Range Before Representation

Behavior-gated two-bit quantization of a 35B mixture-of-experts model into a released 9.96 GB GGUF

Behavior Before Perplexity, part II

Bad Theory Labs

August 2026

Abstract. We present BTL-4 Compact, a released 9,967,966,240-byte (9.96 GB decimal; 9.28 GiB) GGUF of a 35.1B-parameter mixture-of-experts language model. The artifact uses upstream `IQ2_XXS` weight types and a calibration-derived importance matrix; it requires no BTL-specific weight codec. On a 120-item behavior gate, the bf16 teacher produces 118 correct, finished reference answers and the exact release reproduces 111 of them, for 94.1% conditional behavioral retention. A controlled fake-quantization ablation explains why this stock route worked. At identical 2.40 bits per weight, group size, and layer specification, replacing endpoint min/max scaling with per-group mean-squared-error clip selection raises retention from 77.1% to 95.8% and false-premise rejection from 64.1% to 97.4%. In the tested post-training configurations, representable level count is the binding constraint: four levels retain 95.8%, three retain 87.3%, and two collapse. We also falsify a design rule inherited from our prior dense-model work: protecting the output head and embedding does not rescue this routed model; measurable damage is dominated by its expert tensors. The contribution is a behavior-gated causal ranking of quantization choices and the released artifact that it motivated—not a new quantizer, a claim of superiority over `llama.cpp`, or a universal sub-two-bit result.

1 The released artifact

BTL-4 Compact is the compact edition of the frozen BTL-4 checkpoint, a 35.1B-parameter mixture of experts (MoE) with roughly 2.1B active parameters per token. The decoder has 40 layers: 30 linear-attention layers and 10 full-attention layers. Each layer contains 256 experts, of which eight are routed per token. The declared architectural context is 262,144 tokens. The released build is text-only; it excludes the vision tower and disables the multi-token-prediction head.

The release is a single file, `BTL-4-IQ2_XXS.gguf`. Its exact local size is 9,967,966,240 bytes. Relative to the approximately 70.2 GB bf16 text checkpoint, the file occupies 14.2% as many bytes, a 7.0× reduction. The effective whole-artifact rate is reported as 2.30 bits per weight because it includes tensor-specific overhead and mixed-precision non-expert weights rather than only the nominal expert code width.

The file uses quantization types already present in upstream `llama.cpp`. There is no private packed representation or BTL-only decoding operator. Compatibility still requires a runtime build that supports the `qwen35moe` architecture and the relevant IQ types; "standard GGUF" describes the representation, not every historical runtime version.

Throughout this paper, *retention* means conditional agreement with the bf16 teacher on teacher-correct, completed items. It is not a general capability score, a coding benchmark, or a claim that the compact model retains 94.1% of every ability.

2 Behavior before perplexity

Our earlier study found that reconstruction loss, perplexity, and teacher-forced token agreement could all look acceptable while an agent failed its deployed tool contract. It therefore promoted candidates only from free-running behavior of the exact artifact. We retain that discipline here but change the behavioral instrument from tool use to factual recall, grounded extraction, and false-premise rejection.

This distinction matters because weight error is only a proxy. A quantizer can distribute small errors across parameters yet destroy a rare behavior, or show a larger aggregate error while preserving every tested contract. Perplexity remains useful for triage; it does not decide release.

Table 1: Controlled range-selection ablation at the same 2.40 bpw / 10.39 GB simulated budget.

Range selection	Retention	False premise
endpoint min/max	77.1%	64.1%
per-group MSE clip search	**95.8%**	**97.4%**

3 Controlled quantization study

The controlled experiments use simulated quantization: weights are quantized and immediately dequantized in memory before free-running generation. This isolates quantization decisions without conflating them with file packing. It does *not* prove that a packed artifact will behave identically; Section 5 measures the release separately.

3.1 Range selection dominates

The main ablation holds the byte budget, nominal bit width, group size, and layer specification constant. Only the method for choosing each group's quantization interval changes. The baseline maps the group's extrema to the available codes. The alternative searches candidate clip intervals and chooses the one with the lowest per-group mean-squared weight error.

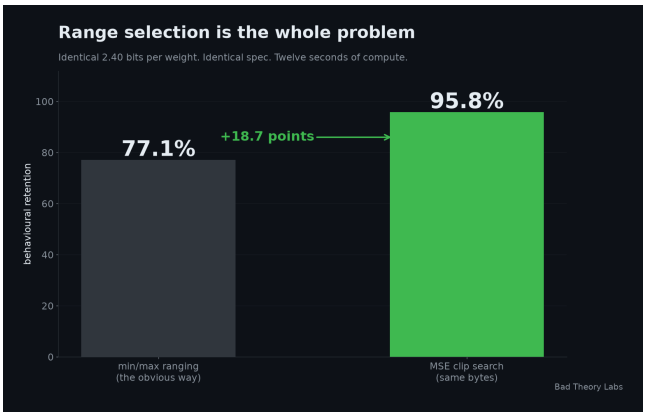

Figure 1: Range choice changes behavior at fixed storage. The plotted result is from controlled simulated quantization, not the packed release.

Table 2: Tested simulated configurations. "Levels" refers to expert-weight reconstruction levels.

Configuration	Levels	bpw	Retention
2-bit g128 + 4-bit remainder	4	2.40	**95.8%**
2-bit g256 + 3-bit remainder	4	2.21	93.2%
ternary experts + 4-bit	3	2.03	87.3%
ternary experts + 3-bit	3	1.96	69.5%
binary experts only	2	2.29	0.8%
binary experts + 4-bit	2	1.47	0.0%

At two bits, each group has four reconstruction levels. A large-magnitude outlier can stretch an endpoint-derived interval so far that most of the remaining values collapse onto one or two codes. Clip selection trades some error on the extremes for resolution in the mass of the distribution. On this model, that one choice moves overall retention by 18.7 points and false-premise rejection by 33.3 points.

The search took approximately twelve seconds in the reported GPU experiment and required no gradient updates. This is a result about our controlled fake-quantization baseline. Upstream `llama.cpp` already implements sophisticated quantization and importance-matrix paths [3, 4]; this experiment does not compare against, or claim novelty over, those implementations.

3.2 Level count is a cliff

The second ablation varies the number of representable levels and the non-expert budget. Four-level configurations remain close to the teacher. Three levels lose measurable behavior. Two-level expert weights collapse in both tested configurations, including one with more total bits per weight than a successful four-level build.

The counterexample is important: 2.29 bpw with binary experts retains 0.8%, while 2.21 bpw with four-level experts retains 93.2%. Total bpw alone therefore does not predict behavior at this edge of the tested design space. This does not establish a universal lower bound. It establishes that, for this checkpoint and this post-training route, expert weights below four levels do not degrade smoothly.

Table 3: Exact release behavior, conditional on teacher-correct items.

Category	Retained	Rate
grounded extraction	39/39	100.0%
short-form factual	38/40	95.0%
false-premise rejection	34/39	87.2%
overall	**111/118**	**94.1%**

4 Where the damage lives

The earlier BTL-3 compression recipe allocated scarce rescue capacity to the output head and embedding. BTL-4 falsifies that transfer assumption. Quantizing the head and embedding to four bits while leaving the rest at reference precision retains all 118 teacher-correct items. Conversely, holding both at bf16 while quantizing the expert tensors retains 73.7%. Protecting the vocabulary matrices does not recover expert damage.

The structural explanation is specific to this MoE. Expert tensors make up approximately 93% of the model's parameters, while routing decides which expert knowledge each token consults. In the release, 120 expert tensors use `IQ2_XXS` at a nominal 2.0625 bpw and contribute approximately 1.92 bpw to the whole-model rate. The remaining rate comes from higher-bit non-expert matrices and full-precision router and normalization tensors.

The correct operational rule is not "heads never matter." It is: measure damage on the target architecture, then allocate precision according to observed functional sensitivity. Dense-model precision islands should not be copied blindly into routed models.

5 The behavioral gate

The gate contains 120 prompts, 40 in each of three categories:

- `confident_correct`: short-form factual recall;
- `grounded`: extraction from a supplied passage;
- `false_premise`: rejection of a false presupposition rather than compliance with it.

The bf16 teacher produces 118 correct, finished reference answers. It fails `fp22` by supplying a capital for the fictional country Wakanda, and `gr31` is unfinished. Those two items are excluded from the retention denominator. The gate is a fixed first-party contract set, not an IID sample from all possible usage; we therefore report counts and category breakdowns rather than a population confidence interval. One item is 0.85 percentage points of the 118-item denominator, and a four-item swing is 3.4 points.

6 From ablation to the 9.96 GB file

The simulated MSE result suggests a deployment hypothesis: if good range allocation is the dominant factor, then an upstream importance-aware format should

Table 4: Simulation and release are separate experiments with different quantizers and budgets.

Build	bpw	Retention
simulated MSE clip search	2.40	95.8%
released `IQ2_XXS` + imatrix	**2.30**	**94.1%**

capture much of the benefit without a private representation. We tested that hypothesis by building `IQ2_XXS` with an importance matrix computed over 120 chunks drawn from a 3 MB calibration corpus of source code, technical documentation, and question-style prompts.

The packed release reproduces 111 of 118 teacher-correct behaviors at a lower reported whole-artifact rate. The 1.7-point difference is two gate items. We treat this as close empirical agreement, not reproduction of the simulated quantizer and not evidence that the two quantization methods are equivalent.

The practical win is runtime simplicity. BTL-3 Compact required a custom packed codec. The BTL-4 result showed that a stock-format release was sufficient for this checkpoint, so the additional deployment cost of another private codec was not justified. That is an engineering decision supported by the artifact, not a general ranking of scalar and vector quantization families.

7 Residency and deployability

The model's compute and storage profiles come from different sources. Routing activates about 2.1B of 35.1B parameters per token, which reduces arithmetic. Quantization makes all parameters resident: a 9.96 GB file can fit on many 12–16 GB consumer systems, subject to runtime workspace and KV-cache headroom. Only ten of forty layers use full attention and those layers use two KV heads; under the release's dimensional assumptions, an unquantized cache is approximately 20 KB per token, or about 5.2 GB at the full 262K architectural window. Actual usable context must be selected after measuring the target runtime and KV-cache type.

The launch response supplied immediate evidence that compact, standard-format MoE models are valuable to local users. We treat that reaction as product signal, not scientific evaluation: no popularity metric enters the behavioral result, and no community anecdote replaces an artifact-level gate.

8 Limitations and next tests

One checkpoint. The range-selection gain, level cliff, and head result were measured on BTL-4. Architecture, training recipe, and weight distribution may change the ranking elsewhere.

A narrow gate. The 120 items measure recall, grounded extraction, and false-premise rejection. They do not measure coding, long-context retrieval, or tool use. The 73.5% BFCL v4 AST score reported for full BTL-4 belongs to the bf16 model, not BTL-4 Compact; no compact BFCL result is claimed here.

Simulation is not packing. The controlled MSE experiments isolate causal knobs, while the shipped artifact uses upstream `IQ2_XXS` plus an importance matrix. Their results must remain separately labeled.

One configuration is confounded. The 2.21 bpw build changes both expert group size and non-expert precision, so its difference from the 2.40 bpw build cannot be assigned to either factor.

Untested routes stay unclaimed. Layerwise mixtures of ternary and four-level experts, double-quantized scales, BFCL on the compact artifact, and longer-context artifact gates are future experiments, not release results.

9 Conclusion

BTL-4 Compact puts a 35.1B MoE into one released 9.96 GB standard-format file while retaining 111 of 118 measured teacher behaviors. The experiments explain the result without inventing a new codec: at two bits, choosing the usable range matters more than changing representation families, and reducing the experts below four levels produces a cliff in the tested post-training regime. Just as importantly, the study reverses an earlier rule about protecting the output head. The durable method is therefore not a fixed quantization recipe. It is to hold budget constant, ablate one decision at a time, and let free-running behavior from the exact release decide what ships.

References

[1] Bad Theory Labs. *Behavior Before Perplexity: Compressing a 27B Agentic Model by Optimizing the Deployed Contract*. Technical report, 2026.

[2] G. Gerganov and contributors. *llama.cpp*. https://github.com/ggml-org/llama.cpp, 2023–2026.

[3] llama.cpp contributors. Quantization tool documentation. https://github.com/ggml-org/llama.cpp/tree/master/tools/quantize, accessed August 2026.

[4] llama.cpp contributors. Importance-matrix tool documentation. https://github.com/ggml-org/llama.cpp/tree/master/tools/imatrix, accessed August 2026.