These are provisional AI-generated references, not certified ground truth. Shared AI generation of training targets and evaluation labels can create correlated stylistic conventions and assumptions.

One reference omitted a defensible newly expressed requirement: the base model recorded it, while the adapted model omitted it despite referring to it in prose. Both chose the same fictional option. The reference's omission therefore favored the adapted record. This label dispute is quantified in section 5.2; no reference or historical score was altered.

2.2 Language coverage

Language label	Training	Held-out
English	164	23
Hinglish	6	0
Tamil-English	2	1
Total	172	24

These are dataset labels, not independent language-quality assessments. The small code-mixed sample cannot establish multilingual reliability.

2.3 Coverage and selection limitations

The examples covered numbers and units, changed instructions, missing information and meaning across multiple messages. These broad descriptions replace the internal category taxonomy; they are not a data-contract catalogue.

The regression cases informed earlier development and are not newly unseen independent tests. Both split results and pooled workload counts are reported below.

3. Smaller-model training

3.1 Model and adaptation configuration

The experiment used LoRA 1 with the configuration below.

Item	Recorded setting
Foundation model	Qwen/Qwen3.5-4B
Public model revision	851bf6e806efd8d0a36b00ddf55e13ccb7b8cd0a
Adapter initialization	Fresh adapter
LoRA rank	16
LoRA alpha; alpha/rank	32; 2
LoRA dropout	0
Adapted projections	Query and value attention projections
Adapted attention layers	8
Trainable tensors	32
Trainable parameter elements	1,835,008
Epochs	3
Optimizer updates	129
Batch size	1
Gradient accumulation	4
Optimizer	AdamW 6
Learning rate	0.0001
Warmup	8 steps
Later schedule	Constant learning rate
Seed	42
Gradient checkpointing	Enabled
Gradient clipping norm	1.0
Weight decay	0
Optimizer betas	0.9 and 0.999
Optimizer epsilon	0.00000001

These settings describe a model-training experiment. They do not disclose application infrastructure, deployment design or a product decision algorithm.

3.2 Numerical path and sequence lengths

The model-loading configuration used FP16 weights with four-bit and eight-bit loading disabled. The recorded execution required an Unsloth FP32 activation path, with automatic mixed precision disabled.

Preprocessing enforced a 1,024-token complete-sample cap. Observed training sequences contained 612–874 tokens and were not truncated. Saved trainer metadata also records a 1,536-token length setting. This distinction prevents the preprocessing cap from being mistaken for the sole runtime length setting. None of the observed training sequences approached either limit.

The recorded hardware was a Tesla T4 with 15,360 MiB reported memory. The experiment recorded Python 3.13.16. These are runtime observations, not minimum resource guarantees.

3.3 Loss objective and execution evidence

Training used response-only loss masking: prompt tokens were excluded, while the assistant continuation, including output termination tokens, was supervised. The source asserts an exact token boundary, and the archived token counts record positive masked and target lengths for all 172 examples. This is source and saved-run evidence, not an independently repeated tokenizer or GPU execution.

The same instruction and model-native chat-template settings were used in training and evaluation. Training appended the target response; evaluation supplied the task input without that target. This is instruction/format parity, not identity of training and held-out input content.

3.4 Execution measurements

Training completed 129 optimizer updates, with 129 recorded gradient checks and 516 finite per-microbatch losses.

Observation	Recorded value
Training time	1,433.7557 seconds
Peak allocated GPU memory	8.8666 GiB, rounded to 8.87
First logged step loss	1.1715
Mean step loss, epoch 1	0.6974
Mean step loss, epoch 2	0.4735
Mean step loss, epoch 3	0.3590
Last ten logged steps, mean loss	0.3339
Average training loss across all 129 logged steps	0.5100
Maximum recorded preclip gradient norm	4.0433
Steps whose preclip norm exceeded 1.0	118/129

All 32 adapter tensors changed, covering all 1,835,008 trainable elements. Independent offline comparison of the preserved adapter arrays confirmed those differences; the maximum absolute element change was approximately 0.01022. The final saved adapter matched the final checkpoint. Recorded before/after foundation checks remained unchanged, and an adapter reload check was recorded as successful.

The checkpoint records included training states at steps 43, 86 and 129. Checkpoint resume was not independently executed or certified. Full foundation weights were not independently compared by the reviewer; the foundation-preservation conclusion relies on saved execution checks.

The recorded preclip norm exceeded the configured threshold in 118/129 steps. Clipping was therefore expected to limit the gradient norm frequently. This observation is neither a diagnosed defect nor causal evidence about the numerical errors; it does not directly bound every AdamW parameter update.

No held-out validation-loss curve or separate task evaluation of the intermediate saved checkpoints is established by the inspected archive. The study therefore does not show when the observed label or budget behavior changed during training.

Lower training loss establishes improved fit to training targets under the objective. It does not establish improved generalization or safer interpretation.

4. Paired evaluation

Terminology

Term used here	Meaning
Base model	Qwen3.5-4B without the experimental adapter
Adapted model	Same foundation with the trained adapter
Record	Accumulated interpretation used for the experiment
Reference	Provisional AI-generated expected answer
Checker	Programmed checks of model output; “checker-gated” means a match counts only if these checks accept the output
Label agreement	Exact equality with a reference label, not human-rated quality

Both comparison arms used matching evaluation prompts and recorded generation settings. The base model was compared with the same foundation plus the fresh adapter. Both arms used the same model-native chat template with thinking mode disabled. Decoding was greedy with sampling disabled, up to 512 new tokens, and generation caching disabled. Template identity was recorded consistently across tokenization and both evaluation arms; neither the template text nor confidential instruction content is disclosed. Each arm completed the same 36 cases, with zero recorded token-limit hits.

Earlier research used different instructions and other configurations. Cross-stage changes must not be presented as controlled evidence of a training-only effect. The internal prompts and their exact structures are withheld.

4.1 Metric definitions

Public metric name	Definition and limitation
Outer structure	Checks that the response parses into the required outer JSON form. This is not full semantic correctness; the underlying schema is not reproduced.
Checker-gated resulting-record agreement	An accepted interpretation must yield a record matching the reference. Rejected outputs receive no match. This incorporates the checker and provisional labels.
Exact next-step-label agreement	The reported conversational-step label equals the reference label. It is counted separately from checker acceptance and does not rate response wording.
Budget-value agreement	On cases with an expected budget change, the predicted amount and price basis match the reference. This does not certify the whole response.
Checker-gated selection agreement	On cases supplying fictional evidence, the selection matches the reference and the checker accepts the output. Expected empty selections are included.
Checker acceptance	The output passes the implemented checks with no recorded token-limit hit. This is not independent factual certification.

These definitions describe evaluation measurements without reproducing the checker implementation, eligibility rules or internal representation. Gating has a material consequence: rejection receives no record or selection match even if an unchanged returned record happens to equal the reference. Those metrics partly measure the checker.

Exact next-step-label agreement can reflect AI-authored label conventions. It is not independently validated conversational quality; response helpfulness and wording were not independently human-scored.

5. Quantitative results

5.1 Combined 36-case comparison

Measure	Base model	Adapted
Outer structure	36/36	36/36
Checker-gated resulting-record agreement	29/36	30/36
Exact next-step-label agreement	11/36	27/36
Budget-value agreement	8/8	6/8
Checker-gated selection agreement	10/13	9/13
Checker acceptance	32/36	31/36
Recorded token-limit hits	0/36	0/36

The outer-structure result combines three reported format observations, each 36/36 in both arms: JSON-object parsing, required outer-key presence and the expected collection type. These observations do not disclose the key names or serialized contract.

5.2 Held-out versus regression results

Measure	Base held-out	Adapted held-out	Base regression	Adapted regression
Resulting-record agreement	19/24	20/24	10/12	10/12
Checker acceptance	22/24	21/24	10/12	10/12
Budget-value agreement	6/6	5/6	2/2	1/2
Selection agreement	5/7	4/7	5/6	5/6

Paired held-out record outcomes were 17 matches in both arms, three adapter-only matches, two base-only matches and two cases matching in neither arm. The regression set had ten shared matches and two shared nonmatches.

The historical held-out difference is one case, approximately 4.17 percentage points; the combined difference is approximately 2.78 points. Exact next-step-label agreement changes by 16/36 cases, approximately 44.44 points. These are different fixed-reference measurements.

Disputed-label sensitivity: Excluding the disputed case, held-out record agreement is **19/23 versus 19/23**. An alternative reference that included the explicitly expressed requirement could favor the base model and reverse the direction. That alternative has not been adjudicated or substituted into the historical results.

Exploratory paired comparisons

“Adapted only” and “base only” count cases satisfying the measure in just that arm. Applicable denominators differ.

Measure and sample	Both satisfy	Adapted only	Base only	Neither	Exact two-sided sign-test p
Record agreement, 24 held-out	17	3	2	2	1.0000

Measure and sample	Both satisfy	Adapted only	Base only	Neither	Exact two-sided sign-test p
Next-step-label agreement, 24 held-out	6	12	2	4	0.0129
Budget agreement, 8 applicable pooled	6	0	2	0	0.5000
Selection agreement, 13 evidence-bearing pooled	8	1	2	2	1.0000
Checker acceptance, 36 pooled	30	1	2	3	1.0000

Held-out next-step-label totals are 8/24 and 18/24; the regression totals are 3/12 and 9/12, giving the unchanged pooled 11/36 and 27/36. Pooled label discordance is 19 adapted-only versus three base-only.

These are exploratory calculations without multiplicity correction. The exact sign test conditions on discordant pairs; a purposively authored, AI-generated fixture is not a random population sample. The label result does not establish independent conversational quality, and one- or two-case differences in either direction cannot establish stable population effects. The two observed tenfold errors remain consequential observations without estimating an error rate.

5.3 Additional behavioral observations

Mean complete output length was approximately 516 characters for the base model and 292 for the adapted model. Shorter output was not independently rated as more helpful or correct. A possible confound is that fewer factual assertions may create fewer opportunities for rejection; this is a hypothesis, not a measured explanation or evidence that brevity caused acceptance differences.

Reference agreement can reward an annotation omission. The omitted-requirement case in section 2.1 is the annotation-sensitive adapter-only match; it is not an established interpretation gain. It requires human adjudication. This edition retains the recorded scores and qualifies their interpretation; it does not silently alter a dataset or a historical result.

6. Observed failures and localized differences

In Indian numbering, 1 lakh is 100,000 and 1 crore is 10,000,000, or 100 lakh. These definitions explain the tenfold differences below.

These descriptions summarize recorded observations without reproducing private case identifiers, raw prompts, output structures or transcripts.

6.1 Budget example A: held-out evaluation

An intended budget of ₹12 lakh, or ₹1,200,000, was represented by the adapter as ₹1.2 crore, or ₹12,000,000. The base model represented the amount correctly. The adapted prose still referred to the intended amount, so apparently reasonable language concealed a tenfold error in the structured value.

The checker rejected the adapted output. This establishes detection of this error in this experiment, not universal protection against numerical mistakes.

6.2 Budget example B: regression evaluation

An intended ₹19 lakh, or ₹1,900,000, was represented by the adapter as ₹1.9 crore, or ₹19,000,000. The base model's amount was correct, but its full response failed a separate check. Both responses therefore illustrate why correctness of one measurement must be separated from whole-output success.

The adapted numerical error was rejected. There were no recorded tenfold errors in the base model's eight applicable budget-value updates.

6.3 Evidence and interpretation failures

The adapter selected a fictional option without sufficient supporting evidence, and that output was rejected. The base model avoided that selection but still gave an unsatisfactory conversational response. Avoiding one failure is not proof of complete correctness.

The three historical adapter-only record matches included a requirement-strength interpretation, multi-turn information preservation and the annotation-sensitive omitted-requirement case in section 2.1. The last is not an established semantic gain. Both models also failed a feature-preference interpretation. Two base-only record matches and the other negative differences deserve the same small-sample caution as the positive differences.

No general vehicle-eligibility algorithm, product rule or exact selection criteria are inferred from these summaries.

7. Teacher-model proof of concept

7.1 Configuration and experimental design

Qwen3-14B belongs to the Qwen3 family 7. The probe used its Unsloth four-bit repackaging 4; the underlying model information and licence are identified in reference 3.

Item	Recorded setting
Model	unsloth/Qwen3-14B-bnb-4bit
Public model revision	b455922269e4678c84f854fe38a6993d73614a1e
Hardware	Tesla T4
Reported hardware memory	15,360 MiB
Compute type	FP16
Thinking mode	Enabled
Sampling temperature	0.6
Top-p	0.95
Top-k	20
Minimum-p	0
Seed	20261111
Maximum new tokens	8,192
Configured context limit	32,768 tokens

The recorded environment included Python 3.13, PyTorch 2.6.0 with CUDA 12.6, Transformers 4.51.3, Tokenizers 0.21.1, Safetensors 0.5.3, Accelerate 1.6.0, bitsandbytes 0.46.1 and huggingface_hub 0.30.2. Runtime warnings were recorded; loading and generation nevertheless completed in both preserved runs. This list describes an observed experiment and is not an installation or deployment prescription.

The smaller-model training was recorded on 7 November 2026; the teacher probes were recorded on 11 November 2026. Each run tested four selected turns from one AI-generated episode. The pre-run preparation record fixed the selected turns before run 1, choosing an existing episode for coverage while avoiding unresolved research cases. This was purposive selection, not random sampling. Evaluation was teacher-forced: every selected turn used the reference conversation history, with earlier generated answers excluded from subsequent inputs. Consequently, these results do not measure autonomous conversational rollout or accumulated state errors.

Run 2 followed revisions informed by run-1 failures on the same four turns, including instructions and evaluation diagnostics. The recorded model and generation settings matched across runs. These are development-sample iterations, not an independent replication; no fresh unseen teacher test was established. No prompt text or confidential training trajectory is reproduced.

7.2 Replayed automated outcomes

Each cell lists pass / fail / unverified counts across four selected turns.

Automated check	Run 1	Run 2
Completeness	4 / 0 / 0	4 / 0 / 0
Required structure	1 / 3 / 0	4 / 0 / 0
Requirement extraction	0 / 0 / 4	4 / 0 / 0
Budget/record comparison	1 / 0 / 3	4 / 0 / 0
Clarification behavior	0 / 1 / 3	2 / 2 / 0
Evidence-selection consistency check	1 / 0 / 3	4 / 0 / 0
Automated grounding check	1 / 0 / 3	4 / 0 / 0
Overall experimental outcome	Fail	Fail

Here “evidence-selection consistency” describes the recorded check of selected fictional options against the expected selection and supplied evidence. It is not a classifier deciding whether the customer message concerns cars; the implementation is not disclosed.

The aggregate verdict is fail if any selected turn has a failed automated check, including clarification. If none failed, the outcome would still be pending human review rather than automatic approval of supervision. Thus run 2's two clarification failures are compatible with four passes on structure.

A structurally invalid response can make a further check unverified. Zero extraction passes in run 1 must not be described as four demonstrated extraction failures.

The observed structural improvement followed a prompt revision. These two small samples do not isolate the causal contribution of each instruction change or establish a general teacher-model success or failure rate.

7.3 Resource observations

Measure	Run 1	Run 2
Model load time	42.5 s	45.6 s
Four selected-turn inference total	378.7 s	462.3 s
Per-turn times, in evaluated order	66.3, 107.5, 80.2, 124.7 s	91.3, 159.4, 89.5, 122.1 s
Generated tokens, including thinking	2,082	2,240
Prompt-token range	816–1,468	1,050–1,702
Recorded token speed range	5.29–5.75 tokens/s	4.47–5.26 tokens/s
Peak inference allocation	9.82 GiB	9.88 GiB

These figures describe the observed quantized experiment, not a service-level benchmark. Both runs completed their selected generations without a recorded token-limit hit.

7.4 Failure interpretation and supervision status

The first run included a mislabeled output element, and the model dropped a requirement the customer had not withdrawn. The second run improved structure, but two clarification checks still failed. A recorded response also treated an unlisted feature as absent while passing the automated grounding check. Thus a recorded grounding pass must not be read as complete semantic verification.

Both runs loaded the model successfully and produced outputs. Qualification failed on output checks, not on model loading. The outputs remained unapproved for student training pending human review, rights review and separate authorization.

Knowledge distillation concerns transferring learned behavior from a teacher to a student 5. The smaller-model training preceded these teacher probes and was not trained from them. Neither teacher run demonstrates completed knowledge distillation.

8. Evidence integrity and reproducibility boundaries

AI-assisted offline verification checked preserved experiment outputs and reproduced the reported evaluation summaries. It was automated technical verification, not independent human peer review. Adapter-array comparison confirmed parameter changes. No new GPU training or model inference was performed during that review.

Windows and Unix source-file line endings differed, changing a file checksum without changing the represented source content. This resolves that byte-format difference, not every limitation in binding source versions to the original execution. Some exact source-execution provenance was not recorded.

The reference material remains AI-generated without established independent human adjudication. The checker is experimental and does not certify real-world factual or decision quality. Independent end-to-end training reproduction, checkpoint resume and production integration were not established.

This edition deliberately withholds proprietary data representations, raw datasets, prompts, source code and internal evidence identifiers. It therefore does not provide enough material for exact independent end-to-end reproduction. No public supplementary dataset or implementation is supplied or licensed by this manuscript.

9. Research Limitations and Human Review Status

EXA's research direction, founder review and editorial approval responsibilities are separate from independent human adjudication of references. Neither complete human reference adjudication nor human peer review is established. AI-based verification does not fill that gap.

1. **Sample size:** One training run, 24 held-out cases and 12 known regression cases provide limited evidence.

2. **Provisional references:** There was no claimed independent human adjudication, and a potential omission affects a scored improvement.

3. **AI-generated coverage:** Constructed examples may differ materially from actual customer language.

4. **Narrow languages:** The English-dominated sample cannot establish regional-language or voice-transcription quality.

5. **Development feedback:** Known regression cases influenced the research process.

6. **Shared assumptions:** Automated overlap checks cannot eliminate semantic leakage or correlated annotation assumptions.

7. **Metric dependence:** Resulting-record and selection scores incorporate a checker; exact conversational labels do not score all response quality.

8. **Cross-stage confounds:** Prompt, data and training changes across earlier experiments prevent single-factor attribution.

9. **Teacher forcing:** The teacher probes do not propagate prior generated mistakes and cover only selected turns of one episode.

10. **Incomplete factual checks:** Automated grounding passes did not establish all statements as true.

11. **Provenance and access:** Publicly withheld assets and incomplete source-execution binding limit exact reproduction.

12. **Unmeasured outcomes:** Commercial cost, throughput, long-term reliability, human satisfaction, fairness and deployment performance were not established.

These limitations restrict the conclusions to the inspected experiment. They do not justify broad claims about all Qwen models, all LoRA adaptations or all teacher–student methods.

10. Scientific conclusions

The recorded run establishes completed training within the measured resource envelope and changed trainable parameters. On the fixed references, label agreement increased and record agreement was one case higher; the latter becomes a 19/23 tie when the disputed case is excluded. Budget, selection and acceptance counts were lower by two, one and one applicable cases. Positive and negative differences of that size require equal caution. The tenfold budget errors are real observations, not a reliable population error-rate estimate.

The teacher probes establish completed quantized model loading and generation followed by failed qualification after development on the same sample. Better structure in run 2 concerns the same 14B model; no controlled 4B-versus-14B comparison on common cases was performed. They do not establish approved supervision or successful distillation.

The central result is mixed: **successful optimization did not demonstrate reliable task performance.** Valid JSON, reduced loss and improved action-label agreement cannot substitute for semantic review and separate examination of consequential failures.

No production, commercial-replacement or cost-savings conclusion follows from these data.

11. About EXA and Its Research Contribution

EXA Ecommerce Private Limited initiated, funded and directed this automotive AI research programme. EXA contributed founder time, financial resources and research infrastructure, and established quality, confidentiality and research-governance requirements.

Angappamoorthy A. — Founder & Managing Director established the research objectives, directed experiments, reviewed findings and made research decisions. These contribution statements reflect the founder's confirmation. EXA retains responsibility for the published research and its scientific claims.

12. AI Assistance and Research Transparency

AI coding agents executed technical workflows under founder instructions. AI systems generated all 172 training examples and all 36 evaluation cases, including customer messages and reference answers, and assisted with research code, analysis, technical verification, manuscript drafting and separate AI-based reviews. No manual EXA authorship of these examples is claimed.

EXA's contributions in direction, funding, resources, governance and research decisions are distinct from AI-generated artifacts. AI tools are not human researchers, and AI-based reviews are not independent human peer review. Founder direction and review do not imply independent human adjudication of every evaluation case or independent human validation of every result.

13. Copyright and Rights Notice

© 2026 EXA Ecommerce Private Limited, to the extent of its applicable rights. All rights reserved.

Underlying AI-generated training-data rights, third-party licences and copyright ownership have not been independently certified by this research article. Public model information identifies Qwen3.5-4B, Qwen3-14B and the four-bit repackaging as Apache-2.0; model notices do not settle the distinct training-data or other asset rights. No licence to datasets, prompts, source code, weights or supplementary data is conferred by this article.

No supplementary dataset is provided or licensed.

References

1. Hu et al. (2021), LoRA: Low-Rank Adaptation of Large Language Models.

2. Qwen, Qwen3.5-4B at the recorded public model revision.

3. Qwen, Qwen3-14B licence information.

4. Unsloth, Qwen3-14B four-bit model information.

5. Hinton, Vinyals and Dean (2015), Distilling the Knowledge in a Neural Network.

6. Loshchilov and Hutter (2019; preprint 2017), Decoupled Weight Decay Regularization, ICLR 2019 (AdamW).

7. Yang et al. (2025), Qwen3 Technical Report.

The references identify public methods and model information. They do not distribute third-party material or confer rights to private research assets. This article does not license private research assets or supplementary materials.