Capability-Stratified Degradation in Ternary Language Models

CLOE V1.1 - OneBit AI

Qwen3.5-0.8B → QAT → ternary Cloe

A capability, representation, learnability, and deployment study of a 752 M-parameter pretrained language model under ternary conversion.

Anirudh Malik | Poojith Devan | M Sparsh Mehra

Executive takeaway. The central result is not that ternarisation produces a uniformly weaker language model. Instead, degradation is **capability-stratified**: specialist factual knowledge and language-modelling quality degrade sharply, while several forms of commonsense judgement, representational capacity, and downstream task learnability remain measurable. Across ten benchmarks where Cloe demonstrably discriminates between candidate answers, it retains a mean **77.1%** of the full-precision teacher. After task-specific fine-tuning, it reaches **95.6%** of an identically fine-tuned teacher on SST-2 and **79.4%** on XSum. The engineering study further shows a **634.7 MB packed artifact** and higher prefill/decode throughput than the measured latent representation, while runtime memory remains a separate constraint.

Abstract

Extreme low-bit inference offers a route toward smaller model artifacts and more constrained deployment. Ternary language models restrict weights to $\{-1, 0, +1\}$, approaching the information-theoretic limit of $\log_2 3 \approx 1.585$ bits per weight. The practical question for an existing pretrained model, however, is not simply whether the weights can be quantised, but *which capabilities survive the transformation* and whether the resulting model remains useful as a substrate for downstream adaptation.

We study this question by converting Qwen3.5-0.8B, a 752 M-parameter pretrained model, to ternary weights using quantisation-aware training (QAT). The resulting model, Cloe, is evaluated across 29 benchmarks and through representation-level diagnostics, matched downstream fine-tuning, and deployment measurements. The evidence shows a non-uniform degradation pattern. A linear probe recovers 43.76% of MMLU answers from the full-precision teacher's final-layer representations but only 26.19% from Cloe, close to the 25% chance level, supporting the interpretation that specialist factual information is substantially lost rather than merely hidden behind an impaired output channel. At the same time, Cloe retains measurable performance on ten tasks, averaging 77.0% of teacher performance on that interpretable subset. More importantly for application-specific deployment, fine-tuning raises Cloe from 61.9% to 89.8% on SST-2, corresponding to 95.6% of the matched fine-tuned teacher, while XSum reaches 79.4% teacher retention.

We also identify an evaluation pitfall: standard answer-letter scoring initially reported 22.97% MMLU accuracy because Cloe emitted "A" on 98.6% of questions. Continuation scoring and majority-class guards are therefore used throughout the interpretable evaluation. The study consequently argues for a capability-aware view of extreme quantisation: a ternary conversion may be unsuitable as a drop-in replacement for a general-purpose pretrained model, yet remain valuable as a compact, adaptable substrate for application-specific models.

Keywords: ternary quantisation, 1.58-bit models, quantisation-aware training, small language models, capability retention, representation learning, model compression, edge AI, efficient inference

1 Introduction

The deployment economics of language models are increasingly shaped by the cost of moving, storing, and executing model weights. Reducing numerical precision is one of the most direct ways to reduce the information carried by a parameter tensor. Ternary models take this idea to an extreme by constraining weights to $\{-1, 0, +1\}$, for which the theoretical information content is $\log_2 3 \approx 1.585$ bits per weight.

The research landscape has already established that models can be trained natively under such low-bit constraints. The more operationally difficult question is what happens when an *already pretrained* model is pushed into this regime. This distinction matters for practitioners because pretrained models contain a large amount of accumulated

factual, linguistic, and task-relevant structure. If that structure cannot survive conversion, a smaller checkpoint is not necessarily a more useful model.

This work therefore takes a deliberately diagnostic perspective. Rather than reducing the outcome to a single aggregate benchmark score, we ask four connected questions:

1. **Capability:** Which behaviours survive ternarisation?
2. **Representation:** Is poor task performance caused by representational collapse, or by loss of specific stored information?
3. **Learnability:** Can the compressed model still acquire a new task through supervised adaptation?
4. **Deployment:** What does the packed representation actually buy in storage and inference measurements?

This produces a more useful engineering picture than a single "accuracy after quantisation" number.

Why this matters for an AI product. A general-purpose model and an application-specific model have different requirements. A general-purpose model must preserve broad knowledge. An application-specific model can instead be judged by whether it can efficiently represent and execute the capability required by a particular product. The distinction between *knowledge retention* and *task learnability* is therefore commercially relevant: it determines whether extreme compression should be viewed only as destructive compression or also as a possible route toward compact specialised models.

2 Research Positioning and Related Work

Prior work on ternary language models can be broadly separated into native low-bit training and conversion of pretrained models. BitNet b1.58 established the feasibility of training language models with ternary weights. More recent approaches such as CAT-Q, PTQTP, ScaleQ-1.58, and Ternary Mamba investigate how pretrained models can be converted or adapted into low-bit regimes. Falcon-Edge represents another direction: native 1.58-bit pretrained models rather than conversion of an existing full-precision checkpoint.

The distinction is important for positioning this study. The contribution here is *not* claimed to be a new state-of-the-art ternary quantiser. Instead, the work uses QAT to create a controlled ternary conversion and then asks what information, behaviour, and adaptability survive.

Table 1: Positioning against major directions represented in the project literature.

Work / direction	Model regime	Core approach	Primary emphasis
BitNet b1.58	Native ternary	Train directly at low precision	Demonstrates viability of ternary training and inference.
CAT-Q	Pretrained conversion	Calibration / post-training conversion	Demonstrates that pretrained models can be brought toward ternary operation.
PTQTP	Pretrained conversion	Dual trit-plane representation	Improves the representation of ternary weights in PTQ settings.
ScaleQ-1.58	Pretrained conversion	Calibration with additional generated signals	Explores stronger post-training ternary conversion.
Ternary Mamba	Pretrained conversion	Grouped quantisation + distillation / QAT	Closest precedent in spirit for converting an existing model and reporting retention.
Falcon-Edge	Native low-bit	Native 1.58-bit pretraining	Shows the alternative strategy of avoiding conversion altogether.
This work	**Pretrained conversion**	**QAT ternarisation + capability diagnostics**	**Characterises what is lost, what survives, and whether downstream learnability remains after conversion.**

The comparison in table 1 defines the paper's central distinction. A conversion method is normally evaluated by asking whether the resulting model retains benchmark performance. Here we additionally ask whether a failed benchmark result reflects missing information, an impaired output channel, representational collapse, or a failure that can be repaired through task-specific training. This makes the study closer to a *capability characterization* than a conventional quantisation leaderboard entry.

3 Model and Ternary Conversion

3.1 Base model

Cloe is derived from Qwen3.5-0.8B, with approximately 752 M parameters. The model contains 24 transformer layers, with 18 linear-attention layers and 6 full-attention layers. Its hidden size is 1024 and its head dimension is 256. The vocabulary contains 248,320 entries, with tied embeddings and language-model head.

Table 2: Core configuration of the Cloe conversion.

Property	**Value**
Base model	Qwen3.5-0.8B
Parameters	752 M
Transformer layers	24
Attention structure	18 linear-attention + 6 full-attention
Hidden size	1024
Head dimension	256
Vocabulary	248,320
Ternary linear layers	**186 / 186**
Distinct weight values per ternary layer	**Exactly 3**
Measured ternary entropy	**1.5821 bits/weight**
Ideal ternary entropy	1.5850 bits/weight
Cumulative QAT tokens	72.4M

The values in table 2 establish that the conversion is not merely a quantisation-inspired approximation. Every targeted linear layer is verified to contain exactly three distinct weight values, and the measured entropy of 1.5821 bits/weight is very close to the theoretical ternary limit. This distinction is important because a model can otherwise be described as "ternary" while retaining continuous-valued weights in parts of its computation.

3.2 Ternary parameterisation

Each targeted linear layer uses abs-mean scaling and a straight-through estimator. The effective quantisation can be represented as

$$s = \text{mean}(|W|), \tag{1}$$

$$Q(W) = \text{round}\left(\text{clamp}\left(\frac{W}{s}, -1, 1\right)\right) s, \tag{2}$$

$$W_{\text{eff}} = W + \lambda \left(Q(W) - W\right)_{\text{detach}}. \tag{3}$$

At inference, $\lambda = 1$. An RMSNorm is applied to activations before the low-precision matrix multiplication.

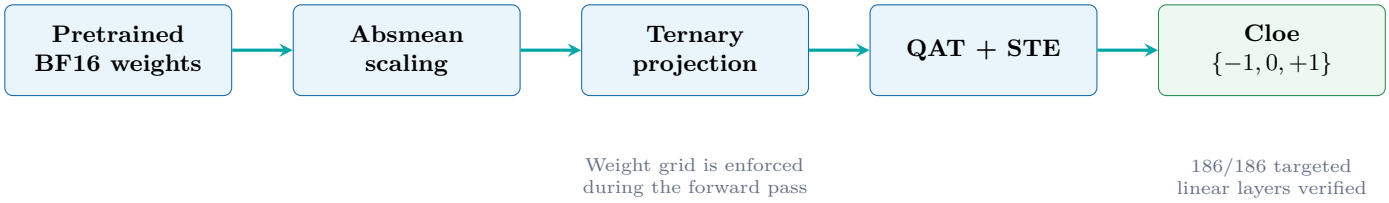


Figure 1: Overview of the conversion pipeline used to obtain Cloe.

3.3 Verification of the ternary execution path

A potential implementation failure in low-bit systems is that a checkpoint may contain a quantisation mechanism while the actual forward pass silently uses latent full-precision weights. The project therefore verifies the loaded quantisation state across all 186 targeted layers. The reconstructed weights differ from the exact ternary grid only by floating-point rounding at a relative scale of approximately 1.8×10^{-6}. Setting $\lambda = 0$ and running the latent weights produces gibberish, providing a direct sanity check that the evaluated model depends on the ternary path.

> **Interpretation.** This verification closes a common reproducibility gap in quantisation experiments: the reported behaviour is attributable to the quantised computation rather than an accidentally retained full-precision path.

3.4 Important accounting distinction

The phrase "1.58-bit model" refers to the ternary weight representation itself, not necessarily to the complete checkpoint. The 248,320-token embedding table remains BF16 and is tied to the language-model head. Consequently, the overall checkpoint is reported in the project as approximately 6.47 bits/parameter on disk, rather than a literal 1.58 bits/parameter for the entire model.

This distinction is preserved throughout the paper because it separates a *weight-representation result* from a *whole-model storage result*.

4 Evaluation Framework

The evaluation is designed around a simple principle: a low-bit model should not be declared capable merely because a raw benchmark number is numerically high. Conversely, a low number should not automatically be interpreted as proof that the underlying capability has disappeared.

Table 3: Evaluation instruments and the question each instrument answers.

Instrument	What it measures	Why it is included
Continuation scoring	Candidate-answer likelihood	Avoids relying on an answer-letter output channel.
Majority baseline	Trivial class-selection performance	Detects cases where raw accuracy is inflated by label imbalance.
Balanced accuracy	Class-balanced discrimination	Prevents constant predictions from being interpreted as capability.
Linear probe	Recoverable task information in hidden states	Separates inaccessible/removed information from output-head problems.
Effective rank	Diversity of representation directions	Tests whether quantisation causes broad representational collapse.
Matched fine-tuning	Downstream task learnability	Tests whether useful task structure can still be acquired after conversion.
Perplexity	Language-modelling quality	Measures the cost to token-level modelling rather than task classification.

The evaluation therefore moves from *behaviour* to *representation* to *learning*. This progression is important: a

benchmark tells us that a model failed a task, while a representation probe and a matched fine-tuning experiment help explain whether the failure reflects lost information or a loss of adaptability.

4.1 The answer-letter failure

The original multiple-choice evaluation produced 22.97% MMLU accuracy, which initially suggested catastrophic failure. Inspection of the generated answer distribution revealed that Cloe emitted "A" for 13,852 of 14,042 questions, or 98.6% of the evaluation set.

Table 4: Answer-letter distribution that exposed the degenerate MMLU output channel.

Model	A	B	C	D
Cloe, pre-SFT	**13,852**	113	14	63
Cloe, post-SFT	6,801	2,505	1,248	3,488
BF16 teacher	301	2,914	2,524	**8,303**

The implication is methodological rather than merely negative. A score produced by answer-letter elicitation is only meaningful if the model distributes its outputs in a way that permits genuine discrimination. The project therefore replaces this instrument with length-normalised continuation log-probability over the candidate answer text. This also makes the reported benchmark results more comparable to the ternary literature represented by BitNet, Ternary Mamba, and CAT-Q.

4.2 False-positive guards

A task is treated as a demonstrated capability only when it exceeds both a chance/majority threshold and a balanced-accuracy threshold by a two-standard-error margin:

$$\text{acc}_{norm} > \max(\text{chance}, \text{majority}) + 2\,SE, \tag{4}$$

$$\text{balanced} > \text{chance} + 2\,SE. \tag{5}$$

This guard is important because CoLA, MRPC, and BoolQ produced superficially strong raw accuracies while emitting a constant label. Balanced accuracy exposed these as degenerate outcomes.

> **How failures are treated in this paper.** Invalid or degenerate measurements are reported where necessary to establish measurement validity, but they are not treated as headline capability results. The main results use the guarded evaluation protocol.

5 Capability Retention

The central behavioural result is that ternarisation does not degrade every capability equally. On the subset of ten benchmarks for which Cloe demonstrably discriminates between candidate answers, mean performance is 77.0% of the full-precision teacher.

5.1 Teacher–Cloe comparison

Table 5: Benchmark performance and capability retention relative to the full-precision teacher. “Bar” denotes the baseline ternary model, while “Over bar” reports the absolute Cloe improvement over that baseline.

Task	Capability	Cloe	Teacher	Retention	Bar	Over bar
race	Reading	**34.0%**	38.3%	**89%**	28.0%	+6.0
piqa	Physical commonsense	**60.7%**	69.0%	**88%**	50.7%	+10.0
commonsenseqa	Commonsense	**34.0%**	39.7%	**86%**	23.7%	+10.3
qasc	Knowledge-easy	**31.3%**	37.3%	**84%**	15.0%	+16.3
sciq	Knowledge-easy	**54.7%**	66.0%	**83%**	27.0%	+27.7
hellaswag	Commonsense	**37.3%**	47.3%	**79%**	28.0%	+9.3
swag	Commonsense	**42.0%**	59.3%	**71%**	28.0%	+14.0
arc_easy	Knowledge-easy	**38.7%**	55.7%	**69%**	30.3%	+8.3
sst2	Sentiment	**59.3%**	88.0%	**67%**	51.7%	+7.7
ag_news	Topic	**36.0%**	66.0%	**55%**	29.3%	+6.7
Mean	**Ten-task benchmark mean**	**43.8%**	**56.3%**	**77.1%**	**31.2%**	**+11.6**

The most important feature of table 5 is the spread rather than any single score. RACE, PIQA, and CommonsenseQA retain 86–89% of teacher performance, while QASC and SciQ remain above 80%. These results suggest that several forms of plausibility, commonsense judgement, and relatively accessible knowledge remain usable after conversion.

At the same time, the 77.0% figure must be read as a *selected interpretable subset mean*, not as a claim that Cloe retains 77% of all general intelligence. Tasks for which both models collapse or for which the measurement instrument is degenerate do not provide a defensible retention denominator.

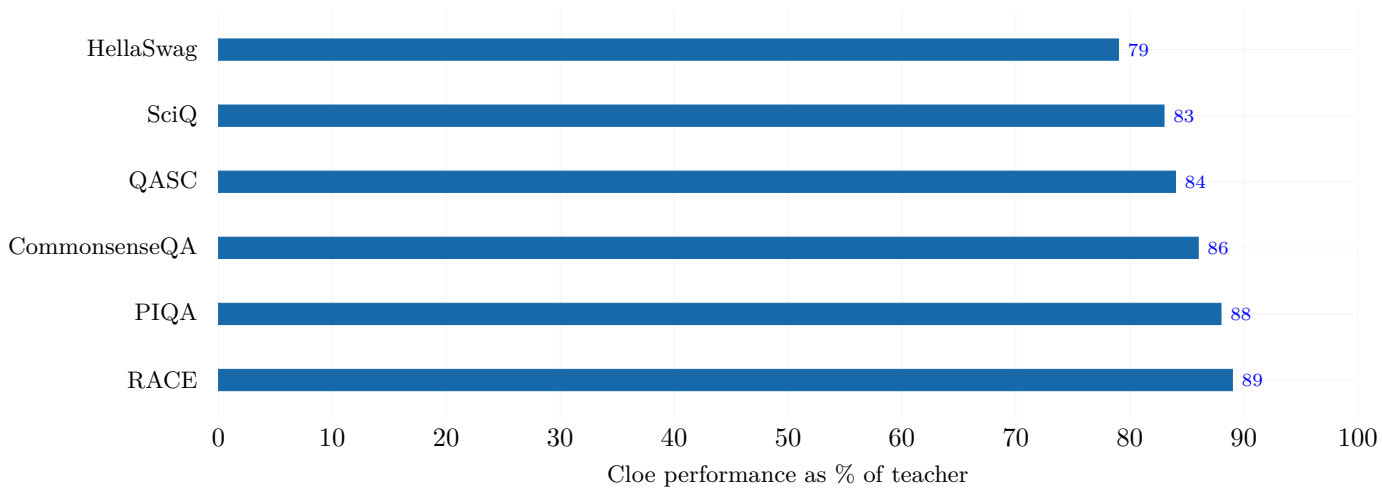


Figure 2: Representative retained capabilities. The main visual message is the non-uniform degradation profile: several commonsense and plausibility-oriented tasks remain substantially above chance after ternarisation.

6 Locating the Lost Information

The behavioural results raise a deeper question. If some benchmarks fail, is the underlying information still encoded in the hidden states but inaccessible to the standard output head, or has the conversion actually removed the information?

6.1 Linear-probe diagnostic

A multinomial logistic-regression probe is trained on final-layer hidden states from 8,000 MMLU auxiliary-training examples and evaluated on all 14,042 test questions. The probe was itself validated on synthetic representations before being applied to Cloe: it recovered 100% accuracy when the answer was explicitly linearly encoded and approximately chance when the signal was absent.

Table 6: Representation-level probe results. Chance for four-way MMLU classification is 25%.

Representation	Probe accuracy	Own answer accuracy	Effective rank
BF16 teacher	**43.76%**	44.72%	557.5
Cloe, post-SFT	26.19%	24.31%	**602.1**
Cloe, pre-SFT	26.68%	22.97%	538.9
Chance	*25.00%*	*25.00%*	–

The probe result is one of the strongest mechanistic observations in the study. Cloe's 26.19% probe accuracy is only about 1.2 percentage points above chance, whereas the teacher reaches 43.76%. Since the probe provides a relatively generous linear readout of the hidden state, the result supports the interpretation that specialist MMLU information is substantially absent from Cloe's representation, rather than simply hidden behind a poorly calibrated output head.

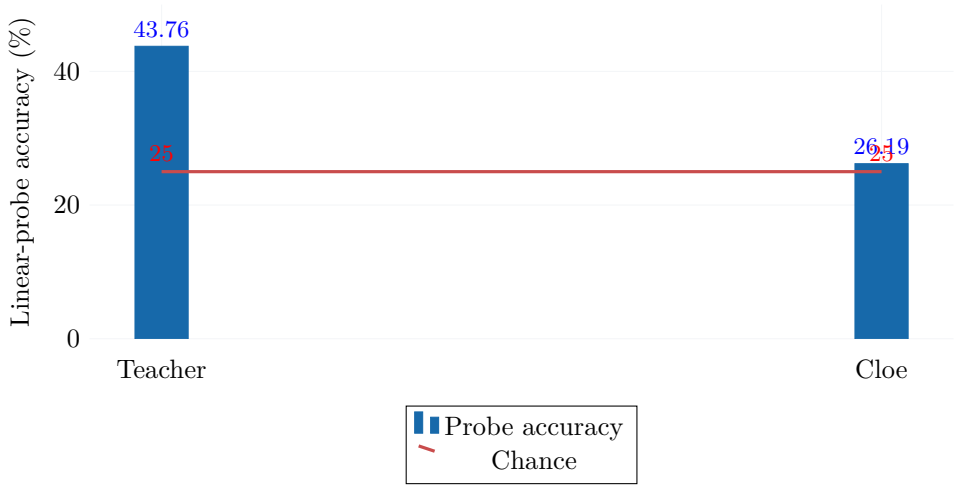


Figure 3: MMLU information recoverable from final-layer representations. Cloe approaches the 25% four-way chance level while the teacher retains substantial recoverable signal.

7 Representation Capacity Survives

The probe result establishes that a particular class of specialist information is lost. It does not, however, imply that the entire representation has collapsed. To test this distinction, we measure effective rank of the hidden-state representations.

Table 7: Effective-rank comparison.

Model	Effective rank	Relative to teacher
BF16 teacher	557.5	100.0%
Cloe, pre-SFT	538.9	96.7%
Cloe, post-SFT	**602.1**	**108.0%**

The result is notable because the post-SFT Cloe representation has an effective rank of 602.1 compared with

557.5 for the teacher. Thus the observed knowledge loss cannot be reduced to a simple "all information collapsed into a few directions" explanation. Cloe has substantial representational diversity even while specific specialist information is poorly recoverable.

> **Key distinction.** The evidence supports a separation between *representational capacity* and *stored knowledge*. Ternarisation can preserve a rich activation geometry without preserving all of the fine-grained factual information that was encoded in the original pretrained weights.

8 Downstream Learnability

The representation result becomes substantially more useful when combined with a learning experiment. If the compressed model has lost broad factual knowledge but retains useful representational structure, can it still acquire a new application through fine-tuning?

We answer this using matched teacher controls: the teacher is fine-tuned under the same downstream setup, so Cloe is compared with the teacher after both have been adapted to the task.

8.1 SST-2 sentiment classification

Cloe begins at 61.9% and reaches 89.8% after fine-tuning. Relative to the matched fine-tuned teacher, this corresponds to 95.6% retention. The absolute gap is reported alongside the retention number because a retention ratio without the teacher's absolute score can exaggerate the apparent quality of a weak baseline.

Table 8: Matched fine-tuning comparison on SST-2.

Stage	Teacher	Cloe	Cloe / Teacher
Before fine-tuning	–	61.9%	–
After fine-tuning	–	**89.8%**	**95.6%**

The important observation is not simply that the score increased. It is that a model whose specialist knowledge is substantially degraded can still reorganise its parameters sufficiently to acquire a downstream discriminative task. The measured learning gap is approximately 4.13 percentage points with an estimated standard error of ±2.39 points in the reported experiment.

8.2 XSum summarisation

The same pattern is observed on a generative task. Cloe improves from 11.1 to 19.0 ROUGE-L and reaches 79.4% of the matched teacher. It also exceeds the LEAD-1 copy/input-order baseline by approximately 2.4×, supporting the interpretation that the improvement reflects actual summarisation learning rather than simple copying.

Table 9: Matched downstream learning results.

Task	Cloe before	Cloe after	Teacher retention
SST-2	61.9%	**89.8%**	**95.6%**
XSum	11.1 ROUGE-L	**19.0 ROUGE-L**	**79.4%**

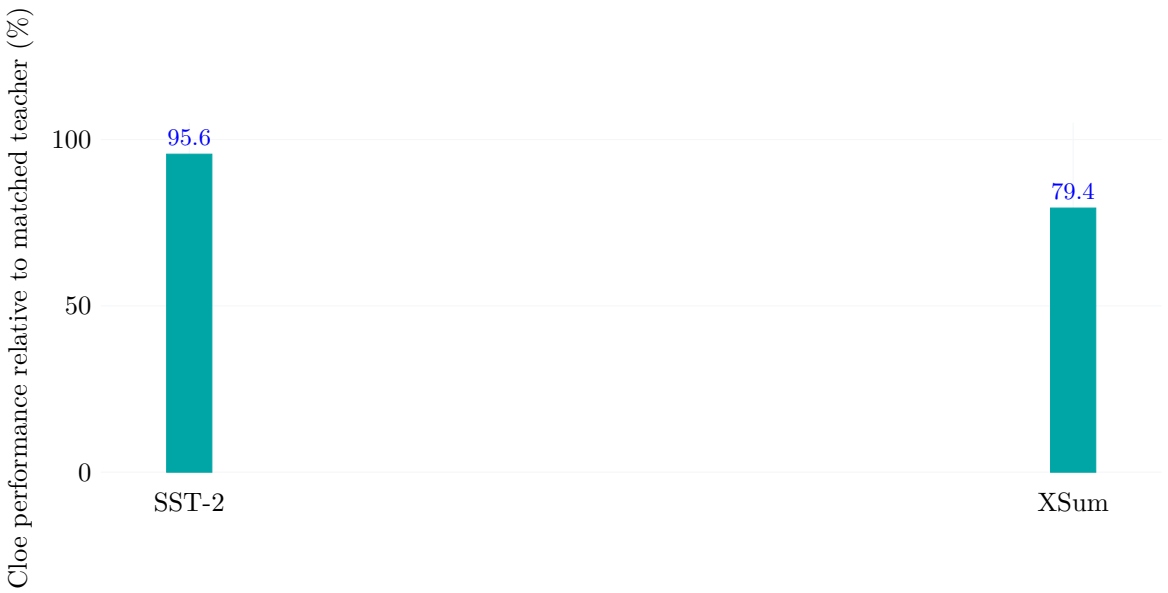


Figure 4: Downstream learnability expressed relative to identically fine-tuned teacher controls. The absolute task scores remain important and are reported in table 9.

Why this matters. The combination of the probe and fine-tuning results gives the paper its strongest conceptual separation: Cloe does not preserve all of the teacher's pretrained knowledge, but it retains enough representational structure to learn useful application-specific behaviour. This is the basis for viewing extreme quantisation as potentially useful for specialised deployment rather than only as an attempt to reproduce a general-purpose teacher.

9 Capability Profile: What Changes and What Survives?

The results can be condensed into a capability map. This table is intentionally more important than a long narrative: the reader should be able to understand the operating envelope of the model before reading the interpretation below it.

Table 10: Capability-stratified outcome profile of ternarisation.

Property	Direction	Evidence	Interpretation
Specialist factual knowledge	↓↓↓	MMLU + linear probe	Much of the teacher's recoverable specialist information is lost.
Arithmetic reasoning	↓↓↓	GSM8K	Not a reliable general arithmetic model after conversion.
Language modelling	↓↓	LAMBADA / WikiText-2	Token-level modelling quality degrades substantially.
Commonsense judgement	→	10-task benchmark subset	Several plausibility-oriented capabilities remain.
Representation diversity	→	Effective rank	No evidence of broad representational collapse.
Downstream learnability	→	SST-2 / XSum	Task-specific adaptation remains effective.
Packed storage	↑↑	Deployment benchmark	Packed artifact is substantially smaller than latent representation.
Inference throughput	↑	Packed benchmark	Prefill and decode are faster in the measured packed-vs-latent comparison.

The table suggests that the conversion should not be interpreted as a single scalar loss of intelligence. Different

properties occupy different layers of the system. The most severe losses appear in specialist factual retrieval, arithmetic, and language modelling. In contrast, the model retains useful performance on a subset of commonsense and plausibility tasks, shows substantial hidden-state diversity, and can be adapted to new downstream objectives.

This stratification is the central empirical result of the study. It also changes how a compressed model should be evaluated. If the intended product only requires a narrow task after supervised adaptation, the relevant question is not whether the compressed model retains all of the teacher's world knowledge. The relevant question is whether the compressed substrate retains enough capacity to learn the required task at the target deployment cost.

10 Deployment and Efficiency Characterisation

The capability study establishes what the model retains. We now examine the engineering consequence of packing the ternary representation.

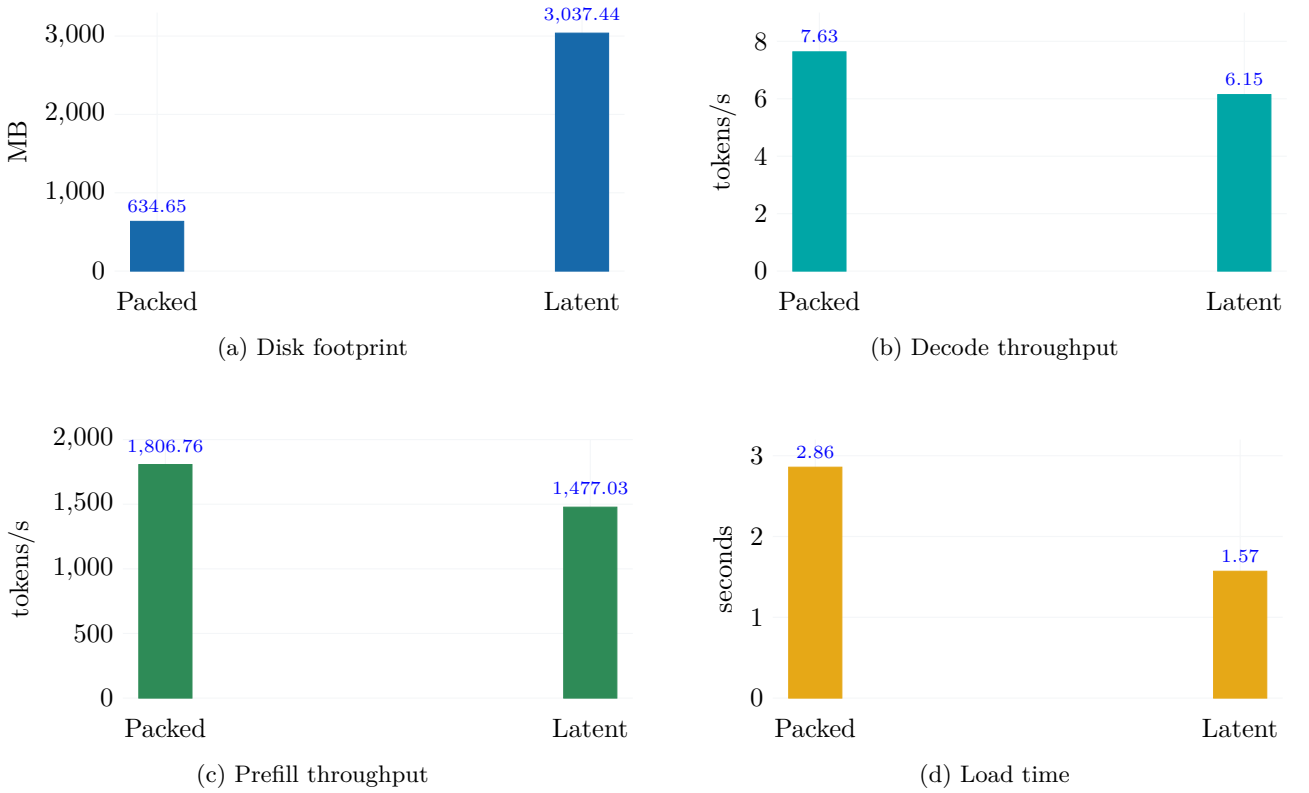
The current deployment comparison is between a *packed* representation and the measured *latent* representation. Because the exact runtime semantics of the latent representation differ from those of a conventional FP/BF16 baseline, we do not use this comparison to claim a universal speedup over all full-precision implementations. Instead, it is reported as a representation-level deployment measurement.

Table 11: Packed versus latent deployment measurements from the current benchmark.

Metric	Packed	Latent	Relative change
Disk footprint	**634.65 MB**	3037.44 MB	**4.79× smaller**
Weight memory	1578.36 MiB	1460.59 MiB	+8.1%
Peak memory	1708.98 MiB	1591.22 MiB	+7.4%
Load time	2.858 s	**1.570 s**	1.82× slower
Prefill	**1806.76 tok/s**	1477.03 tok/s	**+22.3%**
Decode	**7.634 tok/s**	6.145 tok/s	**+24.2%**

Three observations stand out from table 11. First, the packed artifact is dramatically smaller on disk: approximately 634.7 MB versus 3.04 GB for the latent representation. This is the clearest storage benefit in the current measurement. Second, the packed path shows approximately 22.3% higher prefill throughput and 24.2% higher decode throughput in this particular comparison. Third, storage compression does not translate one-to-one into runtime memory: peak memory is approximately 1.67 GiB for the packed path and slightly higher than the latent measurement.

This last point is important for product engineering. A compact model file and a low-memory runtime are related but distinct optimisation targets. The deployment stack may need additional buffers, unpacking structures, kernels, or other runtime state. Consequently, the paper reports both disk footprint and peak memory rather than using one as a proxy for the other.



(a) Disk footprint

(b) Decode throughput

(c) Prefill throughput

(d) Load time

Figure 5: Current packed-versus-latent deployment measurements. The strongest observed benefit is storage footprint, while throughput also improves in the measured execution path. Runtime memory and loading remain separate engineering considerations.

11 Failure Modes as Operating-Boundary Conditions

The study includes negative results because a commercially useful technical paper must identify where the system should *not* be used. These results are therefore presented as operating boundaries rather than as the central narrative.

Table 12: Selected limitations and what they imply for deployment.

Consideration	Current evidence	Engineering implication
Specialist knowledge	MMLU / ARC-C / MedMCQA near chance	Do not treat Cloe as a drop-in replacement for a broad knowledge model.
Arithmetic	GSM8K ≈ 2.0%	Applications requiring reliable symbolic arithmetic need dedicated mechanisms.
Truthfulness	TruthfulQA 28.3%; high unanswerable-question confabulation	Safety-sensitive factual deployment requires task-specific safeguards.
Language modelling	LAMBADA perplexity 12.1× worse; WikiText-2 3.1× worse	General free-form generation quality is materially degraded.
Whole-checkpoint compression	Embeddings remain BF16; overall checkpoint ≈6.47 bits/parameter	The 1.58-bit figure should be reserved for the ternary weight representation.
Runtime memory	Packed peak ≈1.67 GiB	Disk compression should not be interpreted as equal runtime-memory compression.
Fine-tuning uncertainty	Single reported seed	Additional seeds are needed for production-grade variance estimates.
Quantiser comparison	No PTQ control in the current study	The study does not claim QAT is superior to every alternative conversion recipe.

The limitations do not invalidate the main capability-stratification result. Instead, they define its scope. The current Cloe checkpoint should not be presented as a general-purpose replacement for the teacher. The stronger and more defensible interpretation is that extreme ternarisation produces a model with a distinctive capability profile: broad pretrained knowledge is substantially damaged, but representational diversity and task-specific learnability remain useful.

12 Application and Product Implications

The technical findings suggest a deployment paradigm different from simply trying to preserve every capability of a large general-purpose model.

12.1 From knowledge preservation to task-specific adaptation

The combined results can be viewed as a three-stage transformation:

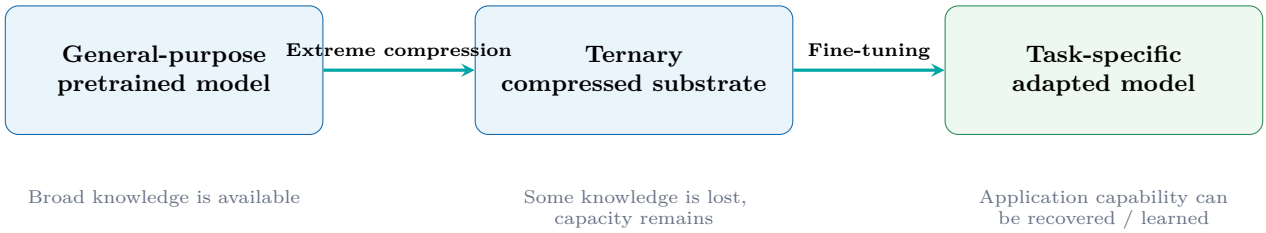


Figure 6: Conceptual deployment thesis suggested by the combined experiments.

The proposed interpretation is therefore not that every application should replace a general-purpose model with Cloe. Instead, applications with narrow or well-defined objectives may value a compact substrate whose task-specific capability can be learned after compression.

Table 13: Potential application mapping. These are deployment hypotheses, not claims that every scenario has been experimentally validated in this study.

Scenario	Why compression matters	Evidence / condition
Edge / constrained inference	Smaller model artifact can reduce storage and transfer burden	Current packed artifact is 634.7 MB.
Private or local assistants	Smaller local model can reduce dependence on remote inference	Requires further end-to-end device validation.
Task-specific classifiers	General knowledge is less important than adaptation	SST-2 reaches 95.6% matched-teacher retention.
Specialised summarisation	Task learnability matters more than broad pretrained knowledge	XSum reaches 79.4% teacher retention after adaptation.
Cost-sensitive inference	Throughput and representation size can affect serving economics	Current packed path shows +22.3% prefill and +24.2% decode versus latent measurement.
General-purpose factual assistant	Broad pretrained knowledge is essential	**Not recommended from current evidence.**

The market-facing implication should therefore be framed around *deployment economics and specialization*, not around an unsupported claim that ternary conversion preserves general intelligence. The current results provide evidence for compact storage, measurable inference benefits in the tested packed path, and strong downstream adaptation on selected tasks. They do not establish a universal advantage across hardware, model scales, or workloads.

13 Discussion

13.1 A capability-stratified view of quantisation

The most useful conceptual outcome of the study is a separation of three properties that are often conflated:

1. **Stored knowledge:** substantially reduced, particularly for specialist factual information.
2. **Representational capacity:** not obviously destroyed, as indicated by effective rank.
3. **Learnability:** remains substantial on the tested downstream tasks.

This separation provides a more informative description than calling the conversion simply "successful" or "failed." It explains why a model can simultaneously perform poorly on broad knowledge benchmarks and still reach near-teacher performance after supervised adaptation on a targeted task.

13.2 Why the evaluation methodology is part of the contribution

The initial MMLU result demonstrates that low-bit models can expose failure modes that conventional evaluation pipelines are not designed to distinguish. Answer-letter bias can convert a degenerate output policy into an apparently meaningful accuracy number. Likewise, retention ratios can become misleading when the teacher itself collapses on a task.

The practical lesson is that low-bit model evaluation should combine behavioural benchmarks with representation-level and learning-level diagnostics. A useful evaluation stack therefore contains:

Benchmark → Guard → Representation probe → Matched adaptation

This is particularly important when the research goal is to understand the *deployment envelope* rather than merely to rank checkpoints.

13.3 What the results do not establish

The study does not establish that QAT conversion is universally superior to PTQ, that Cloe is a general-purpose replacement for Qwen3.5-0.8B, or that the current packed benchmark translates directly into lower cost on every target device. Similarly, the 1.5821 bits/weight entropy describes the ternary weight tensors, whereas the complete checkpoint remains larger because the embedding/LM-head structure is not ternarised.

These boundaries are deliberately explicit because they separate measured facts from the product hypotheses that follow from them.

14 Conclusion

This study investigates what happens when a pretrained small language model is converted into a ternary representation using quantisation-aware training. The answer is not uniform degradation. Instead, the conversion produces a structured capability profile.

The most significant loss is specialist pretrained knowledge: the linear probe falls from 43.76% for the BF16 teacher to 26.19% for Cloe, close to the 25% chance level. At the same time, Cloe retains meaningful performance across a subset of commonsense and plausibility-oriented benchmarks, with a mean 77.0% teacher retention on ten demonstrably interpretable tasks. Effective rank remains high, and downstream adaptation is particularly encouraging: Cloe reaches 95.6% of the matched teacher on SST-2 and 79.4% on XSum.

The engineering measurements add a second dimension. The current packed artifact occupies approximately 634.7 MB on disk and shows higher prefill and decode throughput than the measured latent representation. Runtime memory, however, does not fall in direct proportion to disk footprint, demonstrating why storage, runtime memory, and throughput should be measured separately.

Taken together, the results support a practical thesis:

> **Extreme quantisation can destroy broad pretrained knowledge without destroying the ability to represent and learn useful application-specific behaviour.**

For a startup developing efficient AI systems, this distinction is more valuable than a claim of universal model replacement. It suggests a path toward compact, task-specific models in which the objective is not to preserve every capability of a general-purpose teacher, but to preserve enough computational substrate to learn and execute the capability that the product actually needs.

A Reproducibility Checklist

Table 14: Recommended additions before external publication.

Item	Current status	Publication recommendation
Exact training hyperparameters	Reported internally	Include full configuration.
Random seeds	Single seed for fine-tuning	Add multiple seeds where feasible.
Hardware specification	Required	Report CPU/GPU, RAM/VRAM, software versions.
Kernel implementation	Required	Release or describe packed-kernel implementation.
Complete 29-task table	Available internally	Move full table to supplement.
Raw answer distributions	Available internally	Include diagnostic distributions in supplement.
PTQ baseline	Not included	Add if making a conversion-method claim.
Full BF16 deployment baseline	Not included in current comparison	Add before claiming universal speedup.
Packing memory profile	Partially measured	Report allocation breakdown if available.
Bibliographic metadata	To be finalised	Replace placeholder references with verified citations.

B Reference placeholders

The project materials identify the following literature directions. Bibliographic metadata should be verified and completed from the final reference manager before external submission.

1. BitNet b1.58 — native ternary / 1.58-bit language model training.
2. CAT-Q — ternary conversion of pretrained language models.
3. PTQTP — dual trit-plane representation for ternary post-training quantisation.
4. ScaleQ-1.58 — extension of ternary post-training calibration.
5. Ternary Mamba — pretrained-model conversion using grouped quantisation, distillation, and low-bit adaptation.
6. Falcon-Edge — native 1.58-bit pretrained model family.
7. Straight-through estimator (STE) literature for quantisation-aware training.
8. RMSNorm / low-precision linear-layer implementation literature.

> **Editorial note for the research team.** Before submission, all reference entries, model-version identifiers, hardware specifications, benchmark denominators, confidence intervals, and the exact definition of the “latent” deployment representation should be verified against the reproducibility artifacts. In particular, the current packed-versus-latent numbers should not be relabelled as a BF16 baseline without an explicit benchmark establishing that equivalence.