A Systematic Review on Language Model Compression: Perspectives for Efficiency and Sustainability in Information Systems

Lair Anderson de P. Mesquita [1] **, Saulo Anderson F. de Oliveira** [1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Fortaleza – CE – Brazil

lair.anderson.paula07@aluno.ifce.edu.br, saulo.oliveira@ifce.edu.br

Abstract. Research Context: *The rapid adoption of Large Language Models (LLMs) offers unprecedented opportunities for Intelligent Information Systems. However, their high computational cost creates significant deployment barriers in resource-constrained environments. This limited access for smaller organizations and developing regions reinforces social and economic disparities in the adoption of those systems.* ***Scientific and/or Practical Problem:*** *LLMs demand excessive resources, which hinders their integration into sustainable INFORMATION SYSTEMS. Furthermore, there is a lack of systematic evidence on how compression affects robustness, factuality, and applicability in different contexts.* ***Proposed Solution and/or Analysis:*** *This work presents a systematic literature review analyzing 30 peer-reviewed studies on LLM compression methods. The review identifies trade-offs, evaluate empirical results, and highlight gaps for future research.* ***Related IS Theory:*** *We interpret computational efficiency and accessibility as strategic resources (Resource-Based View) that enable the development of sustainable, and equitable digital infrastructures.* ***Research Method:*** *This study follows PICO structure with inclusion/exclusion criteria to works published between 2018 and January 2025. It was centered around eight research questions addressing efficiency, robustness, hardware compatibility, knowledge preservation, hybridization, architectural adaptation, evaluation metrics, and industrial viability.* ***Summary of Results:*** *Our study shows that efficiency and system-focused methods are the most developed and investigated for LLM compression for training and inference efficiency. Among these techniques, quantization consistently outperforms pruning and error-based methods, balancing performance and design savings. However, research on robustness, generalization, and practical use remains limited. Recent advances in shared hardware designs show potential for scalability and lower power consumption. However, practical studies on the long-term sustainability, ethical issues, and social impacts of LLM compression are rare in current scientific literature.* ***Contributions and Impact to IS area:*** *This work contributes by mapping key challenges in LLM compression research, offering insights to design efficient, sustainable, and socially responsible intelligent systems that align AI's technical progress with goals of inclusion and environmental responsibility.*

1. Introduction

The rapid expansion of LLMs has significantly advanced the field of Natural Language Processing (NLP), enabling state-of-the-art performance across diverse tasks such as text classification, machine translation, summarization, and question answering [Vaswani et al. 2017, Devlin et al. 2018, Brown et al. 2020]. Such models, characterized by their Transformer architectures and extensive parameter counts, often exceeding hundreds of billions, have demonstrated remarkable abilities in learning and generalization across multiple domains [Touvron et al. 2023].

However, the exponential growth in model size and computational demand has raised serious concerns regarding scalability, energy efficiency, and equitable accessibility [Ganesh et al. 2020, Rasheed et al. 2023]. Training and deploying such large models typically require specialized hardware, such as high-end Graphics Processing Units (GPUs) or Tensor Processing Units (TPUs), and substantial energy resources, contributing to environmental and economic constraints [Strubell et al. 2019, Schwartz et al. 2020]. Consequently, their use remains largely restricted to major research institutions and technology corporations, limiting broader adoption in resource-constrained environments such as public universities, educational institutions, and small enterprises.

Model compression has emerged as a response, with quantization, pruning, knowledge distillation, and hardware-aware strategies reducing model size and inference costs while preserving performance [Han et al. 2015, Gholami et al. 2021]. Recent methods, including AWQ [Lin et al. 2024a], GPTQ [Frantar et al. 2023], QLoRA [Dettmers et al. 2023a], and SparseGPT [Frantar and Alistarh 2023], show that compression is not only a computational optimization but also a key enabler for intelligent systems across domains such as healthcare, education, and government. But, open challenges persist, notably regarding robustness, fairness, and knowledge preservation [Rasheed et al. 2023, Hooker et al. 2020, Xu et al. 2024], as well as ethical concerns such as the environmental and social sustainability of AI systems [Strubell et al. 2019, Li et al. 2024, Bender et al. 2021].

In that regard, our systematic review, grounded in the PICO methodology (Population, Intervention, Comparison, and Outcomes), systematically synthesizes studies on LLM compression techniques published between 2018 and January 2025. Our work focuses on evaluating multiple dimensions of model performance, including computational efficiency, predictive accuracy, robustness, knowledge retention, and environmental sustainability. By integrating evidence from empirical research, it aims to map how compression approaches affect the balance between model compactness and functional reliability in NLP applications deployed in organizational environments, where people, processes, and technologies converge to enhance information systems and decision support.

Ultimately, our study seeks to identify trade-offs that reconcile technical innovation with the ethical, organizational, and societal imperatives of deploying intelligent information systems. In doing so, it contributes to the ongoing discourse on responsible AI, highlighting how model efficiency intersects with accessibility and sustainability in the broader ecosystem of large-scale machine learning. This perspective aligns with current reflections on sustainable and ethical AI deployment, which emphasize that innovation must be accompanied by environmental awareness and social responsibility [van Wynsberghe 2021, Sanderson et al. 2024]. Within the Brazilian rese-

arch context, similar discussions have emerged in the field of Information Systems, where ethical principles and governance frameworks for artificial intelligence are increasingly recognized as central to the responsible adoption of intelligent technologies [Siqueira de Cerqueira et al. 2021].

LLM compression goes beyond reducing model size and inference cost: it directly impacts how Information Systems that embed language models are designed, deployed, and governed (e.g., conversational agents, decision support, RAG pipelines, and automation). In IS settings, compression affects operational cost and scalability (SLA-driven latency/throughput), infrastructure choices (cloud vs. edge/on-premise), maintainability and monitoring (updates, drift, observability), and organizational risk management (privacy, security, compliance). Because efficiency gains also reduce compute and energy demand, compression can support more sustainable and inclusive IS by enabling broader access in resource-constrained environments and aligning AI adoption with cost–benefit and environmental targets. By synthesizing evidence through IS-relevant outcomes (efficiency, robustness/safety, and real-world feasibility), this review links LLM compression research to the organizational and societal factors that shape intelligent systems in practice.

The structure of this work follows a systematic and transparent approach to ensure methodological rigor and analytical depth. The **Methodology** section details the research design, the formulation of guiding questions, and the criteria used to include and exclude studies, ensuring the reproducibility of the review process. Subsequently, the search and selection procedures are described, outlining how databases were queried and how data extraction was conducted to synthesize the selected literature. The **Results and Discussion** section explore the findings, offering a multidimensional analysis of trends, trade-offs, and research gaps across LLM compression studies. Finally, the **Conclusion** summarizes key insights, theoretical implications, and directions for future research, reinforcing the contribution of this review to the broader field of responsible and sustainable AI.

2. Methodology

Our study adopts the principles of a Systematic Literature Review (SLR), aiming to ensure transparency, replicability, and comprehensive coverage in the analysis of compression techniques for LLMs. The methodology is structured according to the **PICO** model [Methley et al. 2014], a mnemonic framework widely used in Evidence-Based Medicine (EBM) to formulate clear and well-scoped research questions. PICO decomposes a question into four core components: **Population**, Patient, or Problem (the group or condition of interest); **Intervention** (the treatment, diagnostic test, or exposure under investigation); **Comparison** (the control or alternative condition, such as standard care or no intervention); and **Outcome** (the measurable outcome of interest). Rigorous application of PICO supports systematic and efficient retrieval of relevant literature by translating the research question into explicit, auditable search concepts.

In our context, this structure is adapted to support the systematic formulation of research questions by decomposing them into four key elements: (*i*) the **Population**, defined as Transformer-based LLMs applied to textual Natural Language Processing tasks; (*ii*) the **Intervention**, represented by compression techniques such as quantization, pru-

ning, and knowledge distillation; (*iii*) the **Comparison**, involving uncompressed baselines or alternative compression approaches; and (*iv*) the **Outcomes**, encompassing computational efficiency, model quality, robustness, and sustainability. This structured approach promotes consistent study identification, screening, and evaluation, strengthening reproducibility and enabling comparative synthesis across heterogeneous contributions.

We adopted PICO to structure the research questions and operationalize search and selection because the review focuses on a clear intervention—LLM compression—and compares outcomes expressed as measurable trade-offs (e.g., quality vs. latency, memory, and energy). While more common in evidence-based disciplines, PICO provides methodological clarity by enforcing explicit inclusion/exclusion criteria and making the search strategy auditable through a direct mapping between keywords and question elements. To reduce limitations of a single framework, we combined PICO with a multi-database search, a transparent screening procedure, and a predefined extraction spreadsheet for consistency. In addition, the protocol and synthesis capture not only algorithmic results but also deployment-relevant variables (latency/throughput, memory footprint, hardware, and feasibility evidence), aligning the review with system-level and organizational concerns in Information Systems.

Our goal is to identify, categorize, and evaluate compression methods applied to Transformer-based LLMs, analyzing these techniques not only in terms of technical contributions but also in terms of their broader implications for Information Systems, including efficiency, robustness, sustainability, and responsible deployment.

2.1. Research Questions

The central research question guiding this study is:

Which compression techniques applied to Large Language Models (LLMs) in Natural Language Processing (NLP) are most effective in balancing predictive performance and computational efficiency, while preserving robustness, factual knowledge, and applicability in real-world scenarios?

Building on this central research question, the reviewers formulated a set of eight questions (see Table 1), each addressing a specific analytical dimension of LLM compression. These dimensions capture aspects that are both technically significant to the Machine Learning community and strategically relevant to the Information Systems domain. Accordingly, the reviewers organized the research questions across complementary analytical perspectives, as summarized in Table 2.

2.2. Inclusion and Exclusion Criteria

We established explicit criteria to ensure the relevance and quality of selected studies:

Inclusion Criteria:

- Articles published between 2018 and January 2025;
- Peer-reviewed conference or journal papers;
- Empirical studies applying compression to LLMs with at least 100M parameters;
- Studies reporting at least one relevant outcome: accuracy, perplexity, inference speed, memory usage, robustness, or factual knowledge preservation, and;
- Articles written in English or Portuguese.

Table 1. Research questions Derived from the PICO Framework

ID	Research Subquestion
Q1	What are the most effective trade-offs between accuracy and computational efficiency (memory, latency, and energy consumption), and how do they impact system sustainability?
Q2	How does compression affect robustness against adversarial inputs and the ability to generalize across out-of-domain tasks?
Q3	Which compression techniques are most compatible with deployment on constrained hardware (e.g., Central Processing Units (CPUS), mobile devices), supporting inclusivity and accessibility of intelligent systems?
Q4	To what extent do compression methods preserve factual knowledge and reasoning capabilities compared to original models, guaranteeing trust and transparency in decision-making?
Q5	How do hybrid approaches (e.g., pruning combined with quantization) compare to isolated techniques in balancing efficiency and performance?
Q6	What are the most effective strategies for specific architectures, and how do architectural features influence compression outcomes?
Q7	Which evaluation metrics beyond accuracy (e.g., latency, throughput, usability, energy) provide the most informative assessment for organizational and societal adoption?
Q8	What industrial case studies demonstrate the long-term viability, scalability, and sustainability of compressed LLMs in real-world?

Table 2. Research Questions and Their Analytical Dimensions

ID	Focus Area	Analytical Dimension
Q1	Accuracy vs. Efficiency	Trade-offs in memory, latency, energy consumption.
Q2	Robustness	Impact on adversarial resistance and out-of-domain generalization.
Q3	Hardware Compatibility	Deployment on CPUs, edge devices, and FPGAs.
Q4	Knowledge Preservation	Retention of factual knowledge and reasoning.
Q5	Hybrid Approaches	Effectiveness of combining compression strategies.
Q6	Architecture-Specific Strategies	Adaptations for well-known LLMs
Q7	Evaluation Metrics	Beyond accuracy: latency, energy, usability.
Q8	Industrial Case Studies	Long-term viability and sustainability of compressed LLMs.

Exclusion Criteria:

- Theoretical-only works without empirical validation;
- Studies on small-scale models (<100M parameters);
- Narrative reviews without experimental results;
- Works irrelevant to NLP (e.g., compression of CNNs for computer vision), and;
- Duplicated publications.

2.3. Databases

We conducted searches across the following electronic databases: IEEE Xplore, ACM Digital Library, Scopus, SpringerLink, ArXiv, and Google Scholar (for complementary coverage).

2.4. Search Strategy

The search strategy was designed to maximize coverage and recall while maintaining precision with respect to the scope defined by the PICO framework. To this end, the query strings combined three main groups of terms: (i) synonyms and variations related to LLMs, (ii) different compression techniques, and (iii) evaluation metrics relevant to NLP. This ensured that retrieved studies included not only direct references to compression, but also works addressing efficiency, robustness, and real-world applicability. A representative query string is as follows:

```
(("large language models"OR LLM OR "transformer
models") AND (compression OR quantization OR
pruning OR distillation OR "model slimming"OR
"structured pruning") AND (performance OR
accuracy OR perplexity OR inference OR memory OR
latency OR robustness OR "knowledge retention")
AND (NLP OR "natural language processing"))
```

The terms were iteratively refined through pilot searches to balance recall and specificity. For instance, the inclusion of `model slimming` and `structured pruning` allowed capturing articles that use alternative terminologies for pruning, while `knowledge retention` and `robustness` ensured coverage of studies that evaluate beyond accuracy metrics.

Queries were adapted to the specific syntax and indexing rules of each database. For Scopus and Web of Science, Boolean operators and phrase matching were strictly applied. In IEEE Xplore and ACM Digital Library, controlled vocabularies and metadata fields (title, abstract, keywords) were leveraged to narrow down results. Google Scholar was used only to complement potential gaps, given its broader coverage but lower precision.

To ensure relevance and quality, results were filtered by:

- **Publication period:** 2018–January 2025, capturing both the emergence of LLMs and the rapid evolution of compression techniques.
- **Language:** English and Portuguese.
- **Type of publication:** peer-reviewed journal articles, conference proceedings, and selected high-impact preprints.

The resulting corpus formed the basis for the screening and eligibility stages described in the following subsections.

2.5. Study Selection and Data Extraction

The selection process, illustrated in Figure 1, was designed to ensure rigor, transparency, and reproducibility. It was carried out in three main stages. The first stage, **identification**, involved aggregating all articles retrieved from the selected databases using the predefined search queries into a reference manager. This procedure ensured that all potentially relevant studies were initially considered without exclusion bias. In the **screening** phase, we removed duplicates, and evaluated the titles and abstracts of the remaining works manually to verify compliance with the inclusion and exclusion criteria established in the PICO framework. At this stage, we excluded theoretical works without empirical validation, studies involving models with fewer than 100M parameters, or those not addressing NLP tasks. Finally, in the **inclusion** phase, the remaing articles underwent full-text analysis to confirm their relevance, and final decisions regarding inclusion were made.

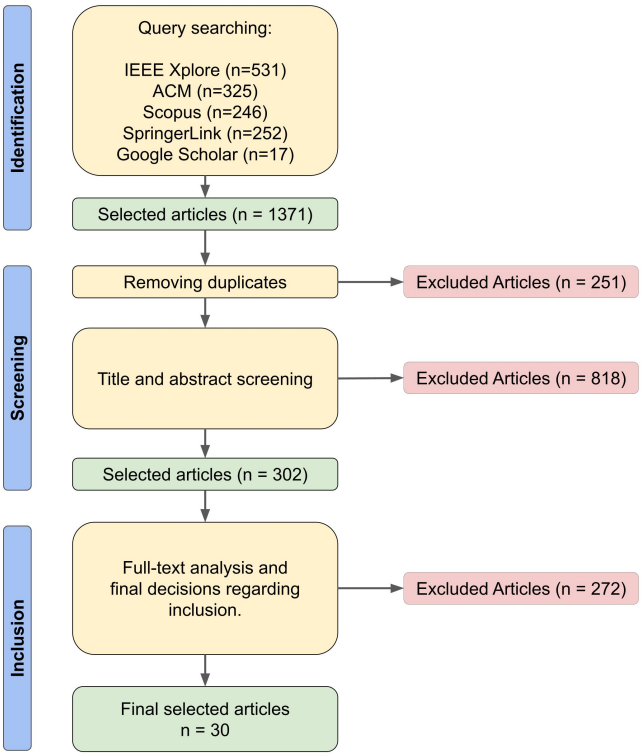

Figure 1. Methodology

After the identification stage, we retrieved 1,371 records. During screening, we removed 251 duplicates and excluded 818 records based on title and abstract. In the eligibility stage, we excluded 272 studies after full-text assessment, resulting in 30 included articles. Two reviewers independently conducted screening and full-text selection using the predefined inclusion and exclusion criteria. Disagreements were resolved through discussion until consensus, and final decisions were recorded in the extraction spreadsheet. The complete extraction table is available in this repository. As a complementary source, we queried Google Scholar after the primary database search to reduce retrieval bias and capture potentially relevant studies not indexed elsewhere. We used the same keyword groups, screened the first 10 result pages (sorted by relevance), applied the same eligibility criteria, and removed duplicates against the main database set before inclusion.

Table 3. Classification criteria for each Research Question in the systematic review, including the corresponding Information Systems (IS) meaning.

ID	0 – Low	0.5 – Moderate	1 – High	IS meaning (score = 1)
Q1	No quantitative trade-off; focuses solely on efficiency or accuracy.	Partial analysis with limited benchmarks or missing metrics.	Clear balance between performance and efficiency across multiple tasks.	Actionable efficiency–quality trade-offs for sizing.
Q2	No mention of robustness or generalization.	Indirect evaluation via zero/few-shot; no stress testing.	Explicit robustness evaluation or mitigation.	Reliability under drift, prompt variation, and attacks.
Q3	Experiments limited to high-end GPUs.	Mentions hardware constraints without empirical validation.	Demonstrates efficiency on constrained or heterogeneous hardware.	Portability/feasibility across heterogeneous infrastructure.
Q4	No analysis of factuality or reasoning; only numerical metrics.	Indirect evaluation using reasoning benchmarks.	Direct assessment of reasoning and factual retention with adaptive methods.	Trustworthiness for knowledge-intensive decision support.
Q5	Purely single-method.	Partial integration with limited interaction.	Hybrid/multi-stage design with demonstrated improvements.	Configurable pipelines for workload- and constraint-aware optimization.
Q6	Single-architecture testing; no structural discussion.	Multi-architecture results without structural sensitivity analysis.	Examines how structure affects compression effectiveness and stability.	Guides model selection and upgrade maintainability.
Q7	Reports only accuracy or perplexity.	Includes limited system-level metrics.	Includes latency, throughput, memory, and energy.	Operational performance/cost monitoring for deployment.
Q8	Conceptual or lab-only experiments.	Deployment potential but no production validation.	Validated in industrial or large-scale inference settings.	Production readiness: MLOps, observability, and scalable serving.

For each included study, the reviewers extracted and organized information using a predefined spreadsheet to ensure consistency and comparability across methods. The

extracted fields covered model characteristics (architecture, parameter count, and task domain) and the applied compression technique, including its class (e.g., post-training quantization, quantization-aware training, pruning, distillation, and hybrid approaches such as pruning+quantization or weight/activation/KV-cache quantization for serving). We also captured the experimental setup, distinguishing calibration, training/fine-tuning, and evaluation datasets when applicable, and recorded the hardware platform used for experimentation and deployment-oriented measurements (e.g., consumer GPUs, A100/H100-class accelerators, multi-GPU settings, and CPU baselines). Finally, we summarized reported outcomes in terms of efficiency (latency, throughput, memory footprint, and model size), quality (perplexity and downstream benchmark scores), robustness-related evaluations when provided, and evidence of factual knowledge and reasoning retention under compression.

Additionally, we classified each research question using a three-level scale (0–Low; 0.5–Moderate; 1–High), as defined in Table 3. This scoring enabled the quantitative aggregation and visualization reported in the **Results and Discussion** section, preserving a clear link between extracted primary evidence and the synthesized findings.

2.6. Threats to validity.

This SLR is subject to common validity threats.

(i) Selection and publication bias: relevant studies may have been missed due to indexing limitations, query formulation, or venue coverage, and the inclusion of selected high-impact preprints may introduce variability in rigor relative to peer-reviewed work.

(ii) Construct validity: widely used proxies (perplexity and benchmark accuracy) may not capture robustness, safety, or behavior under distribution shift, and reported speedups may not transfer across heterogeneous deployment stacks.

(iii) External validity: conclusions may not generalize to proprietary models or industrial workloads shaped by compliance, monitoring, and operational constraints.

To mitigate these threats, we used explicit inclusion/exclusion criteria, a predefined extraction template, and cross-checked key fields (model, technique, datasets, and hardware) during synthesis; nevertheless, some residual bias and incompleteness are unavoidable given the fast-evolving nature of LLM compression research.

3. Results and Discussion

This section is dedicated to the systematic analysis of the Quality Assessment Table (5), which summarizes the performance and methodological comprehensiveness of the 30 most relevant works on LLM compression Table (4). Initial results demonstrate a robust trend toward High-Quality methodologies (average score of 6.01), corroborating the technical maturity of the field. However, the central discussion lies in the divergence of focus observed among the research criteria. While the scientific community has reached saturation in the analysis of the Trade-offs between Accuracy and Efficiency (Q1=1.00) and the Use of Comprehensive Metrics (Q7=1.00), the significantly lower average in areas such as Robustness and Generalization (Q2=0.53) and Industrial Case (Q8=0.57) establishes a clear research agenda for the future. The results will be detailed to identify innovations that optimize operational efficiency, while also exposing gaps that provide scope for future works.

Table 4. Selected articles

Title	Reference
AWQ: Activation-Aware Weight Quantization	[Lin et al. 2024a]
GPTQ: Accurate Post-Training Quantization	[Frantar et al. 2023]
SmoothQuant: Accurate & Efficient PTQ for LLMs	[Xiao et al. 2024]
Wanda: Pruning by Weights and Activations	[Sun et al. 2024a]
LLM-Pruner: On the Structural Pruning of LLMs	[Ma et al. 2023]
SparseGPT: LLMs Pruned in One-Shot	[Frantar and Alistarh 2023]
SpQR: Sparse-Quantized Representation	[Dettmers et al. 2023b]
SqueezeLLM: Dense-and-Sparse Quantization	[Kim et al. 2024]
Sheared-LLaMA: Structured Pruning + Retraining	[Xia et al. 2024]
ZeroQuant: Efficient Post-Training Quantization for Large Transformers	[Yao et al. 2022]
KVQuant: Quantization of Key-Value Cache for Efficient LLM Inference	[Hooper et al. 2024]
QServe: W4A8KV4 Quantization with System Co-Design for LLM Serving	[Lin et al. 2024b]
OmniQuant: Omnidirectionally Calibrated Quantization for LLMs	[Shao et al. 2024]
QLoRA: Efficient Finetuning of Quantized Large Language Models	[Dettmers et al. 2023a]
LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale	[Dettmers et al. 2022]
FlatQuant: Flatness Matters for Large Language Model Quantization	[Sun et al. 2024b]
QuaRot: Outlier-Free 4-bit Inference in Rotated LLMs	[Ashkboos et al. 2024]
QuIP#: Hadamard Incoherence and Lattice Codebooks for LLM Quantization	[Tseng et al. 2024]
AQLM: Additive Quantization for LLMs	[Egiazarian et al. 2024]
LoftQ: LoRA-Finetuning-Aware Quantization	[Li et al. 2023]
SpinQuant: LLM Quantization with Learned Rotations	[Liu et al. 2024b]
EfficientQAT: Efficient Quantization-Aware Training for LLMs	[Chen et al. 2024]
APTQ: Attention-aware Post-Training Mixed-Precision Quantization	[Guan et al. 2024]
OWQ: Outlier-aware Weight Quantization	[Lee et al. 2024]
ZipLM: Inference-Aware Structured Pruning for LLMs	[Kurtic et al. 2023]
RPTQ: Reorder-based Post-Training Quantization for LLMs	[Yuan et al. 2023]
KIVI: Tuning-Free 2-bit Quantization for KV Cache	[Zirui Liu et al. 2023]
LLM-QAT: Data-Free Quantization-Aware Training for LLMs	[Liu et al. 2023]
Efficient Post-Training Quantization with FP8 Formats	[Shen et al. 2024]
MiniCache: Compressing Key-Value Cache Along Depth	[Liu et al. 2024a]

Table 5. Converting qualitative ratings into numerical values with total sum and quality level

Article	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Total	Quality
AWQ (2024)	1	0.5	1	0.5	0	1	1	1	6.0	High
GPTQ (2023)	1	0.5	0.5	0.5	0	1	1	0.5	5.0	Mid
SmoothQuant (2022)	1	0.5	1	1	0.5	1	1	0.5	6.5	High
Wanda (2024)	1	0.5	1	0.5	0.5	1	1	0.5	6.0	High
LLM-Pruner (2023)	1	0.5	0.5	0.5	1	1	1	0.5	6.0	High
SparseGPT (2023)	1	0.5	1	1	1	1	1	0.5	7.0	High
SpQR (2023)	1	0.5	1	0.5	0.5	1	1	0.5	6.0	High
SqueezeLLM (2024)	1	0.5	1	0.5	0.5	1	1	0.5	6.0	High
Sheared-LLaMA (2023)	1	0.5	1	1	1	1	1	0.5	7.0	High
ZeroQuant (2022)	1	0.5	1	1	0.5	1	1	0.5	6.5	High
KVQuant (2024)	1	0.5	1	0.5	0.5	1	1	0.5	6.0	High
QServe (2024)	1	0.5	1	0.5	1	1	1	1	7.0	High
OmniQuant (2024)	1	0.5	1	0.5	0.5	1	1	0.5	6.0	High
QLoRA (2023)	1	0.5	1	1	1	1	1	1	7.5	High
LLM.int8() (2022)	1	0.5	1	0.5	1	0.5	1	0.5	6.0	High
FlatQuant (2024)	1	0.5	1	0.5	1	0.5	1	0.5	6.0	High
QuaRot (2024)	1	0.5	1	0.5	1	0.5	1	0.5	6.0	High
QuIP# (2024)	1	0.5	1	0.5	0.5	0.5	1	0.5	5.5	Mid
AQLM (2024)	1	0.5	1	0.5	1	1	1	0.5	6.5	High
LoftQ (2023)	1	1	0.5	0.5	1	0.5	1	0.5	6.0	High
SpinQuant (2024)	1	0.5	1	0.5	1	1	1	0.5	6.5	High
EfficientQAT (2024)	1	0.5	0.5	0.5	1	1	1	0.5	5.5	Mid
APTQ (2024)	1	0.5	1	0.5	1	1	1	0.5	6.5	High
OWQ (2024)	1	0.5	1	0.5	1	1	1	0.5	6.5	High
ZipLM (2023)	1	0.5	1	0.5	1	1	1	0.5	6.5	High
RPTQ (2023)	1	0.5	0.5	0.5	1	1	1	0.5	5.5	Mid
KIVI (2023)	1	0.5	1	0.5	1	1	1	0.5	6.5	High
LLM-QAT (2023)	1	1	1	0.5	0.5	1	1	0.5	6.5	High
Efficient PTQ FP8 (2024)	1	0.5	1	0.5	0.5	0.5	1	1	6.0	High
MiniCache (2024)	1	0.5	1	0.5	1	0.5	1	0.5	6.0	High

Table 6. Quantitative Distribution of Reviewed Works

Category	Number of Studies	Percentage
High (score ≥ 6.0)	26	87%
Mid ($4.0 \leq score < 6.0$)	4	13 %
Low (< 4.0)	0	0%
Total	30	100%

Examination of Q1 reveals a strong methodological convergence across the literature, where all reviewed studies consistently prioritize the balance between computational efficiency and predictive accuracy, achieving the maximum score (1.0). This consensus underscores that efficiency–performance trade-offs have become a foundational objective in LLM compression research. Most works demonstrate that well-designed compression pipelines, particularly those involving quantization and pruning, can substantially reduce inference cost, memory usage, and energy demand while preserving accuracy within statistically insignificant deviations from full-precision baselines [Lin et al. 2024a, Frantar et al. 2023, Xiao et al. 2024, Shao et al. 2024, Sun et al. 2024a, Frantar and Alistarh 2023]. The prevalence of quantization-based methods, including post-training and quantization-aware approaches, reflects their superior scalability and ease of deployment [Lin et al. 2024a, Frantar et al. 2023, Shao et al. 2024, Chen et al. 2024, Liu et al. 2023]. Techniques such as 4-bit or even sub-4-bit quantization, when combined with calibration or low-rank adaptation, allow models to maintain linguistic coherence and representational fidelity while achieving notable throughput improvements across hardware platforms [Lin et al. 2024a, Frantar et al. 2023, Dettmers et al. 2023a, Lee et al. 2024, Ashkboos et al. 2024, Liu et al. 2024b].

In contrast, analysis of Q2 exposes robustness and generalization as some of the least developed dimensions in current LLM compression research, with an average score of 0.53 indicating insufficient methodological depth. Most studies emphasize efficiency metrics, such as latency, throughput, or parameter reduction, while implicitly assuming that preserved accuracy on standard benchmarks implies maintained robustness [Frantar et al. 2023, Xiao et al. 2024, Shao et al. 2024, Frantar and Alistarh 2023, Sun et al. 2024a]. However, this assumption often masks fragility under real-world perturbations, including adversarial noise, prompt variation, or domain shifts. Few works explicitly evaluate post-compression resilience through controlled stress tests, perturbation-based analysis, or out-of-distribution (OOD) generalization, and evaluations are often limited to perplexity and standard zero-/few-shot benchmarks rather than robustness-specific protocols [Dettmers et al. 2022, Liu et al. 2023, Chen et al. 2024, Zirui Liu et al. 2023, Hooper et al. 2024]. As a result, the field lacks a consistent understanding of how compression-induced modifications, such as quantization noise, weight pruning, or reduced representational redundancy, affect the model's capacity to generalize and maintain reliable behavior under uncertainty [Ashkboos et al. 2024, Liu et al. 2024b, Lee et al. 2024, Ma et al. 2023, Kurtic et al. 2023].

Some studies mitigate these limitations with robustness-aware adaptation, including quantization-aware fine-tuning and calibration-oriented optimization [Chen et al. 2024, Liu et al. 2023, Dettmers et al. 2023a]. They indicate that compression may increase sensitivity to distribution shifts or prompt variation, but retraining, low-rank adaptation, and calibration can reduce downstream degradations [Dettmers et al. 2023a, Liu et al. 2023, Chen et al. 2024]. However, the lack of standardized robustness benchmarks for compressed LLMs still limits cross-study comparability, as most evaluations rely on perplexity and standard zero-/few-shot suites rather than stress tests [Lin et al. 2024a, Frantar et al. 2023, Xiao et al. 2024, Shao et al. 2024]. Progress therefore requires robustness frameworks that include adversarial, stochastic, and domain-level perturbations, which is essential for dependable deployment in Information Systems [Lin et al. 2024b, Hooper et al. 2024, Zirui Liu et al. 2023, Liu et al. 2024a].

In the context of Q3, the result reveals that hardware-aware compression has become one of the most mature and empirically grounded areas within LLM efficiency research, as reflected by the high average score of 0.90. Many studies report clear gains in computational efficiency, reducing latency and memory footprint through low-bit inference and optimized execution paths on modern accelerators [Lin et al. 2024a, Xiao et al. 2024, Frantar et al. 2023, Ashkboos et al. 2024, Liu et al. 2024b, Shen et al. 2024]. This progress indicates that compression is increasingly guided by hardware–algorithm co-design, where methods are optimized not only for accuracy but also for alignment with platform characteristics such as efficient matrix-multiply kernels, memory movement patterns, and attention/KV-cache execution [Lin et al. 2024b, Hooper et al. 2024, Zirui Liu et al. 2023, Liu et al. 2024a]. In particular, serving-oriented system work demonstrates that end-to-end throughput depends on jointly quantizing weights, activations, and KV cache while restructuring the runtime to reduce dequantization overhead and improve kernel efficiency [Lin et al. 2024b]. As a result, hardware-aware strategies contribute directly to scaling large models in research and industrial infrastructures, enabling higher throughput under fixed memory budgets and making compression a practical lever for more sustainable, high-performance AI systems [Dettmers et al. 2022, Hooper et al. 2024, Zirui Liu et al. 2023, Shen et al. 2024].

Despite this progress, systematic evaluation on resource-constrained devices (CPUs, mobile, edge) remains limited, as most studies benchmark primarily on server-grade GPUs, leaving questions about portability, thermal efficiency, and long-term reliability [Lin et al. 2024a, Frantar et al. 2023, Xiao et al. 2024, Shao et al. 2024, Ashkboos et al. 2024, Liu et al. 2024b]. Evidence on constrained settings (e.g., consumer GPUs or CPU measurements) is still fragmented [Dettmers et al. 2022, Dettmers et al. 2023b, Kurtic et al. 2023, Egiazarian et al. 2024, Li et al. 2023], and practical energy savings are often not quantified with operational or environmental metrics [Lin et al. 2024b, Shen et al. 2024]. Recent advances in quantized adaptation and mixed-precision strategies point to more hardware-tailored compression [Dettmers et al. 2023a, Li et al. 2023, Guan et al. 2024, Lee et al. 2024]; future work should extend hardware-aware frameworks to include sustainability, cost, and scalability for real organizational deployments [Lin et al. 2024b, Hooper et al. 2024, Zirui Liu et al. 2023, Liu et al. 2024a, Shen et al. 2024].

The analysis of Q4 highlights a moderate yet uneven methodological commitment to preserving factual knowledge and reasoning integrity in compressed LLMs, reflected in a mean score of 0.57. While many studies report stable downstream performance after compression, they often infer knowledge retention indirectly through broad benchmark suites such as GLUE or MMLU, without isolating factual consistency or reasoning coherence as independent evaluation targets [Kurtic et al. 2023, Li et al. 2023, Dettmers et al. 2023a, Liu et al. 2023]. This reliance on aggregate metrics can obscure subtle degradations in factual recall, logical inference, and domain-specific reasoning that are not well captured by standard accuracy or perplexity proxies [Lin et al. 2024a, Frantar et al. 2023, Xiao et al. 2024, Shao et al. 2024]. In most cases, compression is validated primarily through accuracy-based measures and general-purpose evaluation suites, which may fail to expose epistemic distortions introduced by parameter reduction, quantization noise, or pruning-induced sparsity [Frantar and Alistarh 2023, Sun et al. 2024a, Ma et al. 2023]. As a result, the literature still tends to treat factual and

reasoning preservation as secondary outcomes rather than as central dimensions of model integrity [Dettmers et al. 2022, Ashkboos et al. 2024, Lee et al. 2024].

A subset of works moves beyond this limitation by incorporating fine-tuning, distillation, or calibration stages to mitigate factual drift and reasoning decay [Dettmers et al. 2023a, Li et al. 2023, Liu et al. 2023, Chen et al. 2024]. These studies use adaptation pipelines such as low-rank fine-tuning on quantized backbones, fine-tuning-aware quantization, or data-free quantization-aware training, and evaluate compressed models on knowledge- and reasoning-intensive benchmarks (e.g., MMLU and commonsense suites) to probe factual and reasoning retention [Dettmers et al. 2023a, Li et al. 2023, Liu et al. 2023, Chen et al. 2024]. Their results indicate that degradation is architecture- and method-dependent, influenced by choices such as weight-only vs. weight–activation quantization and by outlier/rotation handling or calibration strategies that preserve critical representational ranges [Ashkboos et al. 2024, Liu et al. 2024b, Lee et al. 2024, Shao et al. 2024]. However, limited standardization of methodologies and domain-sensitive benchmarks still constrains cross-study comparability [Lin et al. 2024a, Frantar et al. 2023, Xiao et al. 2024]. Future work should therefore prioritize systematic frameworks for measuring factual and epistemic reliability and make knowledge-aware evaluation a core component of compression analysis, bridging efficiency and cognitive fidelity for responsible deployment in knowledge-intensive information systems [Lin et al. 2024b, Hooper et al. 2024, Liu et al. 2024a].

The results for Q5 indicate a clear methodological transition toward hybrid compression strategies, with an average score of 0.77 reflecting their growing prominence and maturity. This trend suggests a move beyond single-technique optimization toward multi-objective frameworks that combine quantization, pruning, and distillation or fine-tuning in complementary ways [Frantar and Alistarh 2023, Dettmers et al. 2023b, Yao et al. 2022, Li et al. 2023, Dettmers et al. 2023a]. Such integration allows researchers to balance conflicting goals, reducing model size and inference cost while maintaining (or in some cases improving) predictive performance [Dettmers et al. 2023b, Yao et al. 2022, Li et al. 2023]. Hybrid approaches also enhance adaptability by enabling finer control over compression intensity under different task requirements and hardware constraints [Lin et al. 2024b, Hooper et al. 2024, Zirui Liu et al. 2023]. In several cases, these methods leverage distinct algorithmic strengths: pruning introduces structural sparsity, quantization reduces precision overhead, and adaptation or distillation helps preserve representational capacity after aggressive compression [Frantar and Alistarh 2023, Sun et al. 2024a, Kurtic et al. 2023, Dettmers et al. 2023a, Li et al. 2023].

Despite these advancements, many hybrid frameworks remain exploratory, with limited methodological standardization and reproducible evaluation pipelines [Yao et al. 2022, Dettmers et al. 2023b, Li et al. 2023]. Systematic analysis of interaction effects on robustness, energy efficiency, and cross-architecture transfer is still rare, and validation often relies on perplexity and standard downstream suites rather than stress tests or longitudinal evidence [Lin et al. 2024a, Frantar et al. 2023, Xiao et al. 2024, Frantar and Alistarh 2023]. Moreover, few works evaluate operational feasibility under production constraints and serving stacks [Lin et al. 2024b, Shen et al. 2024, Hooper et al. 2024]. Closing these gaps requires consolidated protocols and context-aware hybrid strategies that align technical gains with organizatio-

nal and sustainability goals [Lin et al. 2024b, Hooper et al. 2024, Zirui Liu et al. 2023, Liu et al. 2024a].

In the context of Q6, the results show that most studies demonstrate strong cross-architecture generalization, with an average score of 0.90, indicating that many compression methods transfer effectively across Transformer-based families and variants (e.g., GPT-/OPT-like models, LLaMA-family, and encoder–decoder backbones) [Frantar et al. 2023, Lin et al. 2024a, Xiao et al. 2024, Yao et al. 2022, Dettmers et al. 2023a, Dettmers et al. 2022]. This consistency suggests reliance on broadly applicable techniques, such as layer-wise post-training quantization, weight-only or weight–activation quantization, and structured pruning or adaptation methods that operate over standard Transformer blocks [Frantar et al. 2023, Lin et al. 2024a, Shao et al. 2024, Sun et al. 2024a, Ma et al. 2023, Kurtic et al. 2023]. However, this generality can come at the expense of deeper architectural understanding, as many works treat models as largely interchangeable and emphasize aggregate performance metrics rather than analyzing how internal mechanisms influence compressibility [Lin et al. 2024a, Frantar et al. 2023, Xiao et al. 2024, Shao et al. 2024].

Only a smaller group of studies adopts architecture-aware design, linking compression outcomes and runtime bottlenecks to attention structure, activation outliers, and KV-cache organization [Hooper et al. 2024, Zirui Liu et al. 2023, Liu et al. 2024a, Ashkboos et al. 2024, Liu et al. 2024b, Lee et al. 2024]. These works show that compressibility depends on architectural choices and memory/activation behavior, so ignoring such dependencies can limit optimization [Ashkboos et al. 2024, Liu et al. 2024b, Hooper et al. 2024, Zirui Liu et al. 2023]. Bridging this gap requires adaptive, architecture-informed compression that accounts for structural diversity and deployment constraints [Lin et al. 2024b, Guan et al. 2024, Shen et al. 2024], which is essential for sustainable and maintainable use in complex Information Systems [Lin et al. 2024b, Hooper et al. 2024, Zirui Liu et al. 2023, Liu et al. 2024a].

Regarding Q7, the result demonstrates complete methodological convergence across the reviewed literature, with all studies attaining the maximum score and evidencing a collective shift toward multidimensional evaluation frameworks. This consolidation reflects the field's maturation: researchers increasingly assess compression efficacy not only through accuracy or perplexity but also through a broader suite of system and deployment indicators, including latency, throughput, memory utilization, and efficiency under constrained execution [Lin et al. 2024a, Frantar et al. 2023, Xiao et al. 2024, Shao et al. 2024, Frantar and Alistarh 2023, Sun et al. 2024a]. This pattern is particularly visible in works that explicitly tie compression to runtime behavior, such as serving-oriented co-design and KV-cache optimizations, where memory movement and kernel efficiency become central drivers of end-to-end performance [Lin et al. 2024b, Hooper et al. 2024, Zirui Liu et al. 2023, Liu et al. 2024a]. By quantifying efficiency across both computational and operational dimensions, these studies establish a coherent methodological standard that links model design choices directly to practical implications for deployment and resource management [Dettmers et al. 2022, Shen et al. 2024, Ashkboos et al. 2024, Liu et al. 2024b].

However, this consolidation also reveals a limitation: metric uniformity can obscure qualitative and contextual aspects of system performance. Most studies emphasize

hardware-level indicators while underrepresenting organizational and environmental dimensions such as long-term energy impact, carbon footprint, and cost–benefit trade-offs [Lin et al. 2024a, Frantar et al. 2023, Xiao et al. 2024, Shao et al. 2024]. Even system-oriented work rarely integrates maintainability, interpretability, or accessibility into evaluation protocols [Lin et al. 2024b, Hooper et al. 2024, Zirui Liu et al. 2023]. Future research should extend multidimensional evaluation beyond technical efficiency to include contextual and ethical dimensions, aligning compression assessment with sustainable and responsible AI needs in Information Systems [Lin et al. 2024b, Liu et al. 2024a, Shen et al. 2024].

The analysis of Q8 exposes a persistent gap between the experimental maturity of LLM compression methods and their real-world operational deployment, as reflected in the modest average score of 0.57. While the technical literature reports substantial progress in efficiency, accuracy preservation, and hardware- or architecture-aware optimization, most studies stop short of validating these advances in production-scale contexts [Lin et al. 2024a, Frantar et al. 2023, Xiao et al. 2024, Shao et al. 2024, Frantar and Alistarh 2023, Sun et al. 2024a]. Experimental evaluations are typically conducted under controlled conditions using benchmark datasets and single-node hardware configurations, which only partially reflect the constraints of enterprise-scale or distributed environments [Frantar et al. 2023, Lin et al. 2024a, Xiao et al. 2024, Ashkboos et al. 2024, Liu et al. 2024b]. Consequently, deployment is often framed as a feasible outcome rather than an empirically validated setting, leaving open questions about maintainability, reliability under sustained inference, and interoperability with serving stacks, monitoring, and data governance processes [Dettmers et al. 2022, Hooper et al. 2024, Zirui Liu et al. 2023, Liu et al. 2024a].

Although sustainability is a recurring motivation, long-term impact is still weakly quantified. Most studies report efficiency proxies (latency, throughput, memory) but rarely translate them into lifecycle-relevant indicators such as energy use, carbon proxies, or cost–benefit trade-offs under sustained demand. These measures are also seldom contextualized within organizational ecosystems where maintainability and interpretability shape total cost of ownership and risk. Future work should adopt deployment-oriented protocols that pair system metrics with lifecycle-aware sustainability reporting to better link compression gains to long-term organizational and environmental outcomes.

Beyond technical constraints, limited practical validation also reflects organizational and socio-technical barriers to real-world adoption. Deploying compressed LLMs at scale requires not only algorithmic efficiency, but also alignment with security, privacy, and sustainability requirements that are often underrepresented in research prototypes. Production settings further demand monitoring, incident handling, and update strategies that fit operational budgets and environmental targets. The lack of longitudinal evidence on cost savings, energy reductions, and end-user impact reinforces the need for evaluation frameworks that connect compression results to Information Systems practices and real deployment constraints.

The results also indicate a methodological evolution in LLM compression research, marked by the consolidation of quantization as a dominant approach for balancing efficiency and model quality, and by an increasing emphasis on scalability, hardware adaptability, and deployment-oriented performance.

However, this maturity is uneven. Limited coverage of robustness, generalization, and industrial or organizational feasibility is driven by technical, economic, and socio-organizational constraints. High-confidence robustness assessment requires explicit stress-testing, yet many studies rely on indirect indicators such as few-shot or zero-shot performance, limiting cross-study comparability. Robust generalization assessment also increases cost because it requires broader pipelines (multiple datasets, OOD settings, and ablations), so secondary effects such as brittleness and sensitivity to prompt variation are often left outside the primary scope.

Industrial feasibility is difficult to document because credible evidence requires production-like validation, while real deployments often rely on proprietary data and infrastructure and face compliance constraints that limit publication of end-to-end results. As a result, despite progress in hardware-aware co-design and inference optimization, the literature still reports limited production-grade validation and limited use of standardized sustainability indicators that connect efficiency gains to measurable operational impact. These gaps motivate future work that complements efficiency with replicable robustness and evaluation under domain and distribution shifts, industrial case studies with operational metrics, and more consistent sustainability reporting for responsible adoption in Information Systems.

Finally, while computational efficiency and hardware awareness are well established, robustness and the preservation of reasoning and factual reliability remain underexplored and are often treated as secondary objectives rather than core requirements. Future work should therefore prioritize holistic compression strategies that make reliability a first-class goal, enabling compressed LLMs to operate dependably in socio-technical environments.

4. Conclusion

This systematic review consolidates the research landscape on LLM updates applied to NLP, synthesizing 30 empirical studies. The findings demonstrate a methodological update in efficiency optimization and trade-off analyses, particularly in the relationship between accuracy and efficiency (Q1) and in the standardization of evaluation metrics (Q7). Quantization-based methodologies have emerged as the most dominant and widespread paradigm in the field. However, the research highlights opportunities for future research, especially in the areas of robustness of compressed models (Q2), the effectiveness of preserving intrinsic and extrinsic knowledge (Q4), and the industrial scalability of solutions (Q8).

The evolution of the field points to a transition toward practical applicability, where hybrid approaches and hardware-aware methods signal a growing focus on deployment in production environments and computational sustainability. From a broader perspective of Information Systems, these results highlight the socio-technical relevance of LLM specifications, which establishes a direct link between the technological efficiency achieved and the organizational adaptability required for sustainable innovation in the context of large-scale basic Artificial Intelligence solutions.

Acknowledgments

The authors acknowledge the academic and computational support that made this study possible. This article employed generative AI tools, specifically ChatGPT, to assist in the preparation of tables.

Reference

Ashkboos, S., Mohtashami, A., Croci, M. L., Li, B., Cameron, P., Jaggi, M., Alistarh, D., Hoefler, T., and Hensman, J. (2024). Quarot: Outlier-free 4-bit inference in rotated llms.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623.

Brown, T., Mann, B., Ryder, N., Subbiah, M., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Chen, M., Shao, W., Xu, P., Wang, J., Gao, P., Zhang, K., and Luo, P. (2024). Efficientqat: Efficient quantization-aware training for large language models.

Dettmers, T., Lewis, M., Belkada, Y., and Zettlemoyer, L. (2022). Llm.int8(): 8-bit matrix multiplication for transformers at scale.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023a). Qlora: Efficient finetuning of quantized llms.

Dettmers, T., Svirschevski, R., Egiazarian, V., Kuznedelev, D., Frantar, E., Ashkboos, S., Borzunov, A., Hoefler, T., and Alistarh, D. (2023b). Spqr: A sparse-quantized representation for near-lossless llm weight compression.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Egiazarian, V., Panferov, A., Kuznedelev, D., Frantar, E., Babenko, A., and Alistarh, D. (2024). Extreme compression of large language models via additive quantization.

Frantar, E. and Alistarh, D. (2023). Sparsegpt: Massive language models can be accurately pruned in one-shot.

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. (2023). Gptq: Accurate post-training quantization for generative pre-trained transformers.

Ganesh, P., Chen, Y., Lou, X., et al. (2020). Compressing large-scale transformer-based models: A survey. *arXiv preprint arXiv:2006.09282*.

Gholami, A., Kim, S., Dong, Z., et al. (2021). A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2103.13630*.

Guan, Z., Huang, H., Su, Y., Huang, H., Wong, N., and Yu, H. (2024). Aptq: Attention-aware post-training mixed-precision quantization for large language models.

Han, S., Mao, H., and Dally, W. J. (2015). Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*.

Hooker, S., Courville, A., Clark, G., Dauphin, Y., and Frome, A. (2020). Compressed to impress: Understanding the effects of model compression on fairness, robustness, and accuracy. In *NeurIPS Workshop on Machine Learning for the Developing World*.

Hooper, C., Kim, S., Mohammadzadeh, H., Mahoney, M. W., Shao, Y. S., Keutzer, K., and Gholami, A. (2024). Kvquant: Towards 10 million context length llm inference with kv cache quantization.

Kim, S., Hooper, C., Gholami, A., Dong, Z., Li, X., Shen, S., Mahoney, M. W., and Keutzer, K. (2024). Squeezellm: Dense-and-sparse quantization.

Kurtic, E., Frantar, E., and Alistarh, D. (2023). Ziplm: Inference-aware structured pruning of language models.

Lee, C., Jin, J., Kim, T., Kim, H., and Park, E. (2024). Owq: Outlier-aware weight quantization for efficient fine-tuning and inference of large language models.

Li, P., Yang, J., Wierman, A., and Ren, S. (2024). Towards environmentally equitable ai via geographical load balancing.

Li, Y., Yu, Y., Liang, C., He, P., Karampatziakis, N., Chen, W., and Zhao, T. (2023). Loftq: Lora-fine-tuning-aware quantization for large language models.

Lin, J., Tang, J., Tang, H., Yang, S., Chen, W.-M., Wang, W.-C., Xiao, G., Dang, X., Gan, C., and Han, S. (2024a). Awq: Activation-aware weight quantization for llm compression and acceleration.

Lin, Y., Tang, H., Yang, S., Zhang, Z., Xiao, G., Gan, C., and Han, S. (2024b). Qserve: W4a8kv4 quantization and system co-design for efficient llm serving.

Liu, A., Liu, J., Pan, Z., He, Y., Haffari, G., and Zhuang, B. (2024a). Minicache: Kv cache compression in depth dimension for large language models.

Liu, Z., Oguz, B., Zhao, C., Chang, E., Stock, P., Mehdad, Y., Shi, Y., Krishnamoorthi, R., and Chandra, V. (2023). Llm-qat: Data-free quantization aware training for large language models.

Liu, Z., Zhao, C., Fedorov, I., Soran, B., Choudhary, D., Krishnamoorthi, R., Chandra, V., Tian, Y., and Blankevoort, T. (2024b). Spinquant: Llm quantization with learned rotations.

Ma, X., Fang, G., and Wang, X. (2023). Llm-pruner: On the structural pruning of large language models.

Methley, A. M., Campbell, S., Chew-Graham, C., McNally, R., and Cheraghi-Sohi, S. (2014). Pico, picos and spider: A comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews. *BMC health services research*, 14(1):579.

Rasheed, F., Karim, A., et al. (2023). A survey on large language models: Applications, challenges, limitations, and future directions. *arXiv preprint arXiv:2307.10169*.

Sanderson, C., Schleiger, E., Douglas, D., Kuhnert, P., and Lu, Q. (2024). Resolving ethics trade-offs in implementing responsible ai. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 1208–1213.

Schwartz, R., Dodge, J., Smith, N. A., and Etzioni, O. (2020). Green ai. *Communications of the ACM*, 63(12):54–63.

Shao, W., Chen, M., Zhang, Z., Xu, P., Zhao, L., Li, Z., Zhang, K., Gao, P., Qiao, Y., and Luo, P. (2024). Omniquant: Omnidirectionally calibrated quantization for large language models.

Shen, H., Mellempudi, N., He, X., Gao, Q., Wang, C., and Wang, M. (2024). Efficient post-training quantization with fp8 formats.

Siqueira de Cerqueira, J. A., Acco Tives, H., and Dias Canedo, E. (2021). Ethical guidelines and principles in the context of artificial intelligence. In *Proceedings of the XVII Brazilian Symposium on Information Systems*, SBSI '21, New York, NY, USA. Association for Computing Machinery.

Strubell, E., Ganesh, A., and McCallum, A. (2019). Energy and policy considerations for deep learning in nlp. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3645–3650.

Sun, M., Liu, Z., Bair, A., and Kolter, J. Z. (2024a). A simple and effective pruning approach for large language models.

Sun, Y., Liu, R., Bai, H., Bao, H., Zhao, K., Li, Y., Hu, J., Yu, X., Hou, L., Yuan, C., Jiang, X., Liu, W., and Yao, J. (2024b). Flatquant: Flatness matters for llm quantization.

Touvron, H., Lavril, T., Izacard, G., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Tseng, A., Chee, J., Sun, Q., Kuleshov, V., and Sa, C. D. (2024). Quip: Even better llm quantization with hadamard incoherence and lattice codebooks.

van Wynsberghe, A. (2021). Ai for sustainability and the sustainability of ai. *AI and Ethics*, 1(3):213–218.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Xia, M., Gao, T., Zeng, Z., and Chen, D. (2024). Sheared llama: Accelerating language model pre-training via structured pruning.

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., and Han, S. (2024). Smoothquant: Accurate and efficient post-training quantization for large language models.

Xu, Z., Wang, H., Zhang, L., et al. (2024). Exploring the robustness of compressed large language models. *arXiv preprint arXiv:2402.07895*.

Yao, Z., Aminabadi, R. Y., Zhang, M., Wu, X., Li, C., and He, Y. (2022). Zeroquant: Efficient and affordable post-training quantization for large-scale transformers.

Yuan, Z., Niu, L., Liu, J., Liu, W., Wang, X., Shang, Y., Sun, G., Wu, Q., Wu, J., and Wu, B. (2023). Rptq: Reorder-based post-training quantization for large language models.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Braverman, V., Beidi Chen, and Hu, X. (2023). Kivi : Plug-and-play 2bit kv cache quantization with streaming asymmetric quantization.