Quantization-aware training: a tradeoff between training and fine-tuning for domain-specific language models

1st Xavier Pillet
LS2N - Nantes Université
Nantes, France
xavier.pillet@ls2n.fr

2nd Cédric Gernigon
LS2N - Nantes Université
Nantes, France
cedric.gernigon@ls2n.fr

3rd Anastasia Volkova
Inria, CITI Laboratory, INSA Lyon,
Villeurbanne, France
anastasia.volkova@inria.fr

4th Richard Dufour
LS2N - Nantes Université
Nantes, France
richard.dufour@ls2n.fr

5th Adeline Granet
Valeuriad
Nantes, France
granet.adeline@valeuriad.fr

6th Nicolas Greffard
Valeuriad
Nantes, France
greffard.nicolas@valeuriad.fr

***Abstract*—Quantization is a widely adopted technique to reduce memory footprint and computational cost in neural networks. While quantizing pre-trained models is effective, retraining is often required for extreme quantization formats. Fine-tuning, on the other hand, enables the adaptation of general-purpose models to specific domains, but quantization can significantly degrade their performance.**

In this work, we investigate the training cost of fine-tuned and quantized language models. By formalizing the computational trade-off between domain adaptation and fine-tuning, we demonstrate that domain-specialized checkpoints exhibit greater robustness to quantization noise. Our findings establish a viable blueprint for deploying high-performance biomedical NLP models in resource-constrained, edge environments.

***Index Terms*—Quantization, QAT, Biomedical NLP, BERT-based model**

I. Introduction

Deploying Transformer-based language models in privacy-sensitive domains such as healthcare or defense presents a dual challenge: strict data governance requirements that mandate on-premise execution, and the high computational cost associated with standard high-precision inference. Although generative Large Language Models (LLMs) have demonstrated remarkable capabilities [4], [9], heir deployment typically relies on large GPU clusters or cloud APIs, making them incompatible with privacy-first, low-latency edge environments. In contrast, encoder-only architectures like BERT [5] remain attractive for Natural Language Processing (NLP) discriminative tasks like classification or Named Entity Recognition thanks to their parameter efficiency and fine-tuning stability [21].

Embedded systems and edge accelerators offer a promising hardware substrate for such scenarios, providing deterministic latency, on-premise deployment, and high energy efficiency [11]. However, porting Transformers to such constrained devices requires aggressive model compression to fit within limited on-chip memory and bandwidth resources. While recent advances in extreme quantization, notably binary and ternary formats [18], [24], show that dramatic reductions in numerical precision are theoretically achievable, their impact on domain-specific language models remains poorly understood. This is especially critical in biomedical NLP, where semantic precision, robustness, and consistency are essential.

This work investigates the trade-off between unsupervised domain adaptation and supervised fine-tuning under quantization constraints. Full pre-training from scratch is avoided due to computational costs, and simple Post-Training Quantization is rejected as it is often destructive at 2 bits. We therefore explore a *Quantization-Aware Continual Pre-training* strategy that allocates a compute budget parameter between domain adaptation and fine-tuning, aiming to realign specialized models with a discrete numerical grid without inducing catastrophic forgetting.

The remainder of this manuscript reviews hardware-aware compression techniques in Section II, presents the co-design methodology in Section III, describes the experimental protocol on French biomedical corpora in Section IV, and analyzes the resilience of general versus domain-specific models under extreme quantization in Section V. The tool and experiments are open-source and available[1].

II. Related Work

This research is situated at the intersection of efficient NLP, quantization algorithms, and hardware acceleration.

[1]The code is available on GitHub.

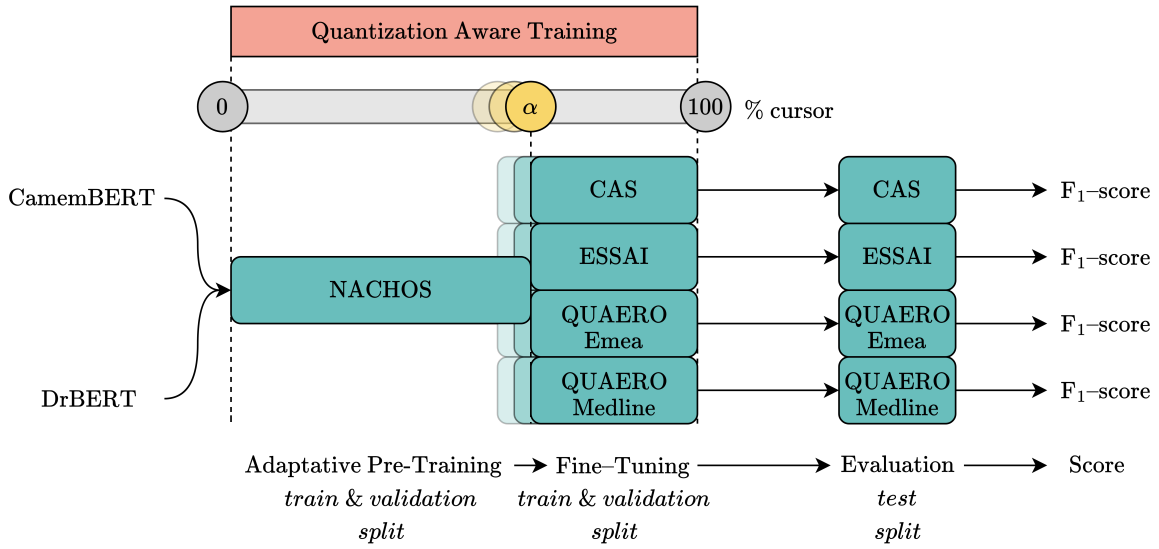

Fig. 1. **Experimental Pipeline for Quantization–Aware Adaptation.** The process integrates the α trade–off within a unified quantization context. **Blue:** Datasets flow (General $\rightarrow$ Domain $\rightarrow$ Task). **Peach:** The Quantization scheme (QAT) spans both phases to force the model to adapt its weights to the discrete grid (e.g., $\{-1, 0, 1\}$) during feature acquisition. **Yellow:** The α cursor modulates the ratio between unsupervised recovery (DAPT) and supervised specialization (Fine–Tuning).

A. Hardware-Aware Model Compression

Early quantization efforts focused on 8-bit inference for CPU and GPU acceleration [13], but research has increasingly turned to sub-4-bit methods to overcome memory and bandwidth limitations. Post-Training Quantization (PTQ) [2] offers a low-cost solution but frequently degrades performance below 4 bits. In contrast, Quantization-Aware Training (QAT) simulates discretization noise during forward passes, allowing the network to learn robust weights. Techniques like LSQ [6] have pushed the boundary of uniform integer quantization down to 2 bits by making the quantization scale a trainable parameter.

B. Ternary Networks and the BitNet Paradigm

Ternary weight networks $(-1, 0, 1)$ offer a more radical form of compression than scalar quantization, eliminating multipliers and enabling highly efficient integer datapaths. The BitNet b1.58 architecture [18] demonstrated that even large generative LLMs can retain strong capabilities with 1.58-bit weights when trained from scratch on massive corpora. Existing studies on BitNet focus primarily on large-scale generative models trained from scratch on trillions of tokens. The applicability of this ternary paradigm to the adaptation of existing medium-sized encoders (BERT [5]) for specialized domains remains an open research question.

C. Domain Adaptation in NLP

Domain specialization of language models is commonly achieved via Domain-Adaptive Pre-Training (DAPT) [10], which continues pre-training on in-domain text before supervised fine-tuning. While effective for full-precision models [14], the interaction between this adaptation phase and extreme quantization is not well understood. Recent work suggests that quantization acts as a regularizer [23] that may require additional data to converge [25], no prior work has formalized the tradeoff between unsupervised DAPT and supervised fine-tuning when models must adapt to discrete, low-precision numerical grids. This tradeoff is central to quantizing domain-specific encoders for constrained hardware.

III. Methodology: Hardware-Aware Quantization and Adaptation

This study bridges the gap between NLP adaptation strategies and Hardware-Software Co-Design for reconfigurable architectures. The methodology is structured around two axes: formalizing the compute budget tradeoff required to recover from quantization noise (α), and defining quantization schemes. The complete experimental workflow is illustrated in Figure 1.

A. The Adaptation Trade-off: The α Parameter

Our central hypothesis posits that the optimal compute budget division between DAPT and FT is strictly dependent on the severity of the quantization scheme. Whereas fine-tuning adjusts decision boundaries for a specific task, unsupervised adaptation leverages data scaling laws [7] and has the potential to recover broader semantic representations degraded by aggressive discretization, especially under ternary weights.

The problem is formalized by defining a total compute budget B_{Total}, measured in optimization steps. A hyperparameter $\alpha \in [0, 1)$ is introduced to govern the fraction

TABLE I

THEORETICAL ESTIMATION OF MEMORY FOOTPRINT AND COMPUTATIONAL COMPLEXITY FOR THE CAMEMBERT ARCHITECTURE (SEQLEN=512). MEMORY GAIN REFERS TO THE REDUCTION IN STORAGE RELATIVE TO FP32 BASELINE. BITOPS GAIN MEASURES THE HARDWARE COST REDUCTION IN STANDARD ARITHMETIC COMPLEXITY.

Configuration	**Precision**	**Memory**		**Arithmetic (BitOPs)**	
	E/W/A (bits)	**Size (MB)**	**Gain**	**Ops (T)**	**Gain**
Baseline					
FP32	32/32/32	417.2	×1.0	44.53	×1.0
Homogeneous quantization					
LSQ E8W8A8	8/8/8	104.3	×4.0	2.78	×16.0
LSQ E4W4A4	4/4/4	52.1	×8.0	0.70	×64.0
LSQ E2W2A2	2/2/2	26.1	×16.0	0.17	×256.0
Heterogeneous quantization					
LSQ E6W2A6	6/2/6	37.4	×11.1	0.52	×85.3
BitNet E6W1.58A6	6/1.58/6	37.4	×11.1	0.52	×85.3

of this budget allocated to the unsupervised adaptation phase:

$$B_{\text{DAPT}} = \alpha B_{\text{Total}}, \tag{1}$$

$$B_{\text{FT}} = (1-\alpha) B_{\text{Total}}, \tag{2}$$

where B_{DAPT} denotes the DAPT budget and B_{FT} the downstream fine-tuning budget. The case $\alpha = 0$ represents the baseline (fine-tuning only), while $\alpha \to 1$ prioritizes domain adaptation and quantization robustness over task–specific specialization. As depicted in Figure 1, the quantization constraints remain active throughout both phases to force the model to adapt its weights to the discrete grid during feature acquisition.

B. Discretization Algorithms

To evaluate hardware efficiency, we implement two complementary quantization paradigms.

Learned Step Size Quantization (LSQ): For integer arithmetic configurations, the LSQ algorithm [6] is adopted. Unlike static methods, LSQ introduces a scaling factor s as a trainable parameter, allowing the network to dynamically adjust the quantization grid during the DAPT phase. The operation is defined as:

$$\hat{x} = \left\lfloor \text{clip}\left(\frac{x}{s}, n, p\right) \right\rceil \cdot s, \tag{3}$$

where $\hat{x}$ is the quantized representation of x, $n = -2^{b-1}$ and $p = 2^{b-1} - 1$ define the integer range for bit-width b and $\lfloor \cdot \rceil$ indicates rounding to the nearest integer.

BitNet b1.58 (Ternary quantization): For extreme compression regimes, the BitNet b1.58 approach [18] is implemented. Weights are constrained to the ternary set $\{-1, 0, 1\}$ via absolute mean normalization:

$$\tilde{x} = \left\lfloor \text{clip}\left(\frac{x}{\gamma + \epsilon}, -1, 1\right) \right\rceil, \ \gamma = \frac{1}{nm} \sum_{ij} |W_{ij}|. \tag{4}$$

During inference, this scheme replace the fundamental Matrix-Multiply-Accumulate (MAC) operation into XNOR-popcount operation, effectively eliminating the requirement for energy-intensive multipliers.

C. Hardware-Aligned Strategy: The Convergence-Efficiency Nexus

The architectural framework of this study is predicated on the structural alignment of quantization precision with the limited resources of modern embedded architectures. To facilitate the discussion, the notation ExWyAz is adopted, where x, y, and z represent the bit-widths for Embeddings (E), Weights of linear layers (W), and Activations (A), respectively.

The selection of 6-bit precision for activations (A=6) is driven by the intersection of algorithmic stability and memory efficiency. From an algorithmic perspective, preliminary empirical observations indicate that reducing activation precision to 4 bits frequently precipitates representation collapse and divergence during domain adaptation. Consequently, 6-bit precision emerges as the requisite lower bound to preserve sufficient dynamic range.

Simultaneously, this precision optimizes the data movement cost, which is the primary bottleneck in edge inference. Using 6-bit activations reduces the memory bandwidth requirements compared to standard 8-bit integer formats, while maintaining a higher representational capacity than 4-bit logic. Therefore, the E6/A6 configuration constitutes a Pareto-optimal "sweet spot" for embedded deployment and convergence.

Crucially, by combining this precision with ternary weights, the BitNet architecture transforms the fundamental operation from Multiplication to Multiplexing and Addition. In constrained embedded processors with limited hardware multiplier units, this allows for the offloading of Matrix Multiplications entirely to the arithmetic logic units (ALUs). This strategy enables higher energy efficiency by bypassing energy-intensive multipliers, validat-

TABLE II

Detailed statistics of the datasets used in this study. The domain source, the number of target labels (classes), the total volume, and the document distribution across Train/Validation/Test splits are reported. HAL holds for the French *HAL Open Archive*, HAS for the French *High Health Authority*, and FTP for *French Treebank* [1].

Dataset	Task	Source Domain	# Labels	Total Size	Train	Val.	Test
Pre-training Corpus							
NACHOS (LARGE)	MLM	HAL (80%), HAS (20%)	-	7.5 GB	2 B Tokens	5 M Tokens	-
DrBenchmark Tasks (Downstream)							
ESSAI	POS	Clinical Trials	∼30 (FTB)	7,247 docs	5,072	725	1,450
CAS	POS	Clinical Cases	∼30 (FTB)	3,790 docs	2,653	379	758
QUAERO (Emea)	NER	Drug Leaflets	10	103k words	429	389	348
QUAERO (Medline)	NER	Scientific Titles	10	103k words	833	832	833

ing the heterogeneous E6W1.58A6 configuration as the optimal trade-off for high-throughput biomedical inference.

We estimate the hardware cost of the different configurations using the BitOPs metric, which accounts for both the weight and activation bit-widths of each linear layer in the model. This metric calculates the number of bitwise multiplications performed in the linear layers. For example, for a given linear layer, the BitOPs metric is defined as:

$$\text{BitOPs}(l) = W \cdot A \cdot |w_l|, \tag{5}$$

where l is a linear layer and $|w_l|$ denotes the number of weights in the linear layer l. Consequently, the total BitOPs for a model is the sum of the BitOPs values for all its linear layers.

IV. Experimental Protocol

The experimental protocol is structured to examine how quantization interacts with domain knowledge by contrasting two adaptation pathways: 1) applying QAT to an already biomedical-specialized encoder (DrBERT [14]) and 2) applying QAT to a generalist model (CamemBERT [19]) that must acquire both domain specialization and quantization robustness within a unified training budget.

A. Adaptation Grid and Optimization Hyperparameters

To capture the non-linear dynamics of quantization recovery, a total computational budget of $B_{\text{Total}} = 30,000$ steps is established. The adaptation ratio α is discretized across the set $\{0.0, 0.1, 0.3, 0.5, 0.7, 0.9\}$ to investigate various regimes. The value $\alpha = 0.0$ serves as the strict fine-tuning baseline, while $\alpha = 0.1$ tests the hypothesis of rapid statistical realignment. Intermediate values probe the trade-off shift caused by quantization severity, and $\alpha = 0.9$ simulates an asymptotic regime where the model is maximally adapted to the domain but minimally specialized for the task.

The goal is not only to compare quantization algorithms but to expose the mechanisms through which model robustness emerges when hardware constraints, training budgets, and domain shift interact.

Optimization is governed by the AdamW algorithm parametrized with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-8}$ [17]. A learning rate of $5 \cdot 10^{-6}$ is applied in conjunction with a linear scheduler incorporating a 10% warm-up phase [5]. Regularization is enforced via a dropout rate of 10% [22] and a weight decay of 0.01. To ensure stability, gradient clipping is set to a maximum norm of 1.0 [8], and the gradient batch size is fixed at 96. Task-specific heads are initialized according to the standard BERT protocol with a range of 0.02.

B. Biomedical NLP Datasets

This study leverages a specialized biomedical corpus for pre-training and a comprehensive benchmark for downstream evaluation. Detailed statistics are provided in Table II.

The quantization-aware pre-training phase relies on the `NACHOS` dataset [14], a large-scale French biomedical corpus comprising 7.5 GB of text. The data is aggregated from high-quality institutional sources, specifically 80% from the French HAL Open Archive (scientific publications) and 20% from the French High Health Authority (HAS) (clinical guidelines). Rigorous filtering was applied to remove Optical Character Recognition noise and ensure language exclusivity.

Downstream evaluation is conducted on the `DrBenchmark` suite [15]. Part-of-Speech Tagging (POS) is assessed on the `ESSAI` corpus (clinical trial protocols) and the `CAS` corpus (clinical cases). Named Entity Recognition (NER) is evaluated using the `QUAERO` corpus, subdivided into `Emea` (drug leaflets) and `Medline` (scientific titles). Due to the lack of official partitions for ESSAI and CAS, a standard random split of 70% Train, 10% Validation, and 20% Test was applied to facilitate reproducibility. Nested entities in QUAERO were flattened to top-level granularity following BigBio standards.

C. Adaptation Strategy: DAPT vs. CPT

To isolate the effect of quantization on domain adaptation, a comparative strategy is implemented using two distinct starting checkpoints that share the CamemBERT architecture. CamemBERT [19] serves as the baseline for

DAPT, tasked with shifting from general web text to biomedical jargon while adapting to the quantization grid. Conversely, DrBERT [14] represents the Continuous Pre-Training (CPT) scenario, where the domain is constant and the objective is solely to adapt pre-existing expert weights to quantization.

A critical architectural distinction exists regarding parameter sharing. DrBERT employs weight tying between the embedding layer and the Masked Language Modeling (MLM) output head, a technique standardized by [12], [20], whereas CamemBERT does not. It is hypothesized that this mechanism may function as a regularization prior, rendering the embeddings of pre-trained models more resilient to aggressive quantization.

V. Results and Discussion

Table III reports the full F_1 results across all quantization configurations. The discussion below analyzes how precision level, model origin, and adaptation ratio (α) jointly modulate robustness in the biomedical domain.

A. Robustness Across Precision Regimes

In the high-precision regime (LSQ E8W8A8), performance remains near-lossless relative to the FP32 baseline across all datasets for both studied models (DrBERT and CamemBERT). Convergence is smooth, confirming that 8-bit integer precision captures sufficient dynamic range for biomedical semantics. As quantization becomes more aggressive with LSQ E4W4A4, instability emerges, particularly for the generalist model on NER tasks. CamemBERT fails to converge on QUAERO-Medline without pre-training ($\alpha = 0$), degrading to near-zero performance, whereas DrBERT maintains 54.6%. However, introducing a pre-training phase ($\alpha \geqslant 0.7$) allows CamemBERT to recover, reaching up to 73.4% on QUAERO-Emea with $\alpha = 0.3$. This finding indicates that adapting a generalist embedding space to a specialized domain under 4-bit constraints is feasible but necessitates a mandatory warm-up phase.

B. The 2-bit Collapse and the BitNet Anomaly

The transition to extreme compression (2-bit weights) exposes the limitations of LSQ. In LSQ E2W2A2 and LSQ E6W2A6 configurations, CamemBERT collapses systematically on NER tasks, regardless of the α value. Conversely, DrBERT maintains usability but suffers degradation compared to 4-bit configurations. This failure suggests that the gradient-based learning of the step-size in LSQ becomes unstable when the discrete space is too sparse. This failure suggests that the gradient-based learning of the step-size in LSQ becomes unstable when the discrete space is too sparse. Moreover, standard LSQ enforces symmetric quantization intervals, which are suboptimal for the asymmetric distributions of Transformer activations (GeLU). This mismatch likely precipitates the collapse of the generalist model, whereas BitNet's magnitude-based structural ternarization yields a smoother and more optimizable landscape and the initial weights are far from the target distribution.

Remarkably, the BitNet configuration (E6W1.58A6) breaks this trend. Despite using fewer weight bits than LSQ E2, ternary-constrained models remain substantially more robust. DrBERT reaches 55.2% on QUAERO-Medline, and CamemBERT—otherwise unusable under LSQ—recovers to 64.4% on QUAERO-Emea at $\alpha = 0.7$. This inversion of expectations, where 1.58-bit ternary weights outperform 2-bit LSQ, supports the hypothesis that structural ternarization yields a smoother and more optimizable landscape than scale-learned quantization when adapting pretrained Transformers to new domains.

C. Specialization as a Robustness Factor

A consistent pattern is the superior resilience of DrBERT compared to CamemBERT at low precision. In the failing LSQ E6W2A6 configuration, where CamemBERT collapses, DrBERT retains significant capabilities ($\sim$ 63.9% F1 on QUAERO-Emea). This resilience is attributed to the domain alignment of DrBERT's pre-trained embeddings and its weight-tied architecture, which likely induces a more cohesive embedding geometry resistant to quantization noise.

D. The Stabilizing Effect of Continuous Pre-Training

The adaptation ratio α emerges as a decisive factor in preventing divergence and maximizing recoverability. Even minimal unsupervised adaptation ($\alpha = 0.1$) consistently improves stability across quantization schemes. In LSQ E2W2A2, for example, DrBERT improves from 53.8% at $\alpha = 0$ to 59.4% at $\alpha = 0.1$. These results support a two-stage interpretation of QAT under domain shift, where the model must first realign its weights to the quantized manifold through low-variance MLM signals (unsupervised phase), and only then can it reliably absorb the high-variance gradients associated with downstream supervision (fine-tuning phase). Thus, increasing α does not simply provide more data—it creates a *structurally safer optimization trajectory* under extreme quantization.

VI. Conclusion and Future Work

This work introduces a hardware–aware quantization framework that unifies domain adaptation and discretization. Across all experiments, three findings consistently characterize the interaction between quantization severity, model provenance, and the allocation of the adaptation budget.

First, the architectural superiority of structural constraints over learned precision is demonstrated in extreme compression regimes. The BitNet E6W1.58A6 configuration, tailored for low-precision embedded constraints, consistently outperforms the learned LSQ E2W2A2 baseline. This confirms that at very low bit-widths, the stability

TABLE III

EVALUATION METRICS ON THE TEST SET FOR QUANTIZED DRBERT & CAMEMBERT ACROSS DIFFERENT TRAINING-FINETUNING RATIOS ($\alpha = 0.3$). NOTATION E-W-A REPRESENTS THE BIT-WIDTH FOR EMBEDDINGS, WEIGHTS, AND ACTIVATIONS. CELLS WITH "–" INDICATE MODEL DIVERGENCE OR COLLAPSE (F_1-SCORE $< 20\%$). BOLD SCORES DENOTE THE HIGHEST PERFORMANCE FOR EACH FINE-TUNED TASK WITHIN A QUANTIZATION SCHEME.

Quantization	α (%)	CAS-POS		ESSAI-POS		QUAERO-emea		QUAERO-medline	
		DrBERT	Camem.	DrBERT	Camem.	DrBERT	Camem.	DrBERT	Camem.
		F_1-score ↑ (%)		F_1-score ↑ (%)		F_1-score ↑ (%)		F_1-score ↑ (%)	
Baseline									
FP32	0	97.3	**97.9**	98.5	**98.9**	61.8	**76.7**	55.5	**58.0**
Homogeneous Quantization									
LSQ E8W8A8	0	97.3	97.4	98.5	98.4	62.9	49.6	55.4	53.0
	10	97.1	97.6	98.5	98.7	63.7	67.4	56.4	56.6
	30	97.4	**97.8**	98.5	98.8	62.2	**76.4**	55.4	55.9
	50	97.3	97.7	98.6	**98.9**	63.2	76.2	55.2	56.5
	70	97.5	97.7	98.5	98.8	64.2	74.6	55.5	**57.3**
	90	97.5	97.6	98.6	98.8	61.8	73.2	55.6	55.3
LSQ E4W4A4	0	97.0	97.2	98.3	95.6	60.0	25.7	54.6	–
	10	97.2	97.2	98.5	98.5	63.8	53.2	55.6	42.2
	30	97.2	97.1	98.5	98.4	60.7	**73.4**	55.3	42.0
	50	97.3	**97.5**	98.5	**98.7**	63.3	71.3	55.7	50.2
	70	97.5	97.3	98.4	98.7	64.0	71.5	55.4	51.8
	90	97.4	97.1	98.6	98.3	62.3	69.2	**56.1**	44.5
LSQ E2W2A2	0	95.9	–	98.0	–	53.8	–	47.1	–
	10	96.9	–	**98.4**	–	59.4	–	50.8	–
	30	96.9	–	98.3	–	58.0	–	52.3	–
	50	97.1	–	98.3	–	**60.5**	–	50.4	–
	70	97.1	–	98.3	–	60.5	–	51.6	–
	90	**97.2**	–	98.4	–	58.8	–	**52.7**	–
Heterogeneous Quantization									
LSQ E6W2A6	0	97.0	–	98.3	–	59.2	–	53.2	–
	10	97.1	–	98.5	–	63.5	–	53.8	–
	30	97.0	–	98.5	–	63.2	–	52.8	–
	50	97.1	–	98.5	–	61.6	–	52.6	–
	70	97.1	–	98.5	–	62.2	–	**55.4**	–
	90	**97.4**	–	**98.6**	–	**63.9**	–	54.7	–
BitNet E6W1.58A6	0	96.6	97.5	98.3	98.4	55.2	51.7	48.8	48.1
	10	97.2	97.6	98.5	98.6	62.2	62.7	53.3	53.0
	30	97.0	97.6	98.4	**98.6**	59.2	60.9	53.6	53.3
	50	97.2	97.6	98.5	98.6	61.7	63.4	54.2	53.8
	70	97.2	**97.6**	98.5	98.6	63.5	**64.4**	53.6	54.5
	90	97.3	97.5	98.6	98.5	64.2	63.5	**55.2**	53.6

of the analytical ternary distribution ($\{-1, 0, 1\}$) provides a more robust inductive bias than the noisy gradient-based optimization of 2-bit scalar quantization. Second, the domain provenance of the model proves critical: specialized checkpoints (DrBERT) exhibit significantly higher resilience to quantization noise than their generalist counterparts (CamemBERT), which succumb to the "double burden" of simultaneous domain adaptation and weight discretization. Third, the introduction of a continuous pre-training phase ($\alpha > 0$) is identified as a necessary condition for convergence in aggressive regimes, allowing the model to realign its internal representation on the discrete grid prior to task-specific fine-tuning.

These results suggest several concrete research directions. A first extension is the translation of ternary-aligned architectures to general-purpose CPUs leveraging AVX/NEON instructions, enabling efficient biomedical inference beyond specialized accelerators. Further reductions in computational footprint may be obtained by integrating structured sparsity with SIMD-friendly tensor formats. Additionally, extending the study to asymmetric quantization schemes (e.g., LSQ+ [3]) could resolve the collapse observed in 2-bit scalar baselines by better capturing the rectified distribution of activations. Beyond encoders, assessing whether the synergy between domain specialization, hardware-constrained quantization, and adaptation strategies generalizes to larger generative models remains an open and compelling question. Finally, a theoretical characterization of adaptation under quantization—particularly the interplay between α, domain shift, and discrete optimization—may yield principled guidelines for budget allocation in low-precision training regimes.

VII. ACKNOWLEDGEMENTS

This work was granted access to the HPC resources of IDRIS under the allocation 2025-AD011014536R1 made by GENCI. It was also partly financially supported by the BPI Partages project and partially funded by the French National Research Agency project PEPR AI under the reference ANR-23-PEIA-0010.

VIII. Ethical and Impact Statement

A. Datasets and Bias

Experiments rely exclusively on anonymized open data [15], in strict adherence to their respective licenses. While focusing on computational efficiency, this study does not audit potential social biases present in pretraining corpora. Although crucial, analyzing the specific interplay between extreme quantization and model fairness remains outside the scope of this work.

B. Environmental Impact (Green AI)

This research aligns with Frugal AI principles by providing heuristics to minimize computational costs. Consistent with transparency, the cumulative carbon footprint for all 175 experimental runs was tracked via CodeCarbon [16]. The total computing duration of 869 hours consumed 435 kWh, resulting in an estimated emission of 1.0 kg CO_2eq. This low footprint benefits from the Jean-Zay (IDRIS/CNRS) supercomputer's decarbonized energy mix. By identifying optimal training regimes and flagging divergent low-bit configurations, this study directly contributes to preventing resource wastage in future hardware-aware NLP research.

References

[1] A. Abeillé, L. Clément, and A. Kinyon, "Building a treebank for French," in *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, M. Gavrilidou, G. Carayannis, S. Markantonatou, S. Piperidis, and G. Stainhauer, Eds. Athens, Greece: European Language Resources Association (ELRA), May 2000.

[2] R. Banner, Y. Nahshan, and D. Soudry, "Post training 4-bit quantization of convolutional networks for rapid-deployment," *Advances in neural information processing systems*, vol. 32, 2019.

[3] Y. Bhalgat, J. Lee, M. Nagel, T. Blankevoort, and N. Kwak, "Lsq+: Improving low-bit quantization through learnable offsets and better initialization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

[4] DeepSeek-AI, A. Liu, and al., "Deepseek-v3 technical report," 2025.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, vol. 1, 2019, p. 2.

[6] S. K. Esser, J. L. McKinstry, D. Bablani, R. Appuswamy, and D. S. Modha, "Learned step size quantization," in *International Conference on Learning Representations*, 2020.

[7] D. Ganguli, D. Hernandez, L. Lovitt, A. Askell, Y. Bai, A. Chen, T. Conerly, N. Dassarma, D. Drain, N. Elhage *et al.*, "Predictability and surprise in large generative models," in *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 2022, pp. 1747–1764.

[8] A. Graves, G. Wayne, and I. Danihelka, "Neural turing machines," 2014.

[9] D. Groeneveld and al., "OLMo: Accelerating the science of language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 15 789–15 809.

[10] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[11] S. Huang, C. Pearson, R. Nagi, J. Xiong, D. Chen, and W.-m. Hwu, "Accelerating sparse deep neural networks on fpgas," in *2019 IEEE High Performance Extreme Computing Conference (HPEC)*, 2019, pp. 1–7.

[12] H. Inan, K. Khosravi, and R. Socher, "Tying word vectors and word classifiers: A loss framework for language modeling," in *International Conference on Learning Representations*, 2017.

[13] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 2704–2713.

[14] Y. Labrak, A. Bazoge, R. Dufour, M. Rouvier, E. Morin, B. Daille, and P.-A. Gourraud, "Drbert: A robust pre-trained model in french for biomedical and clinical domains," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 16 207–16 221.

[15] Y. Labrak, A. Bazoge, O. El Khettari, M. Rouvier, P. C. D. Beaufils, N. Grabar, B. Daille, S. Quiniou, E. Morin, P.-A. Gourraud *et al.*, "Drbenchmark: A large language understanding evaluation benchmark for french biomedical domain," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 5376–5390.

[16] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the carbon emissions of machine learning," *arXiv preprint arXiv:1910.09700*, 2019.

[17] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2017.

[18] S. Ma, H. Wang, L. Ma, L. Wang, W. Wang, S. Huang, L. Dong, R. Wang, J. Xue, and F. Wei, "The era of 1-bit llms: All large language models are in 1.58 bits," *CoRR*, 2024.

[19] L. Martin, B. Muller, P. O. Suarez, Y. Dupont, L. Romary, É. V. de la Clergerie, D. Seddah, and B. Sagot, "Camembert: a tasty french language model," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7203–7219.

[20] O. Press and L. Wolf, "Using the output embedding to improve language models," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 157–163.

[21] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, "Is chatGPT a general-purpose natural language processing task solver?" in *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[23] P. Stock, A. Fan, B. Graham, E. Grave, R. Gribonval, H. Jegou, and A. Joulin, "Training with quantization noise for extreme model compression," in *International Conference on Learning Representations*, 2021.

[24] J. Wang, H. Zhou, T. Song, S. Cao, Y. Xia, T. Cao, J. Wei, S. Ma, H. Wang, and F. Wei, "Bitnet.cpp: Efficient edge inference for ternary LLMs," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 9305–9322.

[25] W. Zhang, L. Hou, Y. Yin, L. Shang, X. Chen, X. Jiang, and Q. Liu, "TernaryBERT: Distillation-aware ultra-low bit BERT," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 509–521.