Can a Small AI Model Learn to Understand Car Buyers?

EXA Automotive AI Research — Public Educational Edition v2.3

Date: 16 November 2026 **Status:** Publication edition. Not human peer reviewed or production approved.
Audience: Students, developers and interested general readers.

This edition presents scientific methods and results without publishing proprietary research records or product implementation. Observed findings below describe one small experiment. Examples marked **illustration** are newly written fiction and are not quotations from a dataset or model transcript.

Reader's guide

Start here	What you will learn
Sections 1–2	The research question and basic AI ideas
Sections 3–5	AI-generated examples, LoRA training and the comparison
Sections 6–8	How to read six different scores and inspect numerical errors
Section 9	What the teacher-model probes did and did not test
Sections 10–13	Limitations, conclusions, glossary and FAQ
Sections 14–16	EXA's contribution, AI assistance and provisional rights notice

1. The question

Can additional training help a relatively small language model understand a car buyer's changing requirements? Understanding a message involves more than finding familiar words. A customer might correct an amount, express a preference or leave an important detail ambiguous.

Illustration — invented for explanation: A fictional shopper says, “Blue would be nice.” Later, the shopper says, “Actually, I prefer red.” Understanding the change requires recognizing that the later preference replaces the earlier one. This example is not a research record, software contract or description of a product's internal behavior.

A smaller model was investigated because a focused task might be learnable without retraining a very large system. That is a scientific possibility to test, not a demonstrated saving in cost, hardware or effort. Adaptation was tested to see whether examples could improve the model's task behavior beyond the shared instructions alone.

The **base model** means Qwen3.5-4B 2 without the experimental adapter; the **adapted model** means that same foundation with the trained adapter. The study compared these two versions. It also examined Qwen3-14B 3 as a possible source of teaching examples. The Qwen3 technical report provides background on that model family 7. These were research experiments; they do not demonstrate a deployed product or a new foundation model.

2. AI ideas, explained step by step

Language, tokens and parameters

A language model generates sequences of **tokens**: pieces of text that may be words, parts of words, punctuation or number fragments. Its **parameters** are learned numbers that influence the calculation. “4B” identifies a model in approximately the four-billion-parameter class. Size alone does not establish quality on a particular task.

A **prompt** supplies instructions and information. The **context window** limits the tokenized material available for one request. A result on short examples does not establish quality on long conversations.

Illustration — invented for teaching; not a study record: “Make it twelve” could refer to twelve days, twelve items or an amount of money. Recognizing the ambiguity is different from knowing how to write a fluent reply.

Structured information and a record

JSON is a common text format for structured data. An answer can be correctly formatted while containing a wrong number. No internal JSON structure is reproduced in this article.

A **record** means the running account of what a customer wants, as interpreted for the experiment. It is a useful way to measure whether meaning is preserved. This definition does not describe a product database or a software interface.

A **checker** is a programmed test of model output. Acceptance means the output passed its implemented checks. It is not a certificate that every statement is true or that a person judged it.

Fine-tuning and LoRA

Fine-tuning is additional training for a narrower task. **LoRA**, or Low-Rank Adaptation 1, learns small added grids of numbers while keeping the base model's main **weights** fixed. Weights are another name for learned parameters; a **matrix** is a grid of numbers. The learned addition is the **adapter**.

Illustration — invented for teaching; not a study record: Imagine a transparent overlay placed over a printed page. Notes on the overlay can change how the page is read, while the printed page stays unchanged. Removing the overlay removes those notes. The analogy explains separation of the original and added parameters; neural-network calculations are not literally printed pages.

Conceptual view	What changes?
Base model alone	Original learned weights are used
Base model plus LoRA adapter	Original weights are used together with learned adjustments
Training the adapter	The added matrices change; the base weights stay fixed

Technically, a large weight adjustment is represented through the product of two smaller matrices. **Rank** controls the size of the two smaller grids used to represent the adjustment; it is not a count of automotive rules. **LoRA alpha** controls scaling. In this experiment alpha was 32 and rank was 16, giving the configured alpha/rank multiplier of 2. The resulting behavior still depends on the learned matrix values.

Dropout randomly suppresses some contributions during training to help reduce memorizing the examples; it was zero here. **Attention** is how a model weighs which earlier text matters when processing a token. This experiment adapted selected attention calculations. The exact customized layer selection is not disclosed.

The adapter contained 1,835,008 trainable elements in 32 tensors, where a **tensor** is a multidimensional array of numbers. Training relatively few added parameters can reduce the amount of trainable state compared with updating every base weight. The recorded run fitted within a Tesla T4's measured resource envelope and completed training in under half an hour. This is a result for this run, not a guarantee for all LoRA jobs or a claim of free training.

Training vocabulary

Term	Plain-language meaning
GPU	A processor suited to many numerical operations in parallel
Batch size	Number of examples processed together; one here
Gradient	Information about how changing trainable numbers affects loss
Gradient accumulation	Combining several small batches before an update
Optimizer step	One update to the trainable numbers
AdamW	A standard optimization method for choosing updates from gradients [6]
Learning rate	A factor controlling update size
Warmup	Starting with smaller learning rates before reaching the configured value
Constant schedule	Keeping the configured learning rate fixed after warmup
Seed	A starting setting for computer-generated choices that imitate randomness; it aids repeatability but cannot guarantee identical execution on every system

Training loss measures mismatch with training targets under the learning objective. It is not a percentage of correct customer answers. A smaller loss usually means a better fit to those training targets; evaluation on other cases is a separate question.

Quantization represents some model values using fewer bits to reduce storage or memory needs. It differs from LoRA. The smaller-model run did not use four-bit or eight-bit loading; the teacher probe used a four-bit model. FP16 and FP32 refer to 16-bit and 32-bit floating-point representations. These numerical settings describe experiments, not product infrastructure.

3. What was trained and tested?

The smaller-model experiment used **172 AI-generated training examples, including their customer messages and target reference answers**. Their recorded provenance attributes the examples to AI tools; no human authorship of these examples is claimed. “Synthetic” means constructed rather than sampled as a representative survey of customer traffic.

A training example pairs an input with a target answer. **Illustration — invented for teaching; not a study record:** Input: “I first preferred blue, but now I prefer red.” Target interpretation in ordinary prose: “The current color preference is red.” This is a fresh teaching example, not a dataset row or an output contract.

Evaluation used 24 held-out cases and 12 previously known regression cases. A **held-out case** is excluded from training. A **regression case** checks behavior already examined during development. The 12 regression cases must not be described as newly unseen tests.

Dataset	Number of examples
Training	172
Held-out evaluation	24
Known regression evaluation	12

Training examples were labeled as 164 English, six Hinglish and two Tamil-English. The held-out set contained 23 English and one Tamil-English example. This coverage cannot establish dependable multilingual performance.

The provenance records also identify **all 36 evaluation cases, including their customer messages and reference answers, as AI-generated**. Independent human adjudication of those references was not established. This does not mean that no person directed the research or reviewed its conclusions; those activities are different from independently checking every label. Automated overlap screening found no flagged cross-split dialogue matches under its implemented checks. That does not prove semantic independence or label correctness.

A generic research workflow

Illustration — invented for teaching; not a study record or a product design:

Write teaching examples → Train a small adapter → Give both models the same test questions → Score separate properties → Inspect mistakes and limitations

This illustrates a research method, not EXA's application architecture. Reviewing an answer key before using it matters: if a target answer omits a meaningful fact, a model can be rewarded for the same omission.

4. What happened during training?

The experiment adapted Qwen3.5-4B using LoRA rank 16, alpha 32 and dropout zero. The adapter contained 1,835,008 trainable parameter elements.

Training setting or observation	Recorded value
Epochs, meaning passes through training data	3
Optimizer updates	129
Batch size	1
Gradient accumulation	4
Optimizer	AdamW
Learning rate	0.0001
Warmup	8 steps, followed by a constant learning rate
Random seed	42
Preprocessing sequence cap	1,024 tokens
Observed training sequence lengths	612–874 tokens
Training time	1,433.76 seconds, approximately 23 minutes 54 seconds
Peak allocated GPU memory	8.87 GiB

Connecting examples, passes and updates

The batch size was one, so each small batch processed one example. Four such batches accumulated before an optimizer update:

• 1 example per batch × 4 accumulated batches = **4 example presentations per update**.

• 172 examples ÷ 4 = **43 updates per epoch**.

• 43 updates × 3 epochs = **129 optimizer updates**.

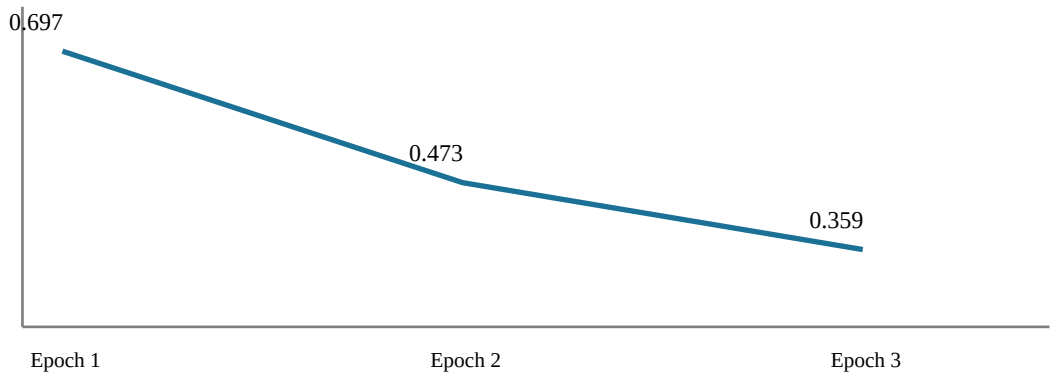
• 129 updates × 4 = **516 example presentations**, representing three passes over the same 172 examples, not 516 unique examples.

Loss was computed on the assistant's response continuation, including its termination tokens; prompt tokens were excluded from the objective. The same instruction and chat-template settings were used in training and evaluation, while the actual example inputs differed.

The saved trainer metadata also records a length setting of 1,536 tokens. That differs from the preprocessing cap; every observed training example was shorter than both limits. Neither number establishes performance on long real conversations.

The recorded GPU was a Tesla T4. Model weights were loaded in FP16, with a required FP32 activation path and automatic mixed precision disabled. These are experimental training conditions, not a description of deployment infrastructure.

Seeing training loss fall



Reconstructed visualization of reported rounded epoch mean losses (not the original source figure).

Observed study measurement: The figure plots the three rounded epoch means. The vertical axis is training loss, not evaluation accuracy.

Average logged losses across the three epochs were **0.697 → 0.473 → 0.359**. The **average training loss across all 129 logged steps was 0.5100**, rounded to four decimal places. All 32 adapter tensors changed; recorded foundation checks remained unchanged. A saved adapter reload was recorded as successful.

Illustration — invented for teaching; not a study record: A student practices with an answer key and makes fewer mistakes on those practice questions. That shows learning of the practice material. It does not tell us the student's score on a different exam, especially if the answer key itself has mistakes.

A loss of 0.359 is a value of the training objective, not “35.9% wrong” or “64.1% accurate.” The falling curve supports improved training fit. The comparison on held-out cases is needed to assess behavior beyond those examples.

5. How was the comparison made?

The base model and the adapted model received matching evaluation prompts and generation settings. This helps separate an adaptation effect from the effect of giving one model different instructions. The prompts themselves are not reproduced.

Both arms used the same model-native chat template with thinking mode disabled in training and evaluation. The public account does not disclose that template or the instruction text. Generation was greedy: the next token was chosen without sampling, with a maximum of 512 new tokens. None of the 36 recorded outputs in either comparison reached that limit.

The study measured several different aspects of behavior. Some measures include the effect of a checker; others inspect a particular part of a response directly. The measures have different denominators and should not be averaged into one “AI accuracy” score.

6. The six main results

How to read a score

In **6/8**, six is the numerator—the number of applicable cases passing the measure—and eight is the denominator—the number tested for that measure. **30/36** answers a different question if it uses a different definition. Eight budget cases are not 36 independent budget tests.

An applicable denominator excludes cases that do not test the property. The same case can pass formatting and fail an amount check, so these ratios cannot be averaged into one accuracy score.

Measure	Base model	LoRA-adapted model
Required outer JSON structure	36/36	36/36
Resulting-record agreement, counts only if the checker accepts the output	29/36	30/36
Exact agreement with the expected next-step label	11/36	27/36
Budget-value agreement, applicable cases	8/8	6/8
Selection agreement, counts only if the checker accepts the output; evidence-bearing cases	10/13	9/13
Checker acceptance	32/36	31/36

Structure: 36/36 for both

The outputs passed the reported outer format checks. This does not mean every value or statement was correct.

Illustration: A form may have every required box filled in while stating the wrong amount. Correct format makes information easier to process; it does not make the information true.

Resulting-record agreement: 29/36 versus 30/36

This measure compares the accumulated interpretation with a reference answer after acceptance by the checker. A rejected output does not receive a match.

Illustration: A fictional record should still contain an earlier preference after an unrelated fact changes. An accurate new fact would not compensate for losing information that should have remained.

The adapter matched one additional record overall. This is agreement with unreviewed reference labels, not a pure measurement of human understanding.

Expected next-step label: 11/36 versus 27/36

This is exact agreement with a reference label for the next conversational step. It improved by 16 cases.

Illustration: Asking for clarification and acknowledging an unambiguous correction are different kinds of response. Matching the expected label does not prove that the wording, factual content or wider conversation is good.

Natural-language helpfulness was not separately rated by independent human reviewers.

Budget-value agreement: 8/8 versus 6/8

Eight cases contained an applicable reference budget change. This measurement compared the budget amount and price basis. It did not certify the whole response.

The two adapter errors were both ten times too large. They are described below. Eight examples are too few to estimate a general budget-error rate.

Selection agreement: 10/13 versus 9/13

Thirteen cases supplied fictional evidence. This measure required agreement with the reference selection and acceptance by the checker. The applicable set included cases where the expected selection was empty; it was not a set of 13 successful recommendations.

Illustration — invented for teaching; not a study record: A fictional worksheet's answer key names option B. A model writes option C, so it does not match the key. Even choosing B would not establish that the answer key was factually right. Selection agreement and verified real-world correctness are different questions.

No real vehicle specification or commercial recommendation is asserted here.

Checker acceptance: 32/36 versus 31/36

An accepted output passed the implemented checks. Acceptance was not independent human certification.

Illustration — invented for teaching; not a study record: A simple form check verifies that a date is written in a recognized format. “1 January 2030” can pass that format check even when the event actually occurred on another date. Passing the implemented check does not settle the unchecked fact.

Acceptance and correctness therefore need separate measurement.

A worked example: one score up, another down

Illustration — invented for teaching; not a study record. The following five travel-planning mini-cases and the two imaginary models are fictional. The ordinary step descriptions are invented teaching labels and do not reproduce the study's labels. Amounts are rupees.

Case	Expected step	Correct amount	Model A step	Model A amount	Model B step	Model B amount
A	Ask a question	10,000	Ask a question	10,000	Ask a question	10,000
B	Confirm what was heard	12,000	Confirm what was heard	12,000	Offer options	12,000
C	Offer options	15,000	Offer options	150,000	Offer options	15,000
D	Ask a question	18,000	Ask a question	18,000	Confirm what was heard	18,000
E	Confirm what was heard	20,000	Confirm what was heard	200,000	Ask a question	20,000

Model A matches all five expected steps, but its amounts are right in only three cases. Model B matches two steps and all five amounts:

Fictional measure	Model A	Model B
Expected step matched	5/5	2/5
Amount correct	3/5	5/5

Model A wins one measure and loses another. Nothing in a correct step label repairs the two wrong amounts. These invented scores are separate from the actual research results.

7. Keep the evaluation splits separate

Measure	Base: 24 held-out	Adapter: 24 held-out	Base: 12 regression	Adapter: 12 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

Among the 24 held-out cases, both models matched the reference record in 17 cases, only the adapter matched in three, only the base model matched in two, and neither matched in two.

A one-case difference on 24 cases is about 4.2 percentage points. The result is sensitive to one disputed AI-generated reference: the base model recorded an explicitly expressed requirement, the adapted model omitted it, and the reference also omitted it. The adapted prose nevertheless referred to that requirement. Both models chose the same fictional option. Exact record agreement therefore rewarded the omission.

Excluding the disputed case, held-out record agreement is 19/23 for the base model and 19/23 for the adapted model—a tie. Under an alternative reference including the disputed requirement, the base model could score higher. This is a sensitivity explanation, not a corrected dataset or an adjudicated new score. All historical scores above remain unchanged.

A paired comparison asks which cases matched in just one of the two models. Cases where both matched or both missed do not appear in the two right-hand columns below.

Measure and sample	Only adapted matched	Only base matched
Next-step labels, 24 held-out cases	12	2
Records, 24 held-out cases	3	2
Budget values, 8 applicable cases	0	2
Selection, 13 evidence-bearing cases	1	2
Checker acceptance, all 36 cases	1	2

The label difference is much larger in this sample; the other differences are one or two cases. A one- or two-case difference in either direction cannot establish a stable population advantage or error rate. The large label-agreement difference may reflect learned AI annotation conventions; it is not independently validated conversational quality.

These comparisons are exploratory. Testing several measures increases the chance of finding an apparent difference by chance; no statistical adjustment for that multiple testing was applied. The examples were constructed, not randomly sampled from real customers.

8. What did the recorded failures teach us?

Lakh and crore

In Indian numbering, **1 lakh = 100,000** and **1 crore = 100 lakh = 10,000,000**. Thus ₹12 lakh is 12 × 100,000 = ₹1,200,000, while ₹1.2 crore is 1.2 × 10,000,000 = ₹12,000,000. The second amount is ten times the first. The symbol ₹ denotes Indian rupees.

The following are **scientific paraphrases of observed findings**, not verbatim prompts, dataset rows or model transcripts.

Observed finding	Intended amount	Adapted amount	Meaning
Budget example A, held-out evaluation	₹12 lakh = ₹1,200,000	₹1.2 crore = ₹12,000,000	10× too large
Budget example B, regression evaluation	₹19 lakh = ₹1,900,000	₹1.9 crore = ₹19,000,000	10× too large

The base model represented both amounts correctly. In the second example, however, its full output still failed a different check. Correct budget extraction did not make its entire answer correct.

Both wrong adapted amounts were rejected by the checker. That is an observed outcome for these cases, not proof that every possible number error would be caught. In the first example, the adapted model's prose still referred to the intended amount, hiding the inconsistent machine-readable value.

Another adapted output selected a fictional option without enough supporting evidence. The selection was rejected. The base model avoided that selection, but its conversational answer was still not fully satisfactory.

The three historical adapter-only record matches included a requirement-strength interpretation, information preservation across messages, and the disputed omitted-requirement case described in section 7. That last case is annotation-sensitive, not an established semantic gain. Both models also failed another feature-preference interpretation.

The two tenfold errors are observed facts and remain consequential examples. Eight applicable budget cases do not estimate a reliable numerical-error rate, just as a one-case record gain does not establish a reliable improvement. “Mixed results” describes these fixed observations, not a statistically established general degradation or improvement.

The practical scientific lesson is that formatting, numbers, recorded requirements, action labels and evidence use can improve or regress separately.

9. Could a larger model supply teaching examples?

A **teacher model** can generate examples that might later help train a smaller student. This is one possible approach within the wider family of knowledge-distillation methods 5. Incorrect teaching examples can transfer errors rather than useful behavior.

The teacher proof of concept used the four-bit model identified as unsloth/Qwen3-14B-bnb-4bit 4 on a Tesla T4 with 15,360 MiB reported hardware memory.

The smaller-model training was recorded on **7 November 2026**; these teacher probes were recorded on **11 November 2026**. Two runs each generated responses to four selected turns from the same AI-generated episode. The four turns were selected in a preparation record before the first run, to cover several language challenges while avoiding unresolved research cases. The selection was purposive, not random or representative. These were **teacher-forced probes**: each turn received the reference conversation history rather than earlier model-generated answers. They were not autonomous conversations, and they did not measure the accumulation of earlier mistakes.

Illustration — invented for teaching; not a study record: Imagine marking a student's answer to question 4 while supplying the teacher's correct answers to questions 1–3. An earlier student mistake cannot spoil question 4's context. In an ongoing conversation, the model might have to live with its own earlier mistakes. Teacher forcing separates those two kinds of test.

Selected checks are shown below; Engineering Research Edition section 7.2 lists all checks. The completion row also reports that generation finished for all selected turns.

Automated measure	Teacher run 1	Teacher run 2
Selected turns completed	4/4	4/4
Required structure passed	1/4	4/4

Automated measure	Teacher run 1	Teacher run 2
Requirement extraction passed	0/4; four unverified	4/4
Clarification behavior passed	0/4; one failed and three unverified	2/4; two failed
Budget/record checks passed	1/4; three unverified	4/4
Automated grounding check passed	1/4; three unverified	4/4
Overall experimental verdict	Fail	Fail

Unverified is not the same as failed. When structure prevented meaningful checking, a missing pass could not be interpreted as proof that every extracted fact was wrong.

The second run followed revisions informed by the first run's failures on these **same four turns**. It is a development-sample re-evaluation, not an independent replication or a test on fresh unseen teacher cases. Better recorded structure does not isolate the causal contribution of any one revision.

The overall recorded verdict was fail if any selected turn failed an automated check, including clarification. Even without a failed check, automated passes would leave human review pending rather than establish approved supervision. Thus four structure passes did not override two clarification failures in run 2.

The automated grounding results also had a concrete limitation: a response treated an unlisted feature as absent and still passed the check. This is another reason to distinguish a programmed pass from complete factual correctness. The raw wording and underlying confidential records are not reproduced.

Recorded resource measure	Run 1	Run 2
Model loading time	42.5 seconds	45.6 seconds
Inference time across four turns	378.7 seconds	462.3 seconds
Peak recorded inference allocation	9.82 GiB	9.88 GiB

Both runs loaded the model and generated outputs. They failed qualification checks; they were not failed model-loading attempts.

The outputs were not approved for student training. No successful teacher-to-student distillation was demonstrated, and the smaller-model adapter described earlier was not trained from these teacher outputs.

10. Research Limitations and Human Review Status

The evidence supports completion of a small LoRA training run, changed adapter parameters and mixed observations on a fixed AI-generated evaluation. Label agreement increased; budget and selection agreement were lower by two and one applicable cases respectively. Differences of one or two cases in either direction need equal caution. None establishes a general model-quality change.

The study does not establish reliable generalization across real customers, multilingual readiness, lower commercial cost, production reliability or successful distillation. A falling loss and valid JSON are insufficient evidence for those conclusions.

The main limitations are:

• Small samples and only one training seed.

• AI-generated training examples and evaluation cases, including messages and references, with shared annotation assumptions.

• No established independent human adjudication of the evaluation labels.

• Development feedback from known regression cases and reuse of the teacher development sample.

- English-dominated data and no broad customer-language sample.
- Scores that partly depend on the checker and provisional answer key.
- Four teacher-forced turns, without measurement of accumulated model errors.
- No customer-outcome, commercial-cost or production-reliability study.

EXA's organizational review and founder research decisions are distinct from independent label adjudication or human peer review. Human peer review has not been established by these manuscripts.

AI-assisted offline verification reproduced the recorded evaluation counts; it was not independent human peer review and did not repeat GPU training or model inference. Windows and Unix source-file line endings can produce different file checksums while representing the same source content; other source-to-execution links remain incomplete.

Private datasets, prompts, source code and raw records are not part of this edition. Exact experimental reproduction cannot be performed from this public narrative alone. The paper should therefore be read as a bounded account of recorded research.

11. Conclusion

This experiment recorded completed adapter training, more exact next-step-label matches and two observed tenfold budget errors. The findings demonstrate why separate measurements matter within this experiment. They do not establish a mechanism explaining the errors or a general advantage or disadvantage for adaptation.

The same Qwen3-14B teacher model produced better-structured answers after revisions informed by its earlier failures, yet still failed qualification on the reused development sample. No controlled comparison between the 4B and 14B models on the same cases was conducted.

The defensible conclusion is narrow: **training completed, results were mixed, and reliability was not demonstrated.** These findings do not condemn LoRA, small models or teacher–student learning as general methods.

12. Glossary

Term	Meaning
Base model	Qwen3.5-4B without this experiment's adapter
Adapter	Added trainable parameters used with the base model
Token	A piece of text processed by the model
Prompt	Instructions and information supplied for a task
Record	Running interpretation of what the customer wants
Reference label	Expected answer used for scoring; not automatically ground truth
Held-out case	An evaluation example excluded from training
Regression case	A known example reused to check established behavior
Loss	Training mismatch objective, not evaluation accuracy
Checker	Programmed checks whose coverage is limited
Lakh	One hundred thousand
Crore	Ten million
Teacher forcing	Testing with reference history instead of generated earlier answers

13. Frequently asked questions

Does 36/36 correct JSON mean the model is accurate? No. It describes format checks; amounts, records and factual claims can still be wrong.

Is 27/36 label agreement enough to justify use? It is one measurement against AI-generated references. It does not replace human quality assessment, numerical reliability or a defined independent evaluation.

Did the teacher train the student? No completed teacher-to-student distillation was demonstrated. The reported smaller-model training preceded the teacher probes and did not use their outputs.

Was the larger model better than the smaller one? These experiments do not answer that question. The two teacher runs compared revisions of the same 14B model on a different probe.

Can a seed guarantee an identical result? It sets the starting point for computer-generated choices that imitate randomness, but software, hardware and numerical differences can still affect repeatability.

Can I reproduce the experiment from this article? Not exactly. Private assets are withheld, and source-to-execution provenance is incomplete. The description supports understanding the method, not an exact public rerun.

Were these results human peer reviewed? No. The stated reviews and technical verification were AI-assisted. EXA's direction and founder review must not be confused with independent adjudication of every reference.

14. 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.

15. 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.

16. Copyright and Rights Notice

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

Rights associated with underlying AI-generated training data, third-party material and model-related assets are separate from rights in this article and are not independently certified here. Public model information identifies the Qwen models and the four-bit repackaging as Apache-2.0; those notices do not clear the training-data rights chain or grant rights in other assets. No dataset, source code, prompt, model-weight or supplementary-data licence is granted by this article.

References

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

2. Qwen, Qwen3.5-4B public model card.

3. Qwen, Qwen3-14B public model card and 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.

Model and method names identify the technologies studied; they do not imply endorsement. This article does not license private research assets or imply production readiness.