

Compressing Onboard Language Models for Space Robots under a Behavioral-Fidelity Requirement

Jia Deng, Zizhe Wang, Xin Lu, Yi Huang, and Zehao Jia

Astroup Lab¹, Beijing, China astroup.com

{dengjia, wangzizhe, luxin, jiazhao}@astroup.com

24B903024@stu.hit.edu.cn

Abstract—Language models are being explored as the interface through which space robots interpret operator commands and select among verified skills. Running them onboard calls for low-bit compression, and flight use adds requirements that generic post-training quantization leaves implicit: a fixed memory budget and storage format, a static artifact under configuration control, and behavior close to the full-precision model validated on the ground. We treat compression as a ground-segment step and study QSelect, which composes one static low-bit model from compatible AWQ, GPTQ, and AutoRound checkpoints by assigning each linear module to one quantizer family and scoring every substitution at the final next-token distribution. The default fidelity mode needs no labels or tuned thresholds and accepts a substitution only when the forward KL divergence from the BF16 reference decreases; an optional target-adapted mode adds answer likelihood on a small labeled set. On DeepSeek-R1-Distill-Qwen-7B, Llama-3.1-8B, and Llama-3.2-3B at 3 and 4 bits, composed models reach lower mean held-out perplexity than every single-family model in all six settings, and the fidelity mode lowers perplexity and KL further. In a matched scaffold, final-output scoring gives lower held-out KL than representation-similarity scoring in all 12 paired runs. With labels from training or development splits of the evaluated task families, the target-adapted mode raises 3-bit pooled accuracy on the two Llama models by 4.3 and 4.5 points over the best single-family model. At 3 bits, unquantized embedding and output layers make up 43–46% of the packed artifact. All evaluations use general-domain benchmarks on a ground GPU; space-domain tasks and flight hardware are left to future work.

Index Terms—Space robotics, onboard language models, post-training quantization, mixed quantizer assignment, on-orbit servicing, model compression.

I. INTRODUCTION

SPACE robots are expected to work with less direct supervision. On-orbit servicing combines approach, capture, and manipulation of client spacecraft, and free-flying robots such as Astrobee serve as test beds for autonomy inside the International Space Station (ISS) [1]. Language and vision-language models offer a natural interface for such tasks: they can map an operator’s request to a sequence of pre-verified skills [2] or, in end-to-end form, to robot actions [3], [4]. Space-specific agent frameworks now apply this interface to on-orbit servicing [5]. These systems are usually evaluated with large models on terrestrial computers; SpaceMind, for example, uses Qwen3-VL-235B-A22B throughout [5], a model whose weights alone would far exceed the memory of current flight computers.

Running the model onboard removes the dependence on communication windows, round-trip delays, and limited link

bandwidth, but onboard compute is scarce [6]. Perseverance’s primary computer, for example, is a radiation-hardened RAD750 running at up to 200 MHz with 256 MB of DRAM [7]; its global localization instead runs on a commercial Snapdragon 801 from the Ingenuity helicopter base station [8]. Language models are now reaching orbit. An AI assistant that combines ground and onboard models was delivered to China’s space station in July 2025 [9], and ASTREA ran a 1.54-billion-parameter Qwen2.5 model quantized to 4 bits on a flight-qualified, CPU-only system-on-chip in an external ISS payload, with response times from about 40 s to over 8 min [10]. For onboard language models, compression is therefore a prerequisite rather than an optimization.

For flight use, the relevant question is not only how small a model can become but also what the compressed model must preserve. We identify three properties. First, the artifact must fit a fixed memory budget and use a storage format that the onboard runtime supports. Second, it must be a static object that can be versioned, verified, and rolled back like other flight software. Current machine-learning guidance for space engineering does not yet cover large language models [11]; until it does, the preparation process has to make these properties explicit. Third, its behavior must stay close to that of the full-precision reference that was validated on the ground. The third property suggests measuring compression error where the validated behavior is defined: at the model’s output distribution.

Standard post-training quantization (PTQ) for large language models (LLMs) applies one quantizer family throughout a model. GPTQ uses approximate second-order error compensation [12], AWQ protects activation-salient channels by equivalent scaling [13], and SignRound, implemented in AutoRound, optimizes rounding and clipping [14]. These families embody different error models, and CALM has shown that choosing a quantization algorithm per Transformer block can help [15]. CALM scores candidates by linear centered kernel alignment (CKA) between intermediate representations [16]. Representation similarity is a reasonable heuristic, but it is not the quantity that an acceptance test of the onboard model would examine.

We therefore treat compression as a ground-segment preparation step (Fig. 1) and study QSelect, which composes one static low-bit model from compatible homogeneous checkpoints. Each linear module is assigned to one quantizer family, and every candidate substitution is scored through the final next-token distribution of the partially assigned model. In its

¹Astroup Lab: 向上跃升（北京）科技有限公司.

default fidelity mode (QSelect-KL), the procedure is label free and accepts a substitution only when the forward KL divergence from the BF16 reference strictly decreases; the rule has no tuned threshold. An optional target-adapted mode (QSelect-TA) adds a target-answer likelihood term for missions that can provide a small labeled set, with fixed margins that bound the KL increase allowed at each step. Scoring with the acceptance quantity itself, rather than a cheaper proxy, is affordable in this setting: a one-pass module-wise selection takes one to a few GPU-hours on the ground for the models studied, excluding candidate generation, is paid once per model release, and never changes the bit width or storage format of the result. The memory limits that make compression necessary also keep onboard models small, and the selection cost shrinks with the model.

Experiments on three open models from 3B to 8B parameters at 3 and 4 bits show four things. Composed models reach lower mean held-out perplexity than every single-family model in all six settings, and the fidelity mode lowers perplexity and KL further. Final-output scoring yields lower held-out KL than representation-similarity scoring in all 12 paired runs of a matched scaffold and lower KL than a local reconstruction criterion in all four tested settings. Module-wise decisions mainly lower 3-bit perplexity, at several times the ground time of layer-wise decisions, and extra passes add little. Packed artifacts are 63–70% smaller than BF16, as for any single-family model at the same bit width, and an analytic breakdown shows that unquantized embedding and output layers make up 43–46% of what remains at 3 bits. All experiments use general-domain language-modeling and reasoning benchmarks on a ground GPU; we do not evaluate space-domain tasks, flight hardware, or radiation effects.

Our contributions are as follows.

- A requirements analysis for compressing onboard language models for space robots. It states five requirements—fixed format and budget, a static artifact, behavioral fidelity to a ground-validated reference, scarce mission labels, and bounded ground compute—and maps them to design choices and to the evidence reported here (Section III).
- QSelect, which builds on CALM’s per-block algorithm selection [15] with a final-output score, finer decision units, a threshold-free and label-free fidelity mode, and an optional target-adapted mode (Section IV).
- An empirical study on three open models that compares QSelect with single-family PTQ, representation-similarity, local-reconstruction, and whole-model selection; characterizes the trade-off between ground time and fidelity; and analyzes packed artifacts for onboard deployment (Sections VI and VII).

II. BACKGROUND AND RELATED WORK

A. Language Models in Space Robotics

Grounding language in executable skills [2] and vision–language–action models [3], [4] have made language a practical interface to robot behavior. In space, SpaceMind organizes an on-orbit servicing agent around natural-language skill

modules and tool calls and evaluates it in simulation and on a ground robot [5]. Space-LLaVA fine-tunes a 13B LLaVA checkpoint to adapt a vision–language model to extraterrestrial visual semantics [17], and language-model agents have controlled simulated spacecraft in the Kerbal Space Program Differential Games challenge [18]. These studies establish the interface; they leave open how a model of useful size can be prepared for the computers that space robots carry.

B. Onboard Computing and Flight Software Assurance

Radiation-hardened processors such as the GR740 support general flight software [19], and NASA’s High-Performance Spaceflight Computing processor is intended to provide at least 100 times the computational capacity of current spaceflight computers [20]. Commercial or radiation-tolerant accelerators provide more throughput for bounded workloads, as with the Myriad 2 that ran a cloud-segmentation network on Φ -sat-1 [21] and the Versal XQR family [22]. Perseverance pairs its radiation-hardened primary computer with a commercial Snapdragon 801 for global localization and mitigates radiation effects in software [8]. ASTREA’s 4-bit language model ran on the IMAGIN-e edge-computing payload [10], [23]; at the other end of the range, a satellite carrying an NVIDIA H100 GPU has run the Gemma model in orbit [24]. The ECSS machine-learning handbook addresses the engineering and assurance of learned components, including safety-cage architectures, but does not yet cover large language models and anticipates updates for generative AI [11]. OPS-SAT, which ended in 2024, provided a reconfigurable payload computer for in-orbit software experiments while the bus monitored the payload and could take back control at any moment [25]; we assume the same supervisory pattern for the onboard language model. These considerations motivate the requirements in Section III.

C. Post-Training Quantization of LLMs

Weight-only PTQ methods include GPTQ’s second-order compensation [12], AWQ’s activation-aware scaling [13], AdaRound’s learned rounding [26], and SignRound’s signed-gradient optimization [14]. SmoothQuant, OmniQuant, and QuaRot additionally transform or quantize activations [27]–[29]. Recent work explores flexible low-rank approximation, spectral decomposition, and global information-flow allocation [30]–[32]. Our setting begins after several compatible homogeneous pipelines have produced candidate checkpoints.

D. Heterogeneous Assignment

HAQ and HAWQ allocate bit widths using hardware feedback or Hessian information [33], [34], and InfoQ uses a global information-flow signal for mixed-precision allocation [32]. CALM keeps the precision fixed and selects a quantization algorithm for each Transformer block with linear CKA [15], [16]. CALM and QSelect both assemble a model sequentially, so later decisions are conditioned on earlier ones. They differ in where a candidate is scored—intermediate block representations for CALM, the final next-token distribution for QSelect—and in granularity: QSelect supports blocks,

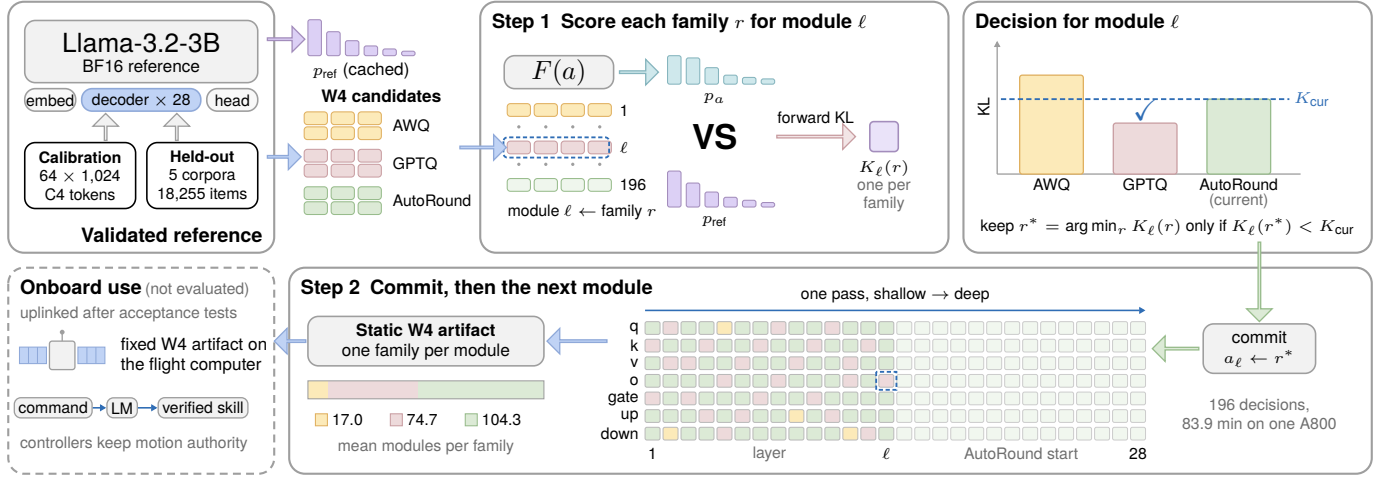


Fig. 1. Overview of QSelect, with Llama-3.2-3B at W4 as the running example. The three W4 candidates share one packed format. Step 1 gives module ℓ of the current model $F(a)$ to family r and scores the result by the forward KL from the cached reference distribution; the lowest-KL family is kept only if it lowers the current KL (right). Step 2 commits the choice and moves on, in one shallow-to-deep pass over all 196 modules. The static artifact is uplinked after acceptance tests; onboard use is not evaluated here. Bar heights and grid colors are schematic; counts, token budgets, and times are measured values (Sections V and VI).

attention or feed-forward components, and individual linear modules. Our CALM baseline is an independent matched-pool reimplementation, not an official reproduction.

E. Output-Aware Calibration and Modular Composition

Knowledge distillation transfers output-distribution information [35], and output-adaptive calibration uses model-output distortion to guide PTQ [36]. AdaRound and BRECQ instead optimize weight-, layer-, or block-local reconstruction losses [26], [37]. QSelect uses the final-output KL only for discrete assignment, not for training candidates. The construction is also related to model stitching [38], although we combine compatible modules from quantized versions of the same checkpoint without learning a stitching layer.

III. GROUND-TO-ORBIT MODEL PREPARATION

A. Deployment Workflow

Fig. 1 summarizes the setting. On the ground, a full-precision (BF16) reference model is validated for its mission role, for example interpreting operator commands and selecting among verified manipulation or inspection skills. Several PTQ pipelines that share a storage format then produce homogeneous low-bit checkpoints. QSelect composes one static model from their modules; the result is packed, recorded with its configuration, and compared with the reference by acceptance tests before it is preloaded before launch or uplinked as a software update. Onboard, the model runs at a supervisory rate inside the autonomy stack. Its outputs pass through the same checks and fallback logic as any other planner output, and conventional controllers keep authority over motion.

B. Requirements

Table I lists five requirements and the corresponding design choices.

R1: fixed memory budget and supported format. The model must fit the memory left after flight software and buffers, and the onboard runtime must be able to execute its storage format. All candidates therefore share the bit width, group size, and symmetric packed integer format, so the footprint of the quantized layers is fixed before selection. AWQ modules additionally carry a static per-channel input scale. The composition introduces no new numerical format, although it may use family-specific module wrappers (Section IV-F). Keeping one bit width also keeps one integer kernel and one memory layout to verify, whereas mixed-precision assignment [33], [34] multiplies both; this is why we select among quantizer families rather than among bit widths.

R2: static artifact under configuration control. Each module keeps exactly one selected candidate. There is no runtime selector, routing, or ensemble, so the onboard network is fixed and can be characterized once on the ground. The assignment vector, with 196 or 224 entries for the models studied here, together with the candidate checkpoint and manifest hashes, identifies the artifact for versioning and rollback.

R3: behavioral fidelity to the validated reference. Ground validation establishes how the reference behaves on mission-relevant inputs. The compressed model should depart from that behavior as little as possible. We measure the departure by the forward KL divergence from the reference next-token distribution to the compressed model's distribution, averaged over text positions; it is zero only when the two distributions agree at every evaluated position. Forward KL weights each position by the reference's own probabilities, so it penalizes most the loss of outputs that the reference considers likely. At one position, the total-variation distance between the two distributions is the smallest probability, over couplings, that the two models sample different next tokens; by Pinsker's and Jensen's inequalities, an average KL of K nats bounds its mean over positions by $\sqrt{K/2}$. The bound holds on reference text rather than on the model's own generations, so it motivates

TABLE I
DEPLOYMENT REQUIREMENTS FOR AN ONBOARD LANGUAGE MODEL AND HOW THEY ARE ADDRESSED

Requirement	Design choice in QSelect	Evidence in this paper
R1 Fixed memory budget and a storage format the onboard runtime supports	All candidates share bit width, group size, and packed integer format; selection never changes the bit width	Table VII, Fig. 3
R2 Static artifact under configuration control	One candidate per module, fixed after selection; no runtime selector, routing, or ensemble; the assignment vector and checkpoint hashes identify the artifact	By construction (Section IV-F)
R3 Behavior close to the ground-validated reference	Candidates are scored by the forward KL from the reference next-token distribution on calibration text; acceptance can check the same quantity on held-out text	Tables II–III
R4 Few or no mission labels	Fidelity mode needs only unlabeled text; the target-adapted mode uses a small labeled set	Tables II and III
R5 Bounded, one-time ground compute	Decision granularity and pass count trade ground time for fidelity	Fig. 2, Table VI

the score rather than guaranteeing free-running behavior. An acceptance test can apply the same quantity to held-out or mission text, alongside task tests, while selection optimizes it on separate calibration text; the remaining gap between selection and acceptance therefore lies in the data rather than in the criterion. For a validated system, a change that raises task accuracy but also raises KL is an unvalidated change in behavior, which is why answer supervision is optional (R4). This paper sets no numerical KL threshold; a mission would derive one from its own validation campaign.

R4: scarce mission labels. Labeled mission data are expensive and often unavailable when a model is prepared. The fidelity mode therefore uses only unlabeled text and the reference model. When a small labeled set exists, the target-adapted mode can use it, but it then departs from the reference on purpose, and acceptance tests must cover that departure.

R5: bounded, one-time ground compute. Selection runs on ground GPUs once per model release, whereas onboard resources are consumed throughout the mission. Spending GPU-hours on the ground is acceptable if it buys fidelity, but the cost should be predictable and adjustable. Decision granularity and the number of passes provide that control, and the cost shrinks with the model, which onboard memory keeps small (Section VI-C).

C. Problem Statement

Let F_{ref} be the validated BF16 reference and let $H^{(1)}, \dots, H^{(R)}$ be homogeneous low-bit checkpoints with the same bit width and format. The task is to choose one candidate family for every quantized linear module so that the composed model minimizes the fidelity score K of Section IV-B, using only unlabeled text. Optionally, a bounded increase in K may be traded for a lower target-answer loss T . Because every candidate for a module has the same bit width and group size, the memory budget is fixed by construction. Exhaustive search is impractical: module-wise assignment of Llama-3.1-8B over three families has 3^{224} possibilities. QSelect therefore uses a greedy search whose cost grows linearly with the number of decision units and candidates.

IV. QSELECT

QSelect operates on a precomputed bank of matched-format module candidates. It scores temporary substitutions in the

current partially assigned model, commits accepted updates according to a fixed search protocol, and exports the resulting static assignment.

A. Candidate Bank and Assignment Space

Let $r \in \{1, \dots, R\}$ index a quantizer family, s a calibration-bank replicate, and $l \in \{1, \dots, L\}$ a target linear module. For every r and s we first construct a homogeneous quantized checkpoint $H_s^{(r)}$. Its module l supplies a family-conditioned candidate

$$\mathcal{B}_{l,s}^{(r)} = \widehat{W}_{l,s}^{(r)}. \quad (1)$$

During assignment, $\widehat{W}_{l,s}^{(r)}$ is stored as a dense effective fake-quantized weight. The notation retains the calibration-bank replicate and the homogeneous pipeline: a candidate is not assumed to be an independently generated function of W_l alone. For AWQ, the selection-time effective weight is

$$\widehat{W}_{l,s}^{(\text{AWQ})} = Q_b(W_l \text{diag}(d_{l,s})) \text{diag}(d_{l,s})^{-1}, \quad (2)$$

where $d_{l,s}$ is the input-channel scale and Q_b is the common symmetric group-wise weight quantizer. A family is supported only when its transformed parameters and static state can be encapsulated as a locally swappable unit with a compatible input–output interface; global rotations, cross-layer transformations, activation quantization, and incompatible activation formats are outside this scope.

Let $\mathcal{C}_l \subseteq \{1, \dots, R\}$ index the candidates available for module l . An assignment $a = (a_1, \dots, a_L)$ chooses $a_l \in \mathcal{C}_l$ for every module. Let $\mathcal{G}$ partition the modules into decision units. The legal families for unit $G \in \mathcal{G}$ are

$$\mathcal{C}_G = \bigcap_{l \in G} \mathcal{C}_l, \quad (3)$$

and the granularity-constrained assignment space is

$$\mathcal{A}_G = \{a : a_l \in \mathcal{C}_l, a_l = a_{l'} \text{ for all } l, l' \in G, G \in \mathcal{G}\}. \quad (4)$$

Singleton units give module-wise assignment; one unit per decoder block gives layer-wise assignment; grouping the query, key, value, and output projections separately from the gate, up, and down projections gives component-wise assignment.

Algorithm 1 QSelect-KL: one-pass fidelity mode

Require: Candidate bank $\{\mathcal{B}_{l,s}^{(r)}\}$, decision units $\mathcal{G}$ in a fixed order, initial assignment a .

```

1:  $K_{\text{cur}} \leftarrow K_s(a)$ 
2: for decision unit  $G$  in the fixed order do
3:    $K_G(r) \leftarrow K_s(a[G \leftarrow r])$  for every  $r \in \mathcal{C}_G$ 
4:    $r^* \leftarrow \arg \min_{r \in \mathcal{C}_G} K_G(r)$ 
5:   if  $K_G(r^*) < K_{\text{cur}}$  then
6:      $a \leftarrow a[G \leftarrow r^*]; K_{\text{cur}} \leftarrow K_G(r^*)$ 
7:   end if
8: end for
9: return static assignment  $a$  and model  $F_s(a)$ .
```

B. Final-Output Fidelity Score

For an assignment a , let $F_s(a)$ be the resulting quantized model. On non-padding next-token positions Ω_{KL} we compute the forward KL divergence from the reference distribution to the quantized-model distribution:

$$K_s(a) = \frac{1}{|\Omega_{\text{KL}}|} \sum_{(x,t) \in \Omega_{\text{KL}}} D_{\text{KL}}(p_{\text{ref},t}(\cdot | x) \| p_{a,t}(\cdot | x)). \quad (5)$$

The distributions use the full vocabulary and temperature one. Reference logits are cached on the CPU in FP16, transferred to FP32, and normalized with `log_softmax`; model logits and KL accumulation use FP32. The cache requires $O(|\Omega_{\text{KL}}|V)$ host memory for vocabulary size V . The reference is evaluated once; every later evaluation runs only the candidate model.

C. Fidelity Mode

The fidelity mode, QSelect-KL, is the default. At each decision unit it evaluates every legal family, proposes the one with the lowest K_s , and accepts it whenever its score is strictly lower than that of the current assignment (Algorithm 1). No labels, margins, or tuned thresholds are involved, so the search is a greedy block-coordinate descent on Eq. (5): the score of the committed assignment never increases. Exact ties follow the fixed order AWQ, GPTQ, AutoRound.

D. Target-Adapted Mode

Some missions can supply prompts x with target answers $y = (y_1, \dots, y_{|y|})$, for example procedure questions with known answers. The target-adapted mode, QSelect-TA, uses the token-averaged target-answer negative log-likelihood (NLL)

$$T(a) = - \frac{\sum_{(x,y) \in \mathcal{D}_{\text{proxy}}} \sum_{j=1}^{|y|} \log p_a(y_j | x, y_{<j})}{\sum_{(x,y) \in \mathcal{D}_{\text{proxy}}} |y|}. \quad (6)$$

This is a teacher-forced answer-likelihood signal, not free-running generation or accuracy. At each unit, QSelect-TA first forms a near-best-KL shortlist,

$$\mathcal{S}_G = \left\{ r \in \mathcal{C}_G : K_G(r) \leq 1.01 \min_{r' \in \mathcal{C}_G} K_G(r') \right\}, \quad (7)$$

and proposes the shortlist candidate with the lowest T . The proposal is accepted through either a fidelity-improvement branch,

$$K_G(r^*) \leq 0.999 K_{\text{cur}}, \quad (8)$$

or a target-adaptation branch,

$$K_G(r^*) \leq 1.002 K_{\text{cur}} \quad \wedge \quad T_G(r^*) \leq 0.997 T_{\text{cur}}. \quad (9)$$

The 1% shortlist, 0.1% KL-improvement margin, 0.2% KL tolerance, and 0.3% NLL-improvement margin were fixed before the evaluated comparisons and shared across all settings; they were not derived theoretically or exhaustively swept. The margins bound how far a single accepted step can move away from the reference. The full procedure, including lazy NLL recomputation, is given in Algorithm 2 in the Appendix.

E. Search Protocol

The main protocol starts from the homogeneous AutoRound assignment, scans decoder blocks from shallow to deep, and processes the query, key, value, output, gate, up, and down projections in that order. It performs one pass. Candidate evaluation is functional, so temporary insertions do not alter the committed assignment. Initialization, granularity, traversal order, and pass budget are protocol variables, and Section VI-D reports their effects.

F. Ground Cost and Static Export

Let $U = |\mathcal{G}|$, let R_u and S_u be the candidate and shortlist sizes for unit u , and let $C_{\text{model}}(N)$ be the cost of a full-model pass over N scored tokens. Assignment requires approximately

$$O\left(\sum_{u=1}^U R_u C_{\text{model}}(N_{\text{KL}}) + \sum_{u=1}^U S_u C_{\text{model}}(N_{\text{proxy}})\right), \quad (10)$$

where the second term is absent in the fidelity mode. Candidate generation and $O(RP)$ candidate-bank storage are separate costs, where P is the number of stored target-module parameters. With fixed width, no prefix or suffix caching, U proportional to depth D , and a full forward pass proportional to D , the direct module-wise implementation grows approximately quadratically with depth. All of this cost is incurred on the ground.

Native export packs the selected modules after assignment and does not alter the selected families. Each packed module stores low-bit integer codes, group scales, and the zero-point field of the packed layout, which is constant under the symmetric format; AWQ modules additionally retain $d_{l,s}$ as static metadata and apply $x \mapsto x \text{diag}(d_{l,s})^{-1}$ before the packed linear operator. The exported model is therefore statically composed of family-specific modules, with no dynamic selector or ensemble. The assignment vector, the candidate checkpoint hashes, and the calibration manifest hashes form the configuration record of the artifact.

V. EXPERIMENTAL SETUP

Scope. The experiments evaluate the ground segment of Fig. 1. Our evaluation contains no space-domain data; general-domain language-modeling and reasoning benchmarks stand in for the reference behavior that a mission would validate. Each job runs on one NVIDIA A800-SXM4-80GB GPU.

Models and low-bit format. We use deepseek-ai/DeepSeek-R1-Distill-Qwen-7B [39], [40] (DS-R1-Qwen-7B), meta-llama/Meta-Llama-3.1-8B [41], and meta-llama/Llama-3.2-3B [42]. BF16 references are loaded without low-bit PTQ. W3A16 and W4A16 (W3 and W4) denote symmetric 3- or 4-bit weights, BF16 activations, no zero point, and input-channel groups of 128. We quantize the query, key, value, output, gate, up, and down projections; embeddings, normalization layers, and the output head remain BF16. Unless stated otherwise, results use dense effective fake-quantized weights; packed artifacts are examined in Section VII.

Candidate generation. The three families use the same replicate-specific C4 manifest [43] of 128 packed sequences of 2,048 tokens. They do not use matched compute or identical effective token budgets. GPTQ uses GPTQModel 7.1.0 with symmetric quantization, desc_act=false, true_sequential=true, and damping 0.05. AutoRound 0.13.0 uses the SignRound recipe of that release with 200 iterations, batch size one, gradient accumulation one, and packed_tensors; unlisted options use that release’s defaults. Our matched-format AWQ implementation gathers at most 1,024 activation tokens per module, searches a 20-point scaling grid, and applies Eq. (2). These configurations are candidate generators in a shared output format, not claims about each method’s strongest published configuration.

Calibration banks and selector data. Replicates 17, 23, and 41 index candidate generation and the 64-sequence, 1,024-token C4 sample used for KL scoring. The labeled proxy and held-out evaluation manifests are fixed. For DS-R1-Qwen-7B, the proxy uses 128 GSM8K training, 128 MATH training, and 256 MMLU development examples; the Llama proxy additionally includes 128 ARC-Challenge training and 256 HellaSwag training examples. GPQA is never included. A shared-core control applies the DS proxy to Llama-3.1-8B.

CALM reimplement. CALM-reimpl. follows sequential block-wise assembly and selects one family per block with biased linear CKA. It compares the BF16 and candidate outputs of `mlp.down_proj`, the pre-residual feed-forward contribution, over 64 C4 sequences of 1,024 tokens, accumulating global centered feature statistics in 1,024-token chunks. The candidate pool is AWQ/GPTQ/AutoRound rather than CALM’s original AWQ/GPTQ/SmoothQuant pool.

Intrinsic evaluation. Each replicate is evaluated on fixed manifests of 128 sequences of length 2,048 from C4, WikiText-2 [44], and three local corpora denoted Code, Math, and Wiki. Tables report the arithmetic average of the five corpus-level perplexity (PPL) values and the corresponding maximum, together with corpus-averaged and maximum KL. We label these columns *Avg.* and *Max-corpus* because they are descriptive summaries rather than pooled language-modeling metrics.

Downstream evaluation. The suite contains GPQA Diamond (198) [45], ARC-Challenge (1,172) [46], GSM8K (1,319) [47], HellaSwag (1,024) [48], MATH-500 (500) [49], and MMLU (14,042) [50]. The protocol is zero-shot, with no chat template or system prompt. Multiple-choice tasks are reformatted as letter-answer questions and scored by average answer-letter log probability, so these scores are not directly comparable with standard continuation-likelihood harness values. GSM8K and MATH-500 use greedy generation with at most 256 new tokens and normalized final-answer matching. Micro pools 18,255 examples and is dominated by MMLU; Macro is the unweighted mean of the six task accuracies. The Appendix gives further protocol details.

Modes evaluated. QSelect-KL is the default mode, but most protocol studies were run with QSelect-TA: the granularity, traversal-order, initialization, candidate-pool, pass-budget, and whole-model controls, the packed export, and the module-wise 7B/8B selection times all use the target-adapted mode. QSelect-KL was run with the main protocol in all six settings. For the four 7B/8B settings, its results are available as paired changes relative to QSelect-TA (Table III), and its absolute values in Table II are derived from them ([†], without standard deviations); on Llama-3.2-3B they are measured directly. Because the fidelity mode evaluates a subset of the target-adapted mode’s forward passes (Section IV-F), the target-adapted selection times bound those of the fidelity mode under the same protocol.

VI. RESULTS

A. Composed Versus Single-Family Models

Table II compares both modes with the homogeneous candidate generators. At W3, QSelect-TA lowers mean Avg. PPL by 7.4%, 5.8%, and 14.6% relative to the best single-family model on DS-R1-Qwen-7B, Llama-3.1-8B, and Llama-3.2-3B, and it raises pooled (Micro) accuracy on the two Llama models by 4.3 and 4.5 points over the best single family. These task gains use labeled proxy examples from training or development splits of five of the six evaluated Llama task families (Section V), so they measure adaptation to known tasks rather than label-free fidelity. On DS-R1-Qwen-7B, PPL and KL improve in every calibration replicate, but the Micro differences from AutoRound are +1.11, +1.12, and −2.47 points (Table XII in the Appendix), so the task result depends on the calibration bank. Changes at W4 are smaller: DS W4 improves PPL but not KL, and both Llama models show metric-dependent task trade-offs.

The fidelity mode lowers mean Avg. PPL and Avg. KL relative to QSelect-TA in all six settings (Table III). Because QSelect-TA is already below every single-family Avg. PPL mean, QSelect-KL has the lowest mean Avg. PPL among the configurations of Table II in every setting, and the lowest Avg. KL in five; in DS W4 its derived KL (0.0913) is comparable to that of homogeneous GPTQ (0.092). The task cost of removing answer supervision is concentrated at Llama-3.2 W3, where every paired replicate loses Micro accuracy and the means fall by 1.57 Micro and 1.03 Macro points; at W4 the changes are −0.16 Micro and +0.23 Macro. Even without labels, the

TABLE II
MAIN INTRINSIC AND DOWNSTREAM RESULTS (MEAN $\pm$ SAMPLE STANDARD DEVIATION OVER CALIBRATION-BANK REPLICATES 17, 23, AND 41)

Model	Bits	Method	Avg. PPL $\downarrow$	Max-corpus PPL $\downarrow$	Avg. KL $\downarrow$	Max-corpus KL $\downarrow$	Micro $\uparrow$	Macro $\uparrow$
DS-R1-Qwen-7B	BF16	Reference	26.68	44.32	0.000	0.000	49.64	41.19
	W3	AWQ	51.13 $\pm$ 1.98	88.63 $\pm$ 3.69	0.885 $\pm$ 0.016	1.058 $\pm$ 0.021	34.70 $\pm$ 0.90	29.31 $\pm$ 0.40
	W3	GPTQ	33.14 $\pm$ 1.06	56.31 $\pm$ 2.47	0.436 $\pm$ 0.005	0.498 $\pm$ 0.016	37.97 $\pm$ 0.99	31.22 $\pm$ 0.18
	W3	AutoRound	35.03 $\pm$ 0.69	58.94 $\pm$ 1.13	0.474 $\pm$ 0.006	0.562 $\pm$ 0.005	41.85 $\pm$ 0.14	33.25 $\pm$ 0.48
	W3	QSelect-TA	30.70 $\pm$ 0.26	51.15 $\pm$ 0.61	0.420 $\pm$ 0.004	0.498 $\pm$ 0.010	41.77 $\pm$ 1.95	34.02 $\pm$ 2.04
	W3	QSelect-KL	29.23 $\uparrow$	–	0.389 $\uparrow$	–	41.62 $\uparrow$	34.77 $\uparrow$
	W4	AWQ	27.68 $\pm$ 0.26	46.16 $\pm$ 0.33	0.132 $\pm$ 0.002	0.160 $\pm$ 0.004	46.94 $\pm$ 0.52	39.82 $\pm$ 0.47
	W4	GPTQ	28.09 $\pm$ 0.15	47.27 $\pm$ 0.43	0.092 $\pm$ 0.001	0.106 $\pm$ 0.003	48.60 $\pm$ 0.13	40.81 $\pm$ 0.34
	W4	AutoRound	29.29 $\pm$ 0.47	49.33 $\pm$ 0.84	0.124 $\pm$ 0.003	0.151 $\pm$ 0.004	48.81 $\pm$ 0.07	40.67 $\pm$ 0.21
	W4	QSelect-TA	27.33 $\pm$ 0.50	45.49 $\pm$ 0.88	0.1059 $\pm$ 0.0016	0.1250 $\pm$ 0.0019	48.83 $\pm$ 0.12	41.19 $\pm$ 0.20
	W4	QSelect-KL	26.87 $\uparrow$	–	0.0913 $\uparrow$	–	48.85 $\uparrow$	40.74 $\uparrow$
Llama-3.1-8B	BF16	Reference	7.01	10.05	0.000	0.000	59.68	45.44
	W3	AWQ	14.97 $\pm$ 0.54	22.08 $\pm$ 0.88	0.734 $\pm$ 0.029	0.898 $\pm$ 0.034	26.69 $\pm$ 0.53	19.64 $\pm$ 0.23
	W3	GPTQ	10.39 $\pm$ 0.13	14.71 $\pm$ 0.16	0.432 $\pm$ 0.014	0.550 $\pm$ 0.007	41.36 $\pm$ 2.13	30.23 $\pm$ 1.23
	W3	AutoRound	11.01 $\pm$ 0.17	15.96 $\pm$ 0.18	0.432 $\pm$ 0.013	0.521 $\pm$ 0.026	43.54 $\pm$ 0.89	31.24 $\pm$ 0.76
	W3	QSelect-TA	9.79 $\pm$ 0.05	14.03 $\pm$ 0.02	0.3486 $\pm$ 0.0075	0.4231 $\pm$ 0.0116	47.81 $\pm$ 0.53	36.24 $\pm$ 0.70
	W3	QSelect-KL	9.58 $\uparrow$	–	0.3336 $\uparrow$	–	46.91 $\uparrow$	35.04 $\uparrow$
	W4	AWQ	7.74 $\pm$ 0.05	11.08 $\pm$ 0.09	0.110 $\pm$ 0.006	0.127 $\pm$ 0.006	56.42 $\pm$ 0.35	42.98 $\pm$ 0.95
	W4	GPTQ	7.689 $\pm$ 0.005	11.188 $\pm$ 0.011	0.0937 $\pm$ 0.0011	0.1204 $\pm$ 0.0002	56.63 $\pm$ 0.19	43.56 $\pm$ 0.29
	W4	AutoRound	7.94 $\pm$ 0.05	11.54 $\pm$ 0.03	0.108 $\pm$ 0.005	0.131 $\pm$ 0.014	56.38 $\pm$ 0.27	44.01 $\pm$ 0.58
	W4	QSelect-TA	7.60 $\pm$ 0.02	10.93 $\pm$ 0.02	0.0835 $\pm$ 0.0011	0.1061 $\pm$ 0.0026	57.34 $\pm$ 0.14	43.73 $\pm$ 0.23
	W4	QSelect-KL	7.55 $\uparrow$	–	0.0810 $\uparrow$	–	57.04 $\uparrow$	43.33 $\uparrow$
Llama-3.2-3B	BF16	Reference	8.70	12.32	0.000	0.000	51.03	38.06
	W3	AWQ	24.44 $\pm$ 2.56	43.71 $\pm$ 5.63	0.962 $\pm$ 0.088	1.305 $\pm$ 0.123	25.82 $\pm$ 0.58	19.81 $\pm$ 0.26
	W3	GPTQ	16.66 $\pm$ 1.93	26.14 $\pm$ 4.84	0.668 $\pm$ 0.124	0.799 $\pm$ 0.166	33.71 $\pm$ 1.23	26.30 $\pm$ 0.98
	W3	AutoRound	15.13 $\pm$ 0.33	23.08 $\pm$ 0.60	0.527 $\pm$ 0.015	0.675 $\pm$ 0.019	36.13 $\pm$ 0.89	25.84 $\pm$ 0.88
	W3	QSelect-TA	12.92 $\pm$ 0.90	19.82 $\pm$ 2.02	0.408 $\pm$ 0.058	0.512 $\pm$ 0.095	40.60 $\pm$ 0.36	30.10 $\pm$ 0.67
	W3	QSelect-KL	12.21 $\pm$ 0.09	18.20 $\pm$ 0.06	0.367 $\pm$ 0.011	0.434 $\pm$ 0.008	39.02 $\pm$ 0.49	29.06 $\pm$ 0.42
	W4	AWQ	9.76 $\pm$ 0.04	14.15 $\pm$ 0.06	0.118 $\pm$ 0.005	0.152 $\pm$ 0.006	48.46 $\pm$ 0.20	37.28 $\pm$ 0.54
	W4	GPTQ	9.73 $\pm$ 0.16	13.71 $\pm$ 0.32	0.111 $\pm$ 0.010	0.143 $\pm$ 0.023	48.24 $\pm$ 0.15	35.12 $\pm$ 0.65
	W4	AutoRound	9.92 $\pm$ 0.03	13.99 $\pm$ 0.07	0.118 $\pm$ 0.003	0.155 $\pm$ 0.009	47.71 $\pm$ 0.50	35.10 $\pm$ 0.80
	W4	QSelect-TA	9.46 $\pm$ 0.05	13.77 $\pm$ 0.31	0.092 $\pm$ 0.006	0.125 $\pm$ 0.025	48.66 $\pm$ 0.80	36.07 $\pm$ 0.67
	W4	QSelect-KL	9.37 $\pm$ 0.10	13.56 $\pm$ 0.45	0.088 $\pm$ 0.007	0.117 $\pm$ 0.028	48.50 $\pm$ 0.69	36.30 $\pm$ 0.76

Reference rows are unquantized BF16 models evaluated once. Avg. PPL and Avg. KL are arithmetic means over five corpora within each replicate; Max-corpus PPL and KL are the corresponding maxima. Micro pools all 18,255 task examples; Macro averages the six task accuracies. Bold marks the best measured mean within each setting and does not indicate statistical significance. $\uparrow$ Derived mean: the QSelect-TA mean shifted by the paired change of QSelect-KL in Table III (relative for PPL and KL, in points for accuracy). Standard deviations and maxima are unavailable, the last digit can differ by one from a direct computation, and derived values are not bolded, although several derived PPL and KL means are lower than the bold ones. For Llama-3.2-3B, where both are available, derived and measured means agree to within 0.01.

fidelity mode exceeds the best single-family Micro accuracy at W3 on both Llama models, by 2.9 points on Llama-3.2-3B and 3.4 points (derived) on Llama-3.1-8B.

B. Why Score the Final Output?

Three comparisons isolate the choice of scoring signal. Held-out KL partly measures what a final-output selector optimizes, so we also report perplexity on the five held-out corpora and task accuracy, which the selectors in the first two comparisons do not optimize.

Representation similarity. Table IV replaces CKA with final-distribution KL inside the same sequential layer-wise scaffold, with the same candidate bank and token budget. The KL selector in this control is a guarded layer-wise variant that accepts a change only when it lowers KL by at least 0.1% (Eq. (8)); it is not the threshold-free fidelity mode. The final-output score lowers held-out Avg. KL in each of the twelve paired calibration-bank runs (Table X in the Appendix). PPL and task accuracy also improve on DS W3 and on both Llama settings, whereas DS W4 trades lower KL for slightly higher PPL and lower task means. Because selection and evaluation both use KL, this control supports a fidelity-oriented scoring

TABLE III
FIDELITY MODE AND A LOCAL RECONSTRUCTION CRITERION, BOTH RELATIVE TO THE MODULE-WISE QSELECT-TA MEAN. PPL AND KL ARE RELATIVE CHANGES OF THREE-REPLICATE MEANS; ACCURACY VALUES ARE PERCENTAGE POINTS. NEGATIVE PPL/KL AND POSITIVE ACCURACY FAVOR THE LISTED CONFIGURATION. LOCAL-MSE WAS RUN ON THE 7B/8B SETTINGS ONLY.

Setting	Configuration	Δ PPL	Δ KL	Δ Micro	Δ Macro
DS W3	QSelect-KL	-4.8%	-7.5%	-0.15	+0.75
	Local-MSE	+8.0%	+1.7%	-2.16	-1.20
DS W4	QSelect-KL	-1.7%	-13.8%	+0.02	-0.45
	Local-MSE	+2.9%	-13.5%	+0.04	+0.32
Llama-3.1 W3	QSelect-KL	-2.1%	-4.3%	-0.90	-1.20
	Local-MSE	-0.7%	+4.3%	-3.60	-3.30
Llama-3.1 W4	QSelect-KL	-0.6%	-3.0%	-0.30	-0.40
	Local-MSE	-0.9%	+5.8%	-0.80	-0.60
Llama-3.2 W3	QSelect-KL	-5.5%	-10.1%	-1.57	-1.03
Llama-3.2 W4	QSelect-KL	-1.0%	-4.5%	-0.16	+0.23

choice within the scaffold rather than a universal superiority claim. The final-output score is not much more expensive at equal granularity: layer-wise selection takes 19.4–21.6 minutes with KL and 16.9–18.1 minutes with CKA (Table IX).

Local reconstruction. The Local-MSE selector collects 8,192 stratified input tokens per module from the BF16

TABLE IV
CKA VERSUS GUARDED FINAL-DISTRIBUTION KL (0.1% MARGIN)
INSIDE THE SAME SEQUENTIAL LAYER-WISE SCAFFOLD. DELTAS ARE
FINAL-KL MINUS CALM-REIMPL.; PPL/KL ARE RELATIVE
PERCENTAGES AND ACCURACY VALUES ARE PERCENTAGE POINTS.

Setting	Δ PPL	Δ KL	Δ Micro	Δ Macro
DS W3	-1.42%	-9.96%	+0.74	+1.17
DS W4	+2.03%	-5.07%	-0.03	-0.46
Llama-3.1 W3	-6.26%	-19.14%	+5.45	+4.29
Llama-3.1 W4	-0.81%	-17.14%	+0.68	+0.88

TABLE V
WHOLE-MODEL FAMILY CHOICES MADE BY THE SAME KL-SHORTLIST
AND ANSWER-NLL RULE. COUNTS ARE OVER CALIBRATION-BANK
REPLICATES 17, 23, AND 41.

Setting	AWQ	GPTQ	AutoRound
DS W3	0	3	0
DS W4	0	3	0
Llama-3.1 W3	0	3	0
Llama-3.1 W4	0	3	0
Llama-3.2 W3	0	0	3
Llama-3.2 W4	0	2	1

reference and unconditionally chooses the family with the lowest mean-squared error between reference and candidate module outputs. It uses static reference activations, so it is not a contextual search and has no initialization or acceptance threshold. It takes about 0.8 minutes on the DS settings. Table III shows that it ends with higher Avg. KL than the fidelity mode in all four tested settings (marginally on DS W4), higher Avg. PPL in three, and lower Micro accuracy in three. Local-MSE is thus a useful quick reference, not a substitute for the output criterion.

Whole-model selection. Applying the same KL shortlist, answer NLL, and acceptance rule to whole homogeneous models selects GPTQ in all four 7B/8B settings, AutoRound in all Llama-3.2 W3 replicates, and GPTQ or AutoRound in two or one Llama-3.2 W4 replicates (Table V). In the five settings where all replicates agree, module-wise QSelect-TA lowers Avg. PPL by 1.22–14.62% and raises Micro and Macro by 0.23–6.45 and 0.17–6.01 points relative to the selected family. This control does not match the search capacity or path dependence of module-wise selection, but it shows that choosing one family with the same rule does not recover the module-wise results.

C. Ground Time Versus Fidelity

Fig. 2 places the evaluated selectors on one time axis for DS W3, and Table VI gives the granularity study. DS layer, component, and module scans take roughly 18, 35–40, and 210 minutes. Finer units are not uniformly better: modules give the lowest W3 PPL, components the highest DS task means, and at W4 layers give the lowest DS KL and the lowest Llama-3.1 PPL. At layer granularity, QSelect-TA gives higher Micro accuracy than CALM-reimpl. in all four settings and lower Avg. KL in three, with DS W4 essentially tied (0.0987 versus 0.0986); CALM-reimpl. gives slightly lower Avg. PPL on both DS settings. Three passes triple the candidate evaluations and raise selection time to 7.2–

9.5 hours, changing Avg. PPL by only -0.12% to -0.75% (Table XIII); one pass remains the default. On one A800 GPU, a one-pass module-wise QSelect-TA run for Llama-3.1-8B W3 takes 3.12 hours, and QSelect-KL on Llama-3.2-3B averages 83.9 minutes, both excluding candidate generation, which is not reported here. Exact final-output scoring is therefore far slower than a local proxy such as Local-MSE, which reuses static reference activations, but at equal granularity it adds only a few minutes over representation-similarity scoring, which also runs the assembled model (Section VI-B); most of the module-wise time comes from the number of decisions. The search itself is deliberately simple: a greedy coordinate descent with one forward pass per candidate. Its absolute cost stays modest because each evaluation is a forward pass of the model being prepared, and onboard memory keeps that model small: Llama-3.2-3B, the smallest model studied here, is still about twice the size of the 1.54-billion-parameter model flown on ASTREA [10]. The budget that makes compression necessary thus also keeps exact scoring affordable. The cost stays on the ground, because the exported model has the bit width and format of a single-family model apart from the AWQ input scales, but it recurs with every model update, because a new calibration bank changes the artifact (Section VII-C).

D. Robustness of the Ground Procedure

Traversal order has a smaller aggregate effect than initialization. Random and reverse scans change Avg. PPL by at most 0.10 in three settings; on DS W3 the increases are 0.77 and 0.53, and mean Macro changes by at most 0.67 points across all orders (Table XVI). Changing the restricted-pool Llama-3.1 W3 start from GPTQ to AWQ, by contrast, increases Avg. PPL by 1.18 and lowers Micro and Macro by 5.49 and 5.14 points (Table XV). Adding AutoRound to the pool has the clearest value at W3: across both models and both AWQ and GPTQ starts, it lowers Avg. PPL by 0.32–1.24 and raises Micro by 2.37–4.02 points, while W4 changes are smaller and mixed (Tables XIV and XV).

The selected models remain heterogeneous. On Llama-3.2-3B, the mean AWQ/GPTQ/AutoRound counts of QSelect-KL are 6.0/65.0/125.0 at W3 and 17.0/74.7/104.3 at W4. Across the three DS W3 calibration banks, pairwise assignment agreement is 73.5–76.0%, and 124 of 196 modules agree in all three. Selection therefore depends on the candidate bank as well as on the model, a point we return to in Section VII.

VII. DEPLOYMENT ANALYSIS FOR SPACE ROBOTS

This section interprets the measured results for onboard use. It adds an analytic storage model but no new experiments.

A. What Occupies the Packed Artifact

Table VII reports the packed exports of the calibration-replicate-17 assignments. Artifact reduction compares the serialized artifact directory with the BF16 reference directory. Packing changes held-out metrics only slightly relative to the fake-quantized model: $|\Delta\text{PPL}| \leq 0.277\%$ and $|\Delta\text{KL}| \leq 1.314\%$, with drift defined as $\Delta_z = 100 [z(M_{\text{nat}}) -$

TABLE VI

CALM COMPARISON AND QSELECT DECISION GRANULARITY. CALM-REIMPL. IS AN INDEPENDENTLY IMPLEMENTED LAYER-WISE CKA SELECTOR WITH THE MATCHED CANDIDATE POOL. THE THREE QSELECT-TA ROWS KEEP THE CANDIDATE BANK, AUTOROUND INITIALIZATION, SCORES, THRESHOLDS, FORWARD ORDER, ONE-PASS BUDGET, AND CALIBRATION-BANK REPLICATES FIXED WHILE CHANGING THE DECISION UNIT. VALUES ARE MEAN $\pm$ SAMPLE STANDARD DEVIATION; BOLD MARKS THE BEST DISPLAYED MEAN WITHIN EACH SETTING AND IS DESCRIPTIVE.

Setting	Selector	Unit	Avg. PPL $\downarrow$	Avg. KL $\downarrow$	Micro $\uparrow$	Macro $\uparrow$
DS-R1-Qwen-7B W3	CALM-reimpl.	Layer	33.34 $\pm$ 1.81	0.474 $\pm$ 0.015	38.49 $\pm$ 1.05	31.91 $\pm$ 0.75
	QSelect-TA	Layer	33.39 $\pm$ 0.65	0.423 $\pm$ 0.006	41.97 $\pm$ 0.66	34.27 $\pm$ 0.84
	QSelect-TA	Component	31.94 $\pm$ 0.95	0.425 $\pm$ 0.008	42.50$\pm$0.35	34.82$\pm$1.22
	QSelect-TA	Module	30.70$\pm$0.26	0.420$\pm$0.004	41.77 $\pm$ 1.95	34.02 $\pm$ 2.04
DS-R1-Qwen-7B W4	CALM-reimpl.	Layer	27.41 $\pm$ 0.23	0.0986 $\pm$ 0.0014	48.55 $\pm$ 0.41	41.07 $\pm$ 0.73
	QSelect-TA	Layer	27.77 $\pm$ 0.62	0.0987 $\pm$ 0.0013	48.81 $\pm$ 0.11	41.08 $\pm$ 0.18
	QSelect-TA	Component	27.84 $\pm$ 0.30	0.103 $\pm$ 0.001	49.09$\pm$0.09	41.27$\pm$0.17
	QSelect-TA	Module	27.33$\pm$0.50	0.106 $\pm$ 0.002	48.83 $\pm$ 0.12	41.19 $\pm$ 0.20
Llama-3.1-8B W3	CALM-reimpl.	Layer	10.56 $\pm$ 0.01	0.451 $\pm$ 0.008	39.28 $\pm$ 3.25	28.84 $\pm$ 2.18
	QSelect-TA	Layer	9.94 $\pm$ 0.08	0.364 $\pm$ 0.005	45.61 $\pm$ 0.50	33.78 $\pm$ 0.60
	QSelect-TA	Component	9.85 $\pm$ 0.13	0.353 $\pm$ 0.008	46.40 $\pm$ 1.20	34.39 $\pm$ 1.48
	QSelect-TA	Module	9.79$\pm$0.05	0.349$\pm$0.008	47.81$\pm$0.53	36.24$\pm$0.70
Llama-3.1-8B W4	CALM-reimpl.	Layer	7.60 $\pm$ 0.09	0.099 $\pm$ 0.007	56.30 $\pm$ 0.28	42.91 $\pm$ 0.47
	QSelect-TA	Layer	7.55$\pm$0.04	0.0842 $\pm$ 0.0007	56.63 $\pm$ 0.77	42.87 $\pm$ 0.91
	QSelect-TA	Component	7.62 $\pm$ 0.05	0.0837 $\pm$ 0.0017	57.25 $\pm$ 0.29	43.20 $\pm$ 0.63
	QSelect-TA	Module	7.60 $\pm$ 0.02	0.0835$\pm$0.0011	57.34$\pm$0.14	43.73$\pm$0.23

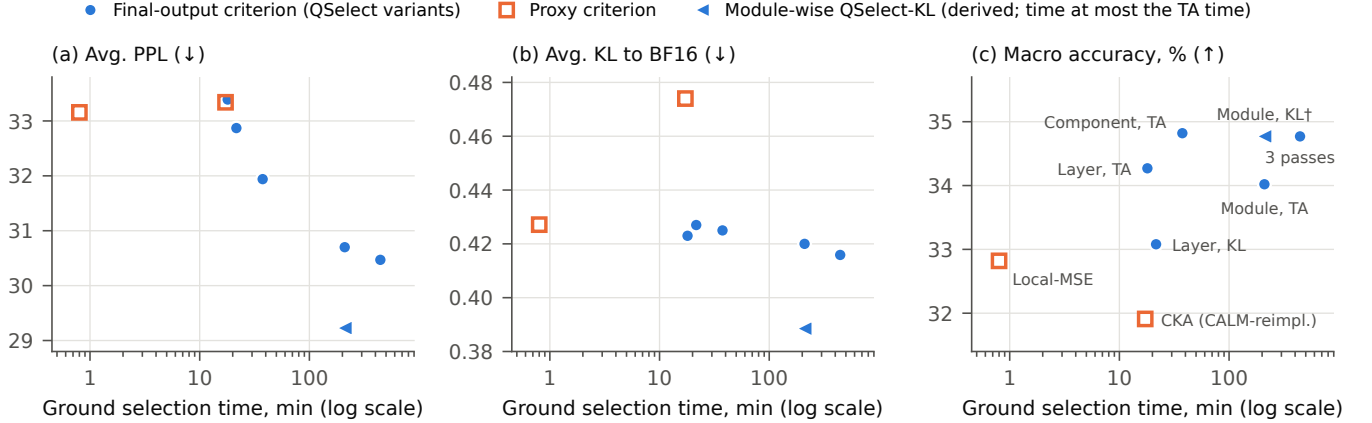


Fig. 2. Ground selection time versus fidelity and task accuracy on DS-R1-Qwen-7B W3; labels in (c) apply to all panels. Filled circles use the final output (layer-, component-, and module-wise QSelect-TA, the guarded layer-wise KL selector, and module-wise QSelect-TA with three passes); hollow squares are proxy criteria (Local-MSE and CALM-reimpl. with CKA). All values come from Tables II, III, VI, IX, and XIII. Local-MSE and three-pass metrics are derived from reported relative changes. Times for Local-MSE and the layer, component, and module scans are the approximate values reported for the DS settings (about 0.8, 18, 35–40, and 210 minutes); the bar on the component point spans 35–40 minutes. The triangle is module-wise QSelect-KL, derived from Table III; its time is not reported, so it is drawn at the module-wise QSelect-TA time, which bounds it from above.

$z(M_{fq})/z(M_{nat})$ for the packed model M_{nat} and the fake-quantized model M_{fq} .

To see where the remaining bytes are, we model the artifact from the public model configurations. Quantized linear layers store b -bit codes, a packed b -bit zero point and an FP16 scale for every group of 128 weights, and an int32 group index per input channel; embeddings, the output head, normalization weights, and biases stay in BF16. This model reproduces the four measured reductions to within 0.08 percentage points (Table VII). It also shows that the BF16 embedding and output layers, which the experiments never quantize, make up 43–46% of the 3-bit artifact and 37–39% of the 4-bit artifact (Fig. 3). Further reductions for an onboard memory budget must therefore address those layers, for example through embedding and head quantization or vocabulary restriction; we have not evaluated either.

TABLE VII

PACKED EXPORT OF THE CALIBRATION-REPLICATE-17 ASSIGNMENTS (QSELECT-TA). ARTIFACT AND PEAK-CUDA ENTRIES ARE REDUCTIONS RELATIVE TO THE BF16 REFERENCE; Δ PPL AND Δ KL ARE PACKED-VERSUS-FAKE-QUANTIZED DRIFT. $\dagger$ FROM THE ANALYTIC STORAGE MODEL (SECTION VII-A), NOT MEASURED.

Setting	Artifact reduction			Peak CUDA	Drift	
	Measured	Est. $\dagger$	BF16 share $\dagger$		Δ PPL	Δ KL
DS W3	68.74%	68.79%	45.9%	26.3%	-0.029%	-0.089%
DS W4	63.37%	63.40%	39.1%	24.2%	-0.028%	+0.024%
Llama-3.1 W3	69.70%	69.78%	43.3%	27.9%	+0.033%	-1.314%
Llama-3.1 W4	64.26%	64.30%	36.7%	25.9%	-0.277%	-0.232%

B. Runtime Memory

Peak CUDA memory falls by only 24.2–27.9% (Table VII), although the serialized artifact shrinks by roughly a factor of three. The absolute sizes also matter: by the analytic model,

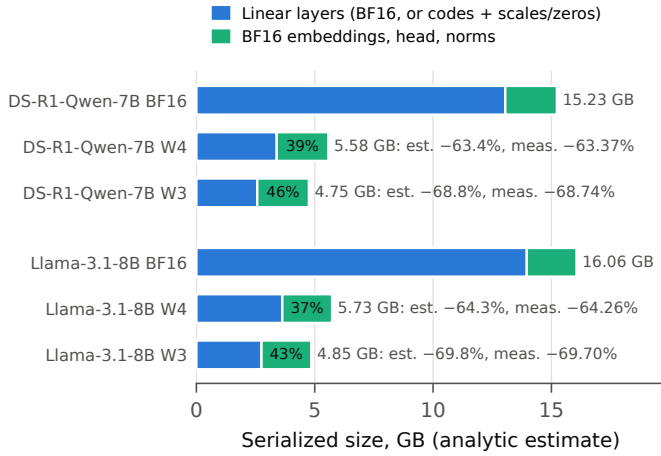


Fig. 3. Analytic composition of the serialized artifact. Green segments are the BF16 embedding, output head, and normalization layers; the percentages give their share of the quantized artifact. Each label compares the estimated reduction with the measured one in Table VII.

the 7B/8B artifacts still occupy about 4.8 GB at 3 bits and 5.6–5.7 GB at 4 bits. They are far beyond the 256 MB of a RAD750-class computer, call for accelerator-class payload computers, and are several times larger than the 1.54-billion-parameter model flown on ASTREA. The measurement uses batch one, sequence length 2,048, and no key–value cache, includes model loading, and runs the packed forward path used here, which dequantizes weights during execution and is not an optimized serving kernel. A flight deployment would need kernels that compute directly from packed weights and a memory plan for activations and the key–value cache, measured on flight-representative hardware. Latency matters as much as memory there: ASTREA’s 1.54-billion-parameter 4-bit model needed from about 40 s to over 8 min per response on a CPU-only flight system-on-chip [10]. We leave these measurements to future work.

C. Choosing an Operating Point

Bit width. At W4, composed models stay close to the reference: QSelect-TA is within 0.81, 2.34, and 2.37 Micro points of BF16 on DS-R1-Qwen-7B, Llama-3.1-8B, and Llama-3.2-3B. At W3 the gaps are 7.87, 11.87, and 10.43 points even after composition. For command interpretation, where the validated reference behavior is the specification, W4 is the realistic default; W3 fits only when memory is the binding constraint and task-level acceptance tests can quantify the loss. At W4, composition buys fidelity rather than accuracy: relative to the best single family, the fidelity mode lowers Avg. KL from 0.0937 to 0.0810[†] on Llama-3.1-8B and from 0.111 to 0.088 on Llama-3.2-3B and matches it on DS-R1-Qwen-7B (0.0913[†] versus 0.092), while its Micro accuracy stays within half a point of the best single-family value.

Mode. The fidelity mode is the natural default for R3: it needs no labels and, among the configurations of Table II, gives the lowest Avg. PPL in every setting and the lowest Avg. KL in five of six. The target-adapted mode is appropriate when the mission supplies labeled examples and its acceptance tests

cover the resulting departure from the reference. At Llama-3.2 W3, for instance, the fidelity mode lowers Avg. KL by 10.1% relative to QSelect-TA but loses 1.57 Micro points.

Updates. Because candidates are calibration-bank specific, re-running selection with a new bank changes the artifact substantially: across the DS W3 banks, only 124 of 196 module choices agree in all three. Even modules whose family is unchanged receive new weights, because each bank regenerates the candidates, so an update replaces the whole quantized artifact. A model update should therefore be treated as a new artifact with full acceptance testing rather than as a patch, and the configuration record (Section IV-F) makes the previous artifact recoverable for rollback.

D. What Flight Use Would Add

Three questions remain open. First, the reference behavior here is general-domain text and reasoning; a mission would validate the reference on its own command corpus, procedures, and skill library, and the fidelity score should then be computed on that text. Second, latency, energy, and thermal load must be measured on flight-representative processors. Third, packed formats change how a radiation-induced upset propagates: a corrupted group scale or zero point affects all 128 weights of its group, so these metadata may deserve stronger protection than the codes. A corrupted AWQ input scale affects a whole input channel of its module, so the exposure also depends on the assignment. None of these questions is answered by the present experiments.

VIII. LIMITATIONS

No space-domain evaluation. All results use general-domain benchmarks on a ground GPU. We do not evaluate space-robot commands, procedures, or skill libraries, flight hardware, latency, energy, or radiation effects.

Scope of the method. The evidence covers fixed-bit symmetric weight-only candidates with group size 128 for 3B–8B text-only models. Candidates inherit homogeneous calibration paths and are dense fake-quantized weights during assignment; family token budgets and compute differ. Global rotations, cross-layer transforms, activation quantization, incompatible runtime formats, vision–language models, and models larger than 8B are not demonstrated. Both Llama references are base models; a command interface would start from an instruction-tuned reference.

Metrics. The arithmetic five-corpus PPL average and raw maximum are descriptive rather than pooled metrics. The downstream protocol is custom, QSelect-TA uses labels from the evaluated task families and token-weights answer NLL differently from Micro and Macro accuracy, GPQA is the only proxy-unseen Llama task, and three calibration-bank replicates do not measure proxy or test-sampling uncertainty. Some QSelect-KL values in Table II are derived from paired changes rather than measured directly.

Search. Assignments depend on initialization, order, passes, the target-adapted margins, and the candidate bank. One pass is a cost choice, not convergence, and rejected NLL proposals do not trigger runner-up tests. Candidate generation and storage

are extra costs, and native export is executable but not an optimized serving kernel. The argument that exact scoring is affordable rests on the 3B–8B sizes studied; larger models or frequent re-preparation would call for prefix caching or a cheaper screening step before final-output scoring, for example a local Fisher-information approximation of the same KL, which we do not evaluate.

IX. CONCLUSION

Onboard language models for space robots need more than a small footprint: the compressed model must be a static artifact in a supported format whose behavior stays close to the reference validated on the ground. Treating compression as a ground-segment step, QSelect composes one low-bit model from compatible quantizer families and scores every substitution through the final next-token distribution. Its label-free fidelity mode has no tuned threshold and reaches lower mean held-out perplexity than every single-family model and the target-adapted mode in all six settings; final-output scoring beats representation-similarity and local-reconstruction criteria on held-out KL; and ground time can be traded against fidelity through decision granularity. The packed artifacts are 63–70% smaller than BF16, and nearly half of what remains at 3 bits (43–46%) is the unquantized embedding and output layers rather than the composed modules. Evaluating the fidelity criterion on mission text and measuring the composed models on flight-representative hardware are the next steps.

ACKNOWLEDGMENT

The authors thank the developers of the open models and quantization libraries used in this work.

REFERENCES

- [1] S. Chapin, K. Stewart, R. Leontie, and C. G. Henshaw, “Autonomous Planning In-space Assembly Reinforcement-learning free-flyer (APIARY) International Space Station Astrobee testing,” in *Proc. Int. Conf. Space Robot. (iSpaRo)*, Sendai, Japan, 2025, pp. 418–424, doi: 10.1109/iSpaRo66239.2025.11436341.
- [2] B. Ichter *et al.*, “Do as I can, not as I say: Grounding language in robotic affordances,” in *Proc. 6th Conf. Robot Learning, Proc. Mach. Learn. Res.*, vol. 205, pp. 287–318, 2023. [Online]. Available: <https://proceedings.mlr.press/v205/ichter23a.html>
- [3] B. Zitkovich *et al.*, “RT-2: Vision-language-action models transfer web knowledge to robotic control,” in *Proc. 7th Conf. Robot Learning, Proc. Mach. Learn. Res.*, vol. 229, pp. 2165–2183, 2023. [Online]. Available: <https://proceedings.mlr.press/v229/zitkovich23a.html>
- [4] M. J. Kim *et al.*, “OpenVLA: An open-source vision-language-action model,” in *Proc. 8th Conf. Robot Learning, Proc. Mach. Learn. Res.*, vol. 270, pp. 2679–2713, 2025. [Online]. Available: <https://proceedings.mlr.press/v270/kim25c.html>
- [5] A. Wu, H. Han, X. Luo, R. Wang, S. He, and X. Wan, “SpaceMind: A modular and self-evolving embodied vision-language agent framework for autonomous on-orbit servicing,” arXiv:2604.14399, 2026. [Online]. Available: <https://arxiv.org/abs/2604.14399>
- [6] G. Furano *et al.*, “Towards the use of artificial intelligence on the edge in space systems: Challenges and opportunities,” *IEEE Aerosp. Electron. Syst. Mag.*, vol. 35, no. 12, pp. 44–56, Dec. 2020, doi: 10.1109/MAES.2020.3008468.
- [7] NASA Science, “Perseverance rover components,” mission documentation. Accessed: Oct. 20, 2026. [Online]. Available: <https://science.nasa.gov/mission/mars-2020-perseverance/rover-components/>
- [8] J. Nash, V. Verma, E. Graser, H. Wang, G. Kubiak, M. Maimone, S. Myint, J. Biesiadecki, and P. Romano, “A software-based approach to radiation mitigation for planetary missions,” in *Proc. IEEE Aerosp. Conf.*, Big Sky, MT, USA, 2026, pp. 1–18, doi: 10.1109/AERO66936.2026.11519792.
- [9] Xinhua, “Space station crew gains AI assistant,” *China Daily*, Aug. 18, 2025. Accessed: Oct. 20, 2026. [Online]. Available: <https://global.china-daily.com.cn/a/202508/18/WS68a27b52a310b236346f217a.html>
- [10] A. D. Mousist, “ASTREA: Introducing agentic intelligence for orbital thermal autonomy,” arXiv:2509.13380v2, Oct. 11, 2025. [Online]. Available: <https://arxiv.org/abs/2509.13380v2>
- [11] European Cooperation for Space Standardization, “Space engineering: Machine learning handbook,” ECSS-E-HB-40-02A, Nov. 15, 2024. [Online]. Available: [https://ecss.nl/wp-content/uploads/2024/12/ECSS-E-HB-40-02A\(15November2024\).pdf](https://ecss.nl/wp-content/uploads/2024/12/ECSS-E-HB-40-02A(15November2024).pdf)
- [12] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, “GPTQ: Accurate post-training quantization for generative pre-trained transformers,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2023.
- [13] J. Lin, J. Tang, H. Tang, S. Yang, W.-M. Chen, W.-C. Wang, G. Xiao, X. Dang, C. Gan, and S. Han, “AWQ: Activation-aware weight quantization for LLM compression and acceleration,” in *Proc. Mach. Learn. Syst. (MLSys)*, 2024.
- [14] W. Cheng, W. Zhang, H. Shen, Y. Cai, X. He, K. Lv, and Y. Liu, “Optimize weight rounding via signed gradient descent for the quantization of LLMs,” in *Findings Assoc. Comput. Linguist.: EMNLP 2024*, 2024.
- [15] J. Zhang, Y. Zhang, and D. Chen, “CKA-guided modular quantization: Beyond bit-width to algorithmic diversity,” arXiv:2512.16282, 2025.
- [16] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, “Similarity of neural network representations revisited,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 3519–3529.
- [17] M. Foutter, D. Gammelli, J. Kruger, E. Foss, P. Bhoj, T. Guffanti, S. D’Amico, and M. Pavone, “Space-LLaVA: A vision-language model adapted to extraterrestrial applications,” in *Proc. IEEE Aerosp. Conf.*, Big Sky, MT, USA, 2025, pp. 1–23, doi: 10.1109/AERO63441.2025.11068757.
- [18] A. Carrasco, V. Rodriguez-Fernandez, and R. Linares, “Large language models as autonomous spacecraft operators in Kerbal Space Program,” *Adv. Space Res.*, vol. 76, no. 6, pp. 3480–3497, 2025, doi: 10.1016/j.asr.2025.06.034.
- [19] Frontgrade Gaisler, “GR740: Quad-core LEON4FT microprocessor,” product documentation. Accessed: Oct. 20, 2026. [Online]. Available: <https://www.gaisler.com/products/gr740>
- [20] NASA, “NASA awards next-generation spaceflight computing processor contract,” Release 22-087, Aug. 15, 2022. Accessed: Oct. 20, 2026. [Online]. Available: <https://www.nasa.gov/news-release/nasa-awards-next-generation-spaceflight-computing-processor-contract/>
- [21] G. Giuffrida *et al.*, “The Φ -sat-1 mission: The first on-board deep neural network demonstrator for satellite Earth observation,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Art. no. 5517414, pp. 1–14, 2022, doi: 10.1109/TGRS.2021.3125567.
- [22] AMD, “Radiation tolerant Versal AI Core series data sheet,” DS946, ver. 1.2, Feb. 19, 2025. [Online]. Available: <https://docs.amd.com/t/en-US/ds946-xqr-versal-ai-core/General-Description>
- [23] Thales Alenia Space, “IMAGIN-e payload unlocks the potential of space edge computing aboard the ISS,” Jan. 30, 2026. Accessed: Oct. 20, 2026. [Online]. Available: <https://www.thalesaleniaspace.com/en/news/imagin-e-payload-unlocks-potential-space-edge-computing-aboard-iss>
- [24] P. Singh, “‘Greetings, earthlings’: Nvidia-backed Starcloud trains first AI model in space as orbital data center race heats up,” *CNBC*, Dec. 10, 2025. Accessed: Oct. 20, 2026. [Online]. Available: <https://www.cnbc.com/2025/12/10/nvidia-backed-starcloud-trains-first-ai-model-in-space-orbital-data-centers.html>
- [25] European Space Agency, “OPS-SAT,” mission documentation. Accessed: Oct. 20, 2026. [Online]. Available: https://www.esa.int/Enabling_Support/Operations/OPS-SAT
- [26] M. Nagel, R. A. Amjad, M. van Baalen, C. Louizos, and T. Blankevoort, “Up or down? Adaptive rounding for post-training quantization,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020.
- [27] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han, “SmoothQuant: Accurate and efficient post-training quantization for large language models,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023.
- [28] W. Shao, M. Chen, Z. Zhang, P. Xu, L. Zhao, Z. Li, K. Zhang, P. Gao, Y. Qiao, and P. Luo, “OmniQuant: Omnidirectionally calibrated quantization for large language models,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.
- [29] S. Ashkboos, A. Mohtashami, M. L. Croci, B. Li, P. Cameron, M. Jaggi, D. Alistarh, T. Hoefler, and J. Hensman, “QuaRot: Outlier-free 4-bit inference in rotated LLMs,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2024.

- [30] H. Gu, L. Hu, S. Niu, and F. Liu, "FLRQ: Faster LLM quantization with flexible low-rank matrix sketching," in *Proc. AAAI Conf. Artif. Intell.*, vol. 40, no. 26, 2026, pp. 21369–21377.
- [31] Z. Zhao, F. Liu, J. Wang, C. Guan, Z. Wang, L. Jiang, and H. Guan, "SpecQuant: Spectral decomposition and adaptive truncation for ultra-low-bit LLMs quantization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 40, no. 34, 2026, pp. 28786–28794.
- [32] M. E. Akbulut, H. H. Y. Shalby, F. Pittorino, and M. Roveri, "InfoQ: Mixed-precision quantization via global information flow," in *Proc. AAAI Conf. Artif. Intell.*, vol. 40, no. 24, 2026, pp. 19598–19606.
- [33] K. Wang, Z. Liu, Y. Lin, J. Lin, and S. Han, "HAQ: Hardware-aware automated quantization with mixed precision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.
- [34] Z. Dong, Z. Yao, A. Gholami, M. W. Mahoney, and K. Keutzer, "HAWQ: Hessian aware quantization of neural networks with mixed-precision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019.
- [35] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv:1503.02531, 2015.
- [36] A. Edalati, A. Ghaffari, M. Ghazvini Nejad, L. Hou, B. Chen, M. Asgharian, and V. Partovi Nia, "OAC: Output-adaptive calibration for accurate post-training quantization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 16, 2025, pp. 16453–16461.
- [37] Y. Li, R. Gong, X. Tan, Y. Yang, P. Hu, Q. Zhang, F. Yu, W. Wang, and S. Gu, "BRECQ: Pushing the limit of post-training quantization by block reconstruction," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
- [38] Y. Bansal, P. Nakkiran, and B. Barak, "Revisiting model stitching to compare neural representations," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 225–236.
- [39] D. Guo *et al.*, "DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning," arXiv:2501.12948, 2025.
- [40] DeepSeek-AI, "DeepSeek-R1-Distill-Qwen-7B model card," 2025. [Online]. Available: <https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-7B>
- [41] A. Dubey *et al.*, "The Llama 3 herd of models," arXiv:2407.21783, 2024.
- [42] Meta AI, "Llama 3.2 model card," 2024. [Online]. Available: https://github.com/meta-llama/llama-models/blob/main/models/llama3_2/MODEL_CARD.md
- [43] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.
- [44] S. Merity, C. Xiong, J. Bradbury, and R. Socher, "Pointer sentinel mixture models," arXiv:1609.07843, 2016.
- [45] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman, "GPQA: A graduate-level Google-proof Q&A benchmark," arXiv:2311.12022, 2023.
- [46] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? Try ARC, the AI2 reasoning challenge," arXiv:1803.05457, 2018.
- [47] K. Cobbe *et al.*, "Training verifiers to solve math word problems," arXiv:2110.14168, 2021.
- [48] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi, "HellaSwag: Can a machine really finish your sentence?" in *Proc. 57th Annu. Meeting Assoc. Comput. Linguist. (ACL)*, 2019.
- [49] H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe, "Let's verify step by step," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.
- [50] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt, "Measuring massive multitask language understanding," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
- [51] Y. Wang *et al.*, "MMLU-Pro: A more robust and challenging multi-task language understanding benchmark," arXiv:2406.01574, 2024.

Algorithm 2 One-pass QSelect-TA assignment

Require: Candidate bank $\{\mathcal{B}_{l,s}^{(r)}\}$, units $\mathcal{G}$, feasible initial assignment a , fixed order, and Eqs. (7)–(9).

- 1: $K_{\text{cur}} \leftarrow K_s(a)$; $T_{\text{cur}} \leftarrow T(a)$; $T_{\text{valid}} \leftarrow \text{true}$.
- 2: **for** decision unit G in the fixed order **do**
- 3: Let r_{cur} be the family currently assigned to G .
- 4: **for** $r \in \mathcal{C}_G$ **do**
- 5: $K_G(r) \leftarrow K_s(a[G \leftarrow r])$.
- 6: **end for**
- 7: Form $\mathcal{S}_G$ with Eq. (7).
- 8: **if** $|\mathcal{S}_G| = 1$ **then**
- 9: Let r^* be its sole family; mark $T_G(r^*)$ unevaluated.
- 10: **else**
- 11: **if** $\neg T_{\text{valid}}$ **then**
- 12: $T_{\text{cur}} \leftarrow T(a)$; $T_{\text{valid}} \leftarrow \text{true}$.
- 13: **end if**
- 14: **for** $r \in \mathcal{S}_G$ **do**
- 15: $T_G(r) \leftarrow T(a[G \leftarrow r])$, reusing T_{cur} for r_{cur} .
- 16: **end for**
- 17: Let r^* minimize $T_G(r)$ under fixed tie breaking.
- 18: **end if**
- 19: **if** $r^* \neq r_{\text{cur}}$ and Eq. (8) holds **then**
- 20: $a \leftarrow a[G \leftarrow r^*]$; $K_{\text{cur}} \leftarrow K_G(r^*)$.
- 21: **if** $T_G(r^*)$ was evaluated **then**
- 22: $T_{\text{cur}} \leftarrow T_G(r^*)$; $T_{\text{valid}} \leftarrow \text{true}$.
- 23: **else**
- 24: $T_{\text{valid}} \leftarrow \text{false}$.
- 25: **end if**
- 26: **else if** $r^* \neq r_{\text{cur}}$, $T_G(r^*)$ was evaluated, and Eq. (9) holds **then**
- 27: $a \leftarrow a[G \leftarrow r^*]$; $K_{\text{cur}} \leftarrow K_G(r^*)$; $T_{\text{cur}} \leftarrow T_G(r^*)$.
- 28: $T_{\text{valid}} \leftarrow \text{true}$.
- 29: **end if**
- 30: **end for**
- 31: **return** a and $F_s(a)$.

APPENDIX A

TARGET-ADAPTED ALGORITHM

Algorithm 2 gives the one-pass QSelect-TA procedure. The implementation caches the current KL and NLL scores. If a proposal is accepted without evaluating NLL, the NLL cache is invalidated and recomputed lazily before its next use. If the minimum-NLL proposal fails both acceptance branches, no runner-up is tried.

APPENDIX B

PROTOCOL DETAILS

The calibration-bank replicates are 17, 23, and 41. Each replicate determines the candidate-generation C4 manifest and the separate C4 sample used for KL scoring; the labeled proxy and held-out evaluation manifests are fixed. Candidate generation uses 128 sequences of length 2,048. KL scoring uses 64 sequences of length 1,024. The five-corpus suite uses 128 sequences per corpus at length 2,048 and evaluation seed 1009. Manifest hashes and candidate checkpoint hashes are recorded in the run metadata.

The proxy builder takes fixed dataset rows without per-run resampling. GSM8K uses the first 128 training rows and the string following “####”; MMLU uses the first 256 development rows and answer letters; ARC-Challenge and HellaSwag use the first 128 and 256 training rows, respectively, with the original option order; MATH cycles through seven training subjects until 128 parseable final answers are collected. Multiple-choice proxy prompts end in `Answer:`, and the target continuation is a leading space followed by the gold letter. GSM8K uses `Question: ... Answer:` and MATH uses `Problem: ... Final answer:`, each followed only by the canonical final-answer string. No reasoning trace is included.

Downstream evaluation is zero-shot with raw text prompts. Multiple-choice prompts list lettered options and end in `Answer:`; the score is the average continuation log probability of a leading-space answer letter. HellaSwag is reformatted as a multiple-choice question with activity and context fields. GSM8K uses `Question: ... Answer: Let’s think step by step.`, and MATH-500 uses `Problem: ... Solution: Let’s solve it step by step.` The final answer is. Generation is greedy with at most 256 new tokens and stops at the tokenizer EOS token. GSM8K uses the final numeric string; MATH-500 first extracts the last boxed expression and then applies whitespace, comma, dollar-sign, and case normalization with a last-number fallback. This custom protocol supports matched internal comparison but should not be read as a standard lm-evaluation-harness score.

The local Code corpus is built from source files in the evaluation repository (Python, C/C++, CUDA, Rust, Go, Java, JavaScript, and TypeScript), excluding common dependency and cache directories. The Math builder uses MMLU-Pro question text [51] when available, and the Wiki builder uses WikiText-103 training text. The frozen files support internally repeatable comparison, but the resolved upstream revisions, license metadata, file counts, language composition, and complete deduplication provenance were not preserved, so these corpora should be treated as local robustness sets rather than standard public benchmarks.

Software versions are PyTorch 2.10.0+cu128, Transformers 5.8.0, Datasets 4.0.0, Accelerate 1.13.0, GPTQModel 7.1.0, AutoRound 0.13.0, and Safetensors 0.7.0. Each reported job uses one NVIDIA A800-SXM4-80GB GPU. The launch metadata records checkpoint names but does not pin model and tokenizer commit revisions, which limits reproducibility.

APPENDIX C

ADDITIONAL RESULTS

A. Per-Task Results

Table VIII lists selected per-task accuracies. The selected rows expose task-level trade-offs but do not compare all candidate families; for Llama, the proxy-unseen group contains only GPQA.

B. Matched Layer-Wise Scoring

Table IX reports the full-budget CALM-reimpl. and final-distribution KL variants discussed in Section VI-B; Table X

gives the paired Avg.-KL differences. Selection time excludes candidate generation. The guarded layer-wise KL selector accepts a change only under the 0.1% margin of Eq. (8).

C. Objective Ablations

Module-wise QSelect-KL selects the minimum-KL candidate and accepts any strict KL decrease. NLL-only ranks all candidates by target-answer NLL and accepts a replacement when NLL improves by at least 0.3%, without a KL condition. NLL-first + KL guard ranks all candidates by NLL, removes the KL shortlist, and applies the default acceptance conditions in Eqs. (8)–(9). All variants retain the same replicate-specific candidate bank, initialization, order, and proxy data (Table XI). NLL-only improves both task aggregates only on DS W3 and degrades PPL and KL in all four evaluated 7B/8B settings, so answer supervision shifts the operating point rather than uniformly improving it.

D. Paired DS-R1-Qwen-7B W3 Analysis

In each calibration-bank replicate, the task baseline with the highest Micro and Macro value is AutoRound. For PPL and KL, the baseline is selected separately within each replicate and metric. Negative intrinsic deltas and positive task deltas favor QSelect-TA (Table XII). The replicate-41 Micro decrease is driven primarily by a 2.91-point MMLU decrease relative to AutoRound; because MMLU contributes 14,042 of 18,255 examples, this term accounts for approximately 2.23 points of the total Micro change. Pairwise assignment agreement is 73.5–76.0%, and 124/196 modules agree across all three replicates. Among the 22 modules for which replicates 17 and 23 agree but replicate 41 differs, 12 change from AutoRound to GPTQ. The observed association does not determine which replacements cause the task decrease.

E. Search Budget and Proxy Control

Table XIII reports differences from the corresponding one-pass main result. Every three-pass row uses exactly three full scans. Shared-core uses the DS proxy subset for Llama-3.1-8B.

F. Initialization Sensitivity

Table XIV varies only the homogeneous starting assignment. Different starts retain different family majorities, and no start is uniformly best across fidelity and task metrics.

G. Candidate Pool $\times$ Initialization

Together with the full-pool AWQ and GPTQ starts in Table XIV, Table XV completes a 2×2 comparison of candidate pool and initialization. The restricted pool contains AWQ and GPTQ only; all other protocol choices remain fixed. The large Llama-3.1 W3 difference between AWQ and GPTQ starts shows that a restricted-pool result cannot be interpreted as a pure source-removal effect.

H. Traversal-Order Sensitivity

Table XVI compares the fixed forward protocol with one fixed random permutation (order seed 2027) and the exact reverse of the full module list. Each row is paired with the same calibration-bank replicate of the forward main result. Deltas are variant minus forward, so negative PPL/KL and positive accuracy favor the alternative order.

TABLE VIII

SELECTED PER-TASK DOWNSTREAM ACCURACY (%). THE 7B/8B ROWS SHOW THE AUTOROUND INITIALIZATION BASELINE AND QSELECT-TA; THE 3B ROWS SHOW ALL THREE HOMOGENEOUS CANDIDATES PLUS QSELECT-KL AND QSELECT-TA. THIS TABLE IS NOT EXHAUSTIVE FOR THE 7B/8B SETTINGS. MICRO POOLS ALL EXAMPLES, WHEREAS MACRO AVERAGES TASKS.

Model	Bits	Method	GPQA	ARC-C	GSM8K	HellaSwag	MATH500	MMLU	Micro	Macro
DS-R1-Qwen-7B	BF16	Reference	31.31	68.94	23.81	50.59	20.80	51.67	49.64	41.19
	W3	AutoRound	23.74±2.20	59.41±0.62	21.13±2.27	34.34±0.46	16.87±2.42	44.03±0.37	41.85±0.14	33.25±0.48
	W3	QSelect-TA	23.23±1.34	58.62±3.08	22.52±2.82	36.20±4.27	19.93±0.90	43.62±1.85	41.77±1.95	34.02±2.04
	W4	AutoRound	32.66±2.33	68.20±0.18	21.15±0.79	48.89±1.13	22.13±1.94	50.97±0.13	48.81±0.07	40.67±0.21
	W4	QSelect-TA	33.33±1.01	67.38±0.88	22.97±1.32	50.36±1.16	22.33±2.25	50.76±0.19	48.83±0.12	41.19±0.20
Llama-3.1-8B	BF16	Reference	28.28	79.01	25.25	60.45	16.40	63.23	59.68	45.44
	W3	AutoRound	27.27±2.81	57.20±1.95	4.65±1.28	37.11±0.93	13.40±2.23	47.82±1.05	43.54±0.89	31.24±0.76
	W3	QSelect-TA	30.30±1.34	63.45±1.07	5.41±0.77	50.55±1.30	16.07±2.55	51.66±0.54	47.81±0.53	36.24±0.70
	W4	AutoRound	31.82±1.01	73.89±0.43	27.37±2.33	55.44±0.54	16.07±1.70	59.49±0.20	56.38±0.27	44.01±0.58
	W4	QSelect-TA	27.95±2.87	76.48±0.94	24.18±4.59	59.02±1.93	14.07±1.51	60.69±0.31	57.34±0.14	43.73±0.23
Llama-3.2-3B	BF16	Reference	25.76	67.32	8.79	58.30	13.40	54.81	51.03	38.06
	W3	AWQ	26.26±3.54	27.73±1.88	1.34±0.16	26.95±0.83	8.07±1.86	28.51±0.70	25.82±0.58	19.81±0.26
	W3	GPTQ	30.64±1.54	41.64±1.74	3.26±0.46	33.46±1.95	12.07±0.99	36.74±1.32	33.71±1.23	26.30±0.98
	W3	AutoRound	25.59±2.10	46.22±2.17	2.35±0.08	28.78±1.04	12.13±0.92	39.99±0.95	36.13±0.89	25.84±0.88
	W3	QSelect-TA	26.94±1.91	51.51±1.02	3.51±0.42	41.89±0.96	12.47±1.50	44.27±0.57	40.60±0.36	30.10±0.67
	W3	QSelect-KL	26.09±0.29	50.88±0.79	3.59±0.70	39.75±2.94	11.60±1.25	42.47±0.63	39.02±0.49	29.06±0.42
	W4	AWQ	30.13±1.05	64.82±0.44	10.06±1.40	54.10±1.18	12.73±1.42	51.82±0.09	48.46±0.20	37.28±0.54
	W4	GPTQ	25.08±1.77	63.88±0.56	9.33±2.36	49.28±1.08	11.00±1.06	52.16±0.21	48.24±0.15	35.12±0.65
	W4	AutoRound	26.26±1.75	63.74±0.84	7.53±2.09	48.44±2.13	13.00±1.40	51.63±0.49	47.71±0.50	35.10±0.80
	W4	QSelect-TA	25.93±1.77	64.25±0.17	8.52±1.38	51.50±0.62	13.73±1.21	52.49±0.88	48.66±0.80	36.07±0.67
	W4	QSelect-KL	27.27±1.34	64.79±0.61	8.49±0.93	52.08±2.83	12.93±0.12	52.20±0.68	48.50±0.69	36.30±0.76

TABLE IX

MATCHED LAYER-WISE SCORING COMPARISON. BOTH SELECTORS USE THE SAME REPLICATE-SPECIFIC CANDIDATE BANK, SEQUENTIAL LAYER-WISE SCAFFOLD, 64×1024 -TOKEN BUDGET, AND THREE CALIBRATION-BANK REPLICATES. VALUES ARE MEAN $\pm$ SAMPLE STANDARD DEVIATION.

Setting	Score	Avg. PPL↓	Max-corpus PPL↓	Avg. KL↓	Max-corpus KL↓	Micro↑	Macro↑	Sel. min
DS W3	CALM-reimpl. (CKA)	33.34±1.81	55.70±3.97	0.474±0.015	0.538±0.028	38.49±1.05	31.91±0.75	17.2
	Final-distribution KL	32.87±0.80	55.12±2.04	0.427±0.006	0.476±0.013	39.23±1.24	33.08±1.38	21.6
DS W4	CALM-reimpl. (CKA)	27.41±0.23	45.40±0.54	0.099±0.001	0.113±0.004	48.55±0.41	41.07±0.73	16.9
	Final-distribution KL	27.97±0.17	46.67±0.53	0.094±0.001	0.107±0.002	48.52±0.30	40.61±0.48	20.4
Llama-3.1 W3	CALM-reimpl. (CKA)	10.56±0.01	14.95±0.15	0.451±0.008	0.569±0.010	39.28±3.25	28.84±2.18	18.1
	Final-distribution KL	9.89±0.11	14.07±0.26	0.365±0.002	0.472±0.019	44.73±1.70	33.13±1.25	19.4
Llama-3.1 W4	CALM-reimpl. (CKA)	7.60±0.09	10.94±0.14	0.099±0.007	0.138±0.011	56.30±0.28	42.91±0.47	16.9
	Final-distribution KL	7.54±0.03	10.81±0.03	0.082±0.001	0.113±0.001	56.98±0.51	43.79±0.45	21.2

TABLE X

PAIRED AVG.-KL DIFFERENCES FOR THE MATCHED LAYER-WISE SCORING CONTROL. VALUES ARE FINAL-DISTRIBUTION KL MINUS CALM-REIMPL.; NEGATIVE VALUES FAVOR FINAL-DISTRIBUTION KL.

Setting	Rep. 17	Rep. 23	Rep. 41
DS W3	-0.0353	-0.0543	-0.0522
DS W4	-0.0058	-0.0029	-0.0062
Llama-3.1 W3	-0.0900	-0.0788	-0.0900
Llama-3.1 W4	-0.0158	-0.0231	-0.0118

TABLE XI

OBJECTIVE VARIANTS RELATIVE TO THE DEFAULT QSELECT-TA AGGREGATE MEAN. PPL AND KL ARE COMPUTED AS $100(\bar{z}_{\text{variant}} - \bar{z}_{\text{TA}})/\bar{z}_{\text{TA}}$ FROM THE THREE-REPLICATE MEANS; ACCURACY VALUES ARE DIFFERENCES IN PERCENTAGE POINTS. NEGATIVE PPL/KL AND POSITIVE ACCURACY FAVOR THE VARIANT.

Setting	Configuration	Δ PPL	Δ KL	Δ Micro	Δ Macro
DS W3	QSelect-KL	-4.8%	-7.5%	-0.15	+0.75
	NLL-only	+18.6%	+27.3%	+1.23	+1.30
DS W4	NLL-first + KL guard	+8.0%	+9.0%	+0.32	+0.10
	QSelect-KL	-1.7%	-13.8%	+0.02	-0.45
Llama-3.1 W3	NLL-only	+6.8%	+22.9%	+0.23	-0.18
	NLL-first + KL guard	+6.0%	+14.2%	-0.05	-0.61
Llama-3.1 W4	QSelect-KL	-2.1%	-4.3%	-0.90	-1.20
	NLL-only	+3.6%	+10.9%	-1.20	-1.30
Llama-3.1 W4	NLL-first + KL guard	+3.9%	+10.7%	-1.00	-0.80
	QSelect-KL	-0.6%	-3.0%	-0.30	-0.40
Llama-3.2 W3	NLL-only	+1.8%	+20.2%	-0.80	+0.30
	NLL-first + KL guard	+1.8%	+20.2%	-0.80	+0.30
Llama-3.2 W3	QSelect-KL	-5.5%	-10.1%	-1.57	-1.03
Llama-3.2 W4	QSelect-KL	-1.0%	-4.5%	-0.16	+0.23

TABLE XII
PAIRED DS-R1-QWEN-7B W3 DIFFERENCES UNDER REPLICATE-SPECIFIC CANDIDATE BANKS. FAMILY COUNTS ARE AWQ/AUTOROUND/GPTQ
MODULE ASSIGNMENTS OUT OF 196.

Rep.	Best-baseline PPL	Δ PPL	Δ PPL (%)	Best-baseline KL	Δ KL	Δ Micro	Δ Macro	A/AR/G
17	32.9196 (GPTQ)	-2.0953	-6.36	0.4307 (GPTQ)	-0.0072	+1.11	+2.11	12/146/38
23	32.2026 (GPTQ)	-1.3345	-4.14	0.4349 (GPTQ)	-0.0193	+1.12	+1.29	13/148/35
41	34.2321 (AutoRound)	-3.8313	-11.19	0.4410 (GPTQ)	-0.0197	-2.47	-1.10	13/146/37

TABLE XIII
SEARCH-BUDGET AND PROXY CONTROLS RELATIVE TO THE MAIN
RESULT. PPL/KL ARE RELATIVE PERCENTAGES; ACCURACY VALUES
ARE PERCENTAGE POINTS.

Setting	Configuration	Δ PPL	Δ KL	Δ Micro	Δ Macro	Sel. min
DS W3	Three-pass	-0.75%	-0.98%	+0.62	+0.75	443.6
DS W4	Three-pass	-0.12%	-2.17%	+0.09	-0.18	432.1
Llama-3.1 W3	Three-pass	-0.32%	-0.55%	+0.34	+0.52	567.1
Llama-3.1 W4	Three-pass	-0.13%	-0.60%	+0.25	+0.49	545.8
Llama-3.1 W3	Shared-core proxy	-0.10%	-0.57%	-0.25	-0.51	138.5
Llama-3.1 W4	Shared-core proxy	+0.04%	+0.60%	-0.29	-0.37	155.6

TABLE XIV
INITIALIZATION SENSITIVITY UNDER THE DEFAULT MODULE-WISE
ONE-PASS RULE. VALUES ARE MEAN $\pm$ SAMPLE STANDARD DEVIATION
OVER THREE CALIBRATION-BANK REPLICATES.

Setting	Start	Avg. PPL $\downarrow$	Avg. KL $\downarrow$	Micro $\uparrow$	Macro $\uparrow$
DS W3	AutoRound	30.70 $\pm$ 0.26	0.420 $\pm$ 0.004	41.77 $\pm$ 1.95	34.02 $\pm$ 2.04
	AWQ	33.58 $\pm$ 1.35	0.515 $\pm$ 0.017	41.70 $\pm$ 0.68	34.48 $\pm$ 0.61
	GPTQ	31.38 $\pm$ 0.32	0.409 $\pm$ 0.002	41.49 $\pm$ 0.52	34.61 $\pm$ 0.80
DS W4	AutoRound	27.33 $\pm$ 0.50	0.106 $\pm$ 0.002	48.83 $\pm$ 0.12	41.19 $\pm$ 0.20
	AWQ	27.18 $\pm$ 0.23	0.113 $\pm$ 0.004	48.99 $\pm$ 0.18	41.13 $\pm$ 0.43
	GPTQ	27.69 $\pm$ 0.13	0.092 $\pm$ 0.002	48.87 $\pm$ 0.14	40.81 $\pm$ 0.90
Llama-3.1 W3	AutoRound	9.79 $\pm$ 0.05	0.349 $\pm$ 0.008	47.81 $\pm$ 0.53	36.24 $\pm$ 0.70
	AWQ	11.02 $\pm$ 0.27	0.475 $\pm$ 0.014	42.93 $\pm$ 1.10	32.66 $\pm$ 1.58
	GPTQ	9.80 $\pm$ 0.05	0.371 $\pm$ 0.009	47.16 $\pm$ 0.98	35.36 $\pm$ 0.62
Llama-3.1 W4	AutoRound	7.60 $\pm$ 0.02	0.084 $\pm$ 0.001	57.34 $\pm$ 0.14	43.73 $\pm$ 0.23
	AWQ	7.61 $\pm$ 0.02	0.096 $\pm$ 0.001	57.50 $\pm$ 0.35	44.81 $\pm$ 1.04
	GPTQ	7.54 $\pm$ 0.03	0.087 $\pm$ 0.001	56.70 $\pm$ 0.48	42.82 $\pm$ 0.95

TABLE XV
AWQ/GPTQ RESTRICTED-POOL SELECTOR UNDER TWO INITIALIZATIONS. VALUES ARE MEAN $\pm$ SAMPLE STANDARD DEVIATION OVER THE SAME
THREE CALIBRATION-BANK REPLICATES AS TABLE II. TOGETHER WITH THE FULL-POOL AWQ/GPTQ ROWS IN TABLE XIV, THESE RESULTS FORM THE
POOL-BY-INITIALIZATION CONTROL.

Setting	Start	Avg. PPL $\downarrow$	Max-corpus PPL $\downarrow$	Avg. KL $\downarrow$	Max-corpus KL $\downarrow$	Micro $\uparrow$	Macro $\uparrow$
DS W3	AWQ	34.83 $\pm$ 0.99	58.53 $\pm$ 2.20	0.563 $\pm$ 0.012	0.654 $\pm$ 0.013	39.33 $\pm$ 1.50	33.93 $\pm$ 0.67
DS W3	GPTQ	31.70 $\pm$ 0.63	53.16 $\pm$ 1.67	0.432 $\pm$ 0.002	0.485 $\pm$ 0.007	38.87 $\pm$ 0.23	32.85 $\pm$ 1.30
DS W4	AWQ	27.43 $\pm$ 0.36	45.78 $\pm$ 0.75	0.111 $\pm$ 0.003	0.131 $\pm$ 0.006	48.77 $\pm$ 0.17	40.74 $\pm$ 0.24
DS W4	GPTQ	27.72 $\pm$ 0.21	46.29 $\pm$ 0.46	0.0928 $\pm$ 0.0010	0.1059 $\pm$ 0.0015	48.63 $\pm$ 0.18	40.75 $\pm$ 0.71
Llama-3.1 W3	AWQ	11.40 $\pm$ 0.23	16.25 $\pm$ 0.56	0.519 $\pm$ 0.012	0.669 $\pm$ 0.024	38.91 $\pm$ 1.43	29.12 $\pm$ 1.73
Llama-3.1 W3	GPTQ	10.22 $\pm$ 0.08	14.22 $\pm$ 0.14	0.419 $\pm$ 0.007	0.590 $\pm$ 0.009	44.39 $\pm$ 0.43	34.26 $\pm$ 0.90
Llama-3.1 W4	AWQ	7.621 $\pm$ 0.018	11.017 $\pm$ 0.033	0.0983 $\pm$ 0.0024	0.1282 $\pm$ 0.0028	57.21 $\pm$ 0.19	44.67 $\pm$ 1.05
Llama-3.1 W4	GPTQ	7.540 $\pm$ 0.007	10.844 $\pm$ 0.028	0.0882 $\pm$ 0.0004	0.1299 $\pm$ 0.0015	56.72 $\pm$ 0.47	43.68 $\pm$ 0.64

TABLE XVI
TRAVERSAL-ORDER SENSITIVITY RELATIVE TO THE FORWARD MODULE-WISE ONE-PASS PROTOCOL. VALUES ARE PAIRED MEAN $\pm$ SAMPLE STANDARD
DEVIATION ACROSS CALIBRATION-BANK REPLICATES 17, 23, AND 41.

Setting	Order	Δ Avg. PPL	Δ Max PPL	Δ Avg. KL	Δ Max KL	Δ Micro	Δ Macro
DS W3	Fixed random	+0.768 $\pm$ 0.259	+1.444 $\pm$ 0.950	+0.0183 $\pm$ 0.0034	+0.0238 $\pm$ 0.0040	+0.66 $\pm$ 2.12	+0.36 $\pm$ 1.80
DS W3	Full reverse	+0.526 $\pm$ 0.327	+0.993 $\pm$ 1.139	+0.0041 $\pm$ 0.0058	+0.0038 $\pm$ 0.0026	+0.13 $\pm$ 1.67	+0.67 $\pm$ 1.10
DS W4	Fixed random	+0.095 $\pm$ 0.366	+0.265 $\pm$ 0.681	-0.0004 $\pm$ 0.0017	+0.0008 $\pm$ 0.0026	-0.00 $\pm$ 0.22	-0.07 $\pm$ 0.41
DS W4	Full reverse	+0.055 $\pm$ 0.372	+0.122 $\pm$ 0.708	+0.0010 $\pm$ 0.0027	+0.0022 $\pm$ 0.0030	-0.02 $\pm$ 0.31	-0.25 $\pm$ 0.31
Llama-3.1 W3	Fixed random	-0.017 $\pm$ 0.024	+0.027 $\pm$ 0.018	-0.0044 $\pm$ 0.0049	-0.0040 $\pm$ 0.0094	+0.01 $\pm$ 0.69	+0.36 $\pm$ 0.79
Llama-3.1 W3	Full reverse	+0.058 $\pm$ 0.163	+0.025 $\pm$ 0.186	+0.0000 $\pm$ 0.0106	+0.0074 $\pm$ 0.0163	-0.54 $\pm$ 1.85	-0.33 $\pm$ 1.63
Llama-3.1 W4	Fixed random	-0.017 $\pm$ 0.038	-0.012 $\pm$ 0.041	-0.0000 $\pm$ 0.0005	-0.0006 $\pm$ 0.0022	-0.18 $\pm$ 0.53	-0.26 $\pm$ 1.46
Llama-3.1 W4	Full reverse	-0.005 $\pm$ 0.020	-0.020 $\pm$ 0.023	-0.0009 $\pm$ 0.0009	-0.0011 $\pm$ 0.0014	-0.07 $\pm$ 0.22	-0.56 $\pm$ 0.40