
PERFORMANCE TRADE-OFFS OF OPTIMIZING SMALL LANGUAGE MODELS FOR E-COMMERCE

Josip Tomo Licardo

Faculty of Informatics

Juraj Dobrila University of Pula
Zagrebačka 30
52100 Pula, Croatia
jllicardo@unipu.hr

Nikola Tanković

Faculty of Informatics

Juraj Dobrila University of Pula
Zagrebačka 30
52100 Pula, Croatia
ntankov@unipu.hr

October 28, 2025

ABSTRACT

Large Language Models (LLMs) offer state-of-the-art performance in natural language understanding and generation tasks. However, the deployment of leading commercial models for specialized tasks, such as e-commerce, is often hindered by high computational costs, latency, and operational expenses. This paper investigates the viability of smaller, open-weight models as a resource-efficient alternative. We present a methodology for optimizing a one-billion-parameter Llama 3.2 model for multilingual e-commerce intent recognition. The model was fine-tuned using Quantized Low-Rank Adaptation (QLoRA) on a synthetically generated dataset designed to mimic real-world user queries. Subsequently, we applied post-training quantization techniques, creating GPU-optimized (GPTQ) and CPU-optimized (GGUF) versions. Our results demonstrate that the specialized 1B model achieves 99% accuracy, matching the performance of the significantly larger GPT-4.1 model. A detailed performance analysis revealed critical, hardware-dependent trade-offs: while 4-bit GPTQ reduced VRAM usage by 41%, it paradoxically slowed inference by 82% on an older GPU architecture (NVIDIA T4) due to dequantization overhead. Conversely, GGUF formats on a CPU achieved a speedup of up to 18 \times in inference throughput and a reduction of over 90% in RAM consumption compared to the FP16 baseline. We conclude that small, properly optimized open-weight models are not just a viable but a more suitable alternative for domain-specific applications, offering state-of-the-art accuracy at a fraction of the computational cost.

Keywords Large Language Models · E-commerce · Fine-tuning · QLoRA · Quantization · GPTQ

1 Introduction

The field of artificial intelligence has been fundamentally reshaped by the advent of Large Language Models (LLMs), sophisticated deep learning systems that demonstrate a remarkable capacity to understand, generate, and reason with human language [1]. Foundational models such as Meta’s Llama 3 series [2], Google’s Gemma 3 [3], and Alibaba’s Qwen3 [4] now represent the state-of-the-art, powering applications that span from complex code generation to nuanced creative writing and scientific discovery. Their advanced capabilities have unlocked new paradigms for human-computer interaction and process automation across countless industries.

One of the most promising domains for LLM application is e-commerce, a highly competitive landscape where user experience is a critical determinant of success. The ability of a system to accurately interpret and act upon user requests expressed in natural, often informal, language is a significant competitive advantage. This paradigm, known as conversational commerce, aims to create more intuitive and personalized shopping experiences [5]. Tasks such as robust product classification from noisy user descriptions [6] and understanding complex user intent are central to this vision [7]. Major industry players like eBay have already begun developing their own in-house, domain-specific LLMs to capitalize on these opportunities, underscoring the strategic importance of this technology [8].

However, the deployment of the most powerful, state-of-the-art commercial models, such as OpenAI's GPT-4, presents significant practical and economic barriers. These models are computationally massive, and their use via API services generates continuous operational costs, creates vendor lock-in, and raises concerns regarding data privacy and security. Furthermore, the immense energy, water, and carbon footprint associated with serving billions of queries from these large-scale models is a growing concern for sustainable AI development [9, 10, 11]. The high cost and resource intensity of these "frontier" models thus create a substantial obstacle to their widespread adoption, particularly for small to medium-sized enterprises.

In response to these challenges, a powerful trend has emerged: the rise of smaller, highly optimized, open-weight language models. A growing body of research demonstrates that these smaller models, when specialized for a specific domain, can achieve performance comparable or even superior to that of much larger, general-purpose models. For instance, fine-tuned small models have been shown to outperform GPT-4 on specific tasks like arithmetic [12], mental health understanding [13], and developing pedagogical tools [14]. A large-scale study fine-tuning over 300 models confirmed that specialized models can consistently rival GPT-4 on narrow tasks, suggesting that a "many small models" approach is a viable and efficient strategy [15]. This paradigm shift from monolithic, "one-size-fits-all" models to a diverse ecosystem of specialized agents motivates our research.

Two key technologies are central to unlocking the potential of these smaller models: parameter-efficient fine-tuning (PEFT) and post-training quantization (PTQ). PEFT techniques allow a pretrained model to be adapted to a new task by training only a small fraction of its parameters. Low-Rank Adaptation (LoRA) has become a standard approach [16], and its successor, Quantized LoRA (QLoRA), further democratized this process by enabling the fine-tuning of large models on consumer-grade hardware through aggressive 4-bit quantization of the base model during training [17].

Concurrently, post-training quantization has become essential for efficient inference. PTQ methods reduce the numerical precision of a model's weights (and sometimes activations) after training, leading to significant reductions in memory footprint and potential increases in inference speed. Techniques like GPTQ [18] and AWQ [19] have demonstrated the ability to compress models to 4-bit precision with minimal accuracy loss. However, recent research highlights that the benefits of quantization are not universal; they are highly dependent on the model size, task, and underlying hardware architecture, creating a complex web of trade-offs between performance, energy, and output quality [20, 21, 22].

This paper bridges these research threads by conducting a comprehensive, hardware-aware investigation into the practical viability of a small, optimized open-weight model for a critical e-commerce task. We hypothesize that by combining QLoRA-based fine-tuning with aggressive post-training quantization, it is possible to specialize a 1 billion parameter model to achieve accuracy parity with the much larger, state-of-the-art GPT-4.1 on a structured intent recognition task. Crucially, we extend beyond mere accuracy to analyze the real-world performance trade-offs, measuring inference speed, memory consumption, and energy efficiency on both GPU and CPU environments to provide a holistic view of the model's operational characteristics.

Our contributions are fourfold. First, we introduce an end-to-end methodology for building a highly efficient, specialized language model for multilingual e-commerce intent recognition, including a novel synthetic data generation process based on "metaprompting". Second, we supply empirical evidence that a properly fine-tuned 1-billion-parameter Llama 3.2 model reaches 99% accuracy on this task, matching GPT-4.1. Third, we present a detailed, hardware-aware performance analysis that surfaces nuanced, sometimes counter-intuitive effects of quantization, showing that 4-bit GPTQ can slow inference on older GPUs while GGUF formats yield significant speedups on CPUs. Fourth, we release our multilingual synthetic dataset for e-commerce intent recognition to support further research and reproducibility.

The remainder of this paper is organized as follows. Section 2 discusses the background and related work in model specialization and optimization. Section 3 details our methodology, including task definition, dataset generation, and the experimental pipeline. Section 4 presents our results on model accuracy and operational performance. Section 5 discusses the implications of our findings, and Section 6 concludes the paper with a summary and directions for future work.

2 Background and Related Work

The rapid evolution of LLMs is characterized by a dual trend: the scaling of massive, general-purpose models and the proliferation of smaller, highly specialized models. While large models set performance benchmarks, their practical deployment is often limited. This has spurred significant research into methodologies for creating and optimizing smaller models that are both powerful and efficient. Our work is situated at the intersection of three key research areas: domain specialization through fine-tuning, performance optimization via quantization, and the generation of high-quality data for these processes.

2.1 From Generalists to Specialists: The Power of Fine-Tuning

The prevailing paradigm is shifting from relying on a single, monolithic model to deploying smaller models tailored for specific tasks. A growing body of evidence shows that this specialization is highly effective. Large-scale studies, such as the one conducted in LoRA Land, have demonstrated that fine-tuning a diverse set of 7B models with LoRA can result in performance that consistently rivals or exceeds that of GPT-4 on many narrow tasks [15]. This effect holds across various domains; for instance, the Goat model, a fine-tuned LLaMA, surpassed GPT-4 on complex arithmetic tasks [12], and specialized models have shown comparable performance to GPT-4 in sensitive fields such as mental health understanding [13] and for creating pedagogical tools [14]. This trend is also prominent in industry, with companies such as eBay developing their own in-house, e-commerce-focused LLMs to gain a competitive edge [8].

The primary mechanism for achieving this specialization is fine-tuning. However, full fine-tuning, which updates all of a model's billions of parameters, is resource-prohibitive. This has led to the dominance of Parameter-Efficient Fine-Tuning (PEFT) methods. Low-Rank Adaptation (LoRA) is a foundational PEFT technique that freezes the pretrained model weights and injects small, trainable low-rank matrices, dramatically reducing the number of trainable parameters and memory requirements [16]. Building on this, Quantized LoRA (QLoRA) introduced a breakthrough by quantizing the base model to an aggressive 4-bit precision during fine-tuning. This, combined with innovations such as the NormalFloat4 (NF4) data type and paged optimizers, enables the fine-tuning of very large models (e.g., 65B) on a single consumer GPU while maintaining the performance of 16-bit fine-tuning [17]. The PEFT landscape continues to evolve with more advanced techniques such as DoRA, which dynamically allocates rank during training [23], further enhancing the toolkit for creating specialized models.

2.2 Post-Training Quantization for Efficient Inference

While QLoRA uses quantization during training, Post-Training Quantization (PTQ) is a critical step for optimizing models for efficient deployment. The primary goals of PTQ are to reduce the model's memory footprint, decrease inference latency, and lower energy consumption [10, 24]. Numerous PTQ methods have been developed, with GPTQ and AWQ being among the most prominent. GPTQ is a one-shot weight quantization method that uses approximate second-order information to iteratively quantize weights while compensating for errors, achieving high accuracy even at 3 or 4-bit precision [18]. In contrast, Activation-aware Weight Quantization (AWQ) identifies that a small fraction of weights are disproportionately important for performance. It protects these salient weights by scaling them up before uniform quantization, a simple yet highly effective strategy [19]. Other methods such as SmoothQuant tackle the problem by mathematically migrating quantization difficulty from activations, which often have problematic outliers, to the more easily quantized weights [25].

However, the benefits of quantization are not a "free lunch." Recent comprehensive benchmarks reveal a complex landscape of trade-offs. The effectiveness of a given PTQ method is highly dependent on the model size, the target task, and the underlying hardware [20, 26, 22]. For instance, smaller models often suffer greater accuracy degradation from 4-bit quantization than larger ones [20]. Furthermore, quantization can disproportionately affect certain capabilities, such as mathematical reasoning [27], and may even alter a model's truthfulness under certain conditions [28]. A systematic characterization of LLM quantization shows that choices regarding tensor parallelism and GPU architecture can dramatically alter the realized gains in latency and energy, underscoring the need for hardware-aware evaluation [21].

2.3 Synthetic Data and Structured Output Generation

The success of fine-tuning is fundamentally dependent on the quality and volume of the training data. When domain-specific data is scarce, synthetic data generation using LLMs has become an indispensable technique [29]. The Self-Instruct methodology first demonstrated that an LLM could bootstrap its own instruction-following dataset from a small seed set [30]. This concept has evolved into sophisticated, agentic pipelines such as AgentInstruct [31] and MetaSynth [32], which use multi-LLM systems to generate, refine, and diversify synthetic data at a massive scale. The quality of the data generator is key, as different LLMs exhibit complementary strengths in this role [33].

A specific challenge within this domain, and one central to our work, is generating structured outputs such as JSON. This capability is critical for integrating LLMs into automated workflows and RAG systems. Recent research has focused on benchmarking this capability [34, 35] and developing techniques to enforce strict schema adherence, such as reinforcement learning strategies [36]. Studies comparing fine-tuned small models against prompted large models for generating structured low-code workflows have found that targeted fine-tuning often yields more robust and domain-aware results [37]. However, it is also known that imposing strict format restrictions can sometimes impair a model's underlying reasoning ability, suggesting a trade-off between structural rigidity and performance [38]. Our work

builds on these insights by using an advanced LLM to generate a structured, noisy, multilingual dataset tailored for a real-world e-commerce task, providing the foundation for our specialization experiments.

3 Methodology

To systematically evaluate the performance trade-offs of optimizing small language models, we designed a multi-stage experimental pipeline. The process encompasses task definition, synthetic dataset generation, model selection, fine-tuning, post-training quantization, and a dual-pronged evaluation of both accuracy and operational performance.

3.1 Task Definition and Dataset Generation

The core task of our experiment is structured intent extraction from natural language user queries in an e-commerce context. The model's objective is to parse a user's free-form text request for managing a shopping cart and output a structured JSON object containing three key fields: 'action' (e.g., "add" or "remove"), 'product' (the canonical name of the item), and 'quantity' (an integer).

Given the lack of publicly available, multilingual datasets for this specific task, we generated a high-quality synthetic dataset. We employed a "metaprompting" strategy, using the GPT-4.1 model as a sophisticated data generator. A Python script systematically constructed detailed prompts that guided the generator model to produce diverse and realistic examples.

For each example, the script programmatically defined the ground-truth parameters: the language (cycling through English, Croatian, and Spanish), the action (either 'add' or 'remove'), a product randomly selected from a predefined catalog, and a quantity chosen via a weighted random selection to mimic realistic purchasing patterns. To introduce stylistic diversity, a wide range of prompt templates was employed, generating user expressions that varied from formal requests to brief imperative commands.

To further enhance realism and robustness, the script strategically injected noise. This included linguistic noise such as typos or slang like "pls" and "thx", contextual noise such as greetings, emojis, or unrelated brand names, and instances of code-switching, for example embedding English phrases like "free shipping" into Croatian sentences.

This pipeline produced a dataset of 3,000 examples. An example of a data point is shown below:

```
User input: "Can you delet 12 lip balms for me?"
->
JSON: {"action": "remove", "product": "Lip Balm", "quantity": 12}
```

The dataset, named `jtlicardo/ecommerce-intent-3k`, was published on the Hugging Face Hub¹ and split into a 90% training set and a 10% validation set.

3.2 Models and Baselines

Our investigation centered on optimizing a small, open-weight model. The primary model for fine-tuning and quantization was **Llama 3.2 1B**, a recent and efficient model from Meta [2]. To contextualize its performance, we also evaluated a representative set of other open-weight models, including **Gemma 3 1B** and **Gemma 2 2B** from Google [3], and **Qwen 2.5 1.5B** and **Qwen 2.5 3B** from Alibaba [4].

To establish a state-of-the-art performance benchmark, we used leading commercial models from OpenAI's **GPT-4.1 series**. These proprietary models serve as an upper bound for accuracy on the given task and were evaluated using a few-shot prompting approach, where several examples were provided in the prompt to guide the model's response format.

3.3 Experimental Pipeline

The core of our experiment involved a two-stage process: specializing the base model for our task via fine-tuning, and then optimizing the specialized model for efficient deployment via quantization.

¹<https://huggingface.co/datasets/jtlicardo/ecommerce-intent-3k>

3.3.1 Fine-tuning

We employed **QLoRA (Quantized Low-Rank Adaptation)** [17], a parameter-efficient fine-tuning technique, to train our model. The process was orchestrated using the ‘transformers’, ‘peft’, and ‘bitsandbytes’ libraries. The base Llama 3.2 1B model was loaded with its weights quantized to 4-bit precision using the NormalFloat4 (NF4) data type. A LoRA adapter was then applied with the following key hyperparameters: rank (r) of 8, alpha (`lora_alpha`) of 16, and a dropout rate of 0.1. The adapter targeted all linear projections within the self-attention and feed-forward network blocks. The model was trained for 5 epochs using the ‘SFTTrainer’ with a batch size of 8 and a learning rate of 2e-5. Crucially, the loss was calculated only on the completion part of the sequence (the JSON output), focusing the learning process exclusively on generating the correct structured data.

3.3.2 Post-Training Quantization

After fine-tuning, the trained LoRA adapter was merged with the original, full-precision (FP16) Llama 3.2 1B base model. This step created a single, unified, specialized model in FP16 format, which served as the foundation for all subsequent quantization.

1. **GPTQ for GPU:** We used the ‘auto-gptq’ library to apply **GPTQ (Generative Pre-trained Transformer Quantization)** [18]. Using a calibration set of 300 random samples from our training data, we quantized the merged FP16 model to 4-bit precision, creating a version optimized for GPU inference.
2. **GGUF for CPU:** We used the ‘llama.cpp’ toolchain to convert the merged FP16 model into the GGUF format, which is highly optimized for CPU inference. We generated three distinct versions to analyze the impact of bit depth: an aggressive 3-bit version (Q3_K_M), a balanced 4-bit version (Q4_K_M), and a high-quality 5-bit version (Q5_K_M).

3.4 Evaluation Framework

To ensure a rigorous and fair comparison, we used a separate, unseen test set of 100 examples, [jtlicardo/ecommerce-intent-eval](https://huggingface.co/datasets/jtlicardo/ecommerce-intent-eval)², structured identically to the training data.

3.4.1 Accuracy Metric

Performance was measured using **Exact Match Accuracy**. For a prediction to be considered correct, the generated output had to be a syntactically valid JSON object, and all three key-value pairs (‘action’, ‘product’, and ‘quantity’) had to be identical to the ground truth labels. This strict metric was chosen because partial correctness in this e-commerce task can lead to significant functional errors in a production system.

3.4.2 Performance Metrics and Hardware

Beyond accuracy, we profiled the key operational characteristics of our fine-tuned Llama 3.2 1B model in its FP16 and quantized forms. We measured inference speed, reported in tokens generated per second, memory consumption, expressed in gigabytes (GB) of VRAM for GPU tests and RAM for CPU tests, and energy consumption, where power draw in Watts was monitored using `nvidia-smi` during GPU tests to calculate energy efficiency. GPU-based evaluations (FP16 vs. GPTQ) were conducted on an **NVIDIA T4** GPU within the Google Colab environment, while CPU-based evaluations (GGUF variants) were performed on a local machine equipped with an **AMD Ryzen 7 5800HS** processor.

4 Results

This section presents the empirical findings of our study. We first compare the accuracy of all evaluated models to establish a performance baseline and validate our primary hypothesis. We then provide a detailed analysis of the operational characteristics of our optimized model, examining the hardware-dependent trade-offs of quantization on both GPU and CPU platforms.

²<https://huggingface.co/datasets/jtlicardo/ecommerce-intent-eval>

4.1 Model Accuracy Comparison

The primary goal of our experiment was to determine if a small, specialized open-weight model could achieve accuracy parity with large-scale commercial models. The exact match accuracy scores for all models on the unseen test set are presented in Table 1.

Table 1: Comparison of Exact Match Accuracy Across All Evaluated Models.

Model Family	Model Variant	Accuracy
<i>Commercial Baselines (Few-shot)</i>		
GPT (OpenAI)	GPT 4.1	1.00
	GPT 4.1-mini	1.00
	GPT 4.1-nano	0.99
<i>Open-Weight Models (Base and Fine-tuned)</i>		
Qwen (Alibaba)	Qwen 2.5 1.5B	0.86
	Qwen 2.5 1.5B Instruct	0.87
	Qwen 2.5 3B	0.96
	Qwen 2.5 3B Instruct	0.96
Gemma (Google)	Gemma 3 1B	0.78
	Gemma 3 1B (finetune)	0.90
	Gemma 2 2B	0.86
Llama (Meta)	Llama 3.2 1B	0.82
	Llama 3.2 1B (finetune)	0.99
	Llama 3.2 1B (finetune, GPTQ 4-bit)	0.99
	Llama 3.2 1B (finetune, GGUF 3-bit)	0.60
	Llama 3.2 1B (finetune, GGUF 4-bit)	0.89
	Llama 3.2 1B (finetune, GGUF 5-bit)	0.99

As expected, the commercial GPT-4.1 series models set a high bar, achieving near-perfect scores. The base open-weight models showed varied but generally respectable performance, with the Llama 3.2 1B base model achieving an accuracy of 0.82. The most significant finding is the dramatic performance increase after fine-tuning. The Llama 3.2 1B model, after fine-tuning with QLoRA, saw its accuracy jump to 0.99, an improvement of 21%. This result places the specialized 1B parameter model on equal footing with the much larger GPT-4.1-nano and effectively matches the performance of the top-tier commercial models. This outcome provides strong support for our central hypothesis. The overall performance landscape is visualized in Figure 1.

4.2 The Impact of Quantization on Accuracy

A key aspect of our investigation was to determine how post-training quantization affects the accuracy of the specialized model. The results, shown in Table 1, reveal that modern quantization techniques can preserve performance with high fidelity. Both the 4-bit GPTQ version and the 5-bit GGUF (Q5_K_M) version fully retained the 0.99 accuracy of the FP16 fine-tuned model.

However, more aggressive quantization introduces a clear trade-off. The 4-bit GGUF (Q4_K_M) version experienced a noticeable drop in accuracy to 0.89. The most aggressive 3-bit GGUF (Q3_K_M) variant suffered a catastrophic performance collapse, with accuracy plummeting to 0.60. This demonstrates the existence of a "quantization cliff," [39] a threshold below which the model loses the numerical precision required to execute its specialized task effectively.

4.3 GPU Performance Profile (FP16 vs. GPTQ)

We profiled the FP16 fine-tuned model against its 4-bit GPTQ version on an NVIDIA T4 GPU to assess performance in a typical accelerated environment. The results reveal a nuanced and counterintuitive relationship between quantization, memory, and speed on this specific hardware architecture.

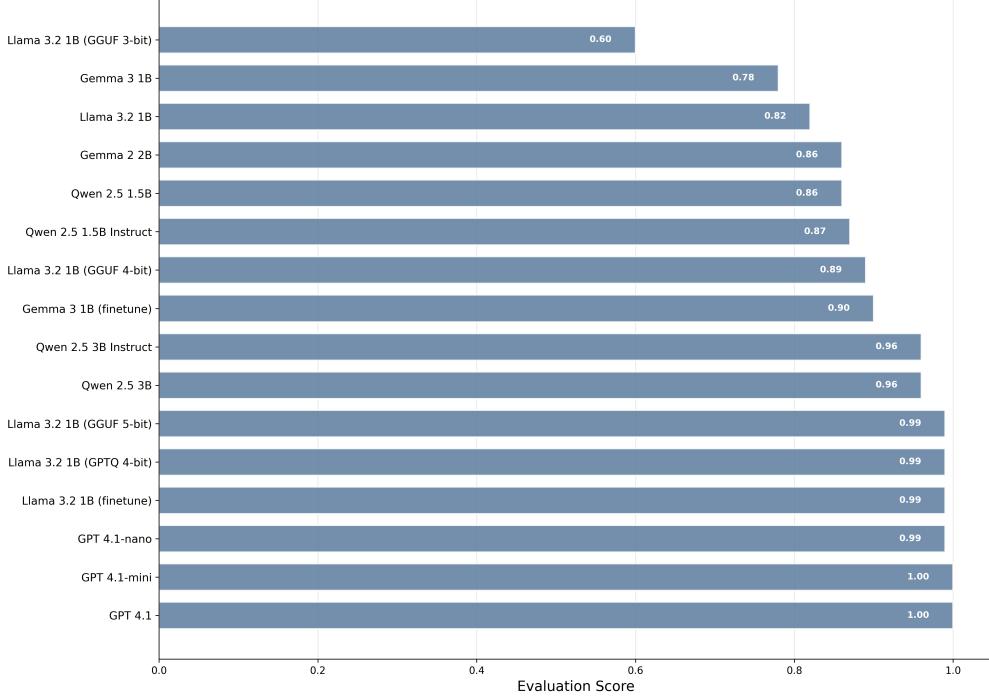


Figure 1: Overall Model Performance Comparison. The fine-tuned Llama 3.2 1B model and its high-quality quantized variants achieve accuracy scores comparable to the top commercial baselines.

As shown in Figure 2, quantization delivered substantial memory savings. The 4-bit GPTQ model reduced total VRAM consumption from 3.27 GB to 1.93 GB (a 41% reduction) and cut the model’s parameter-only footprint from 2.30 GB to just 0.96 GB.

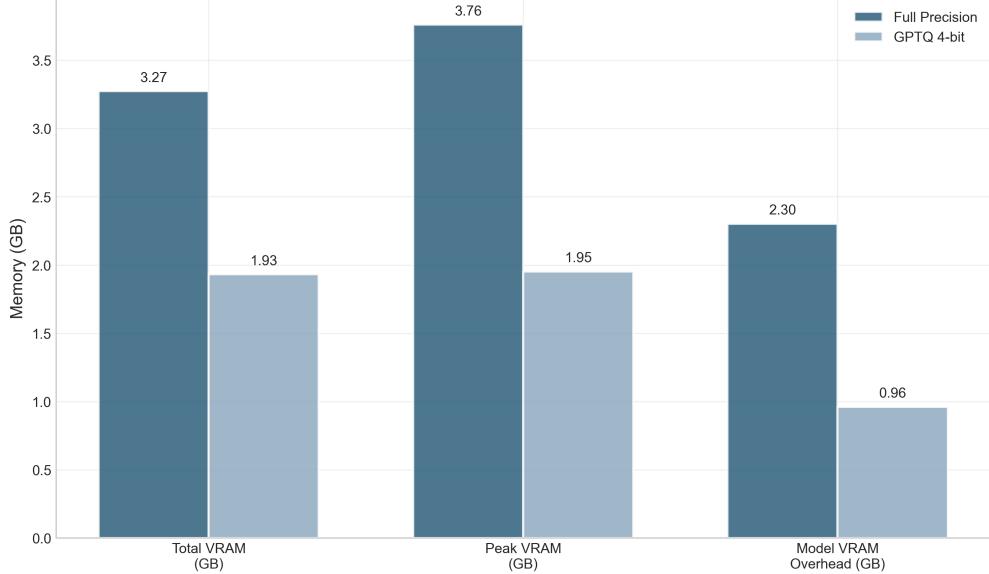


Figure 2: GPU Memory Comparison. GPTQ 4-bit quantization significantly reduces total VRAM, peak VRAM, and the memory required for model parameters alone.

Despite these memory benefits, the GPTQ model was significantly slower during inference, as detailed in Figures 3 and 4. While the model load time decreased dramatically by 93.4% (from 16.95 s to 1.12 s), the inference speed dropped

from 44.56 tokens/second for the FP16 model to just 7.92 tokens/second for the GPTQ version, an 82.2% slowdown. The slowdown suggests that the T4 GPU did not execute computations directly in 4-bit. Instead, it appears that the quantized weights were converted back to higher precision during inference, adding overhead. Consequently, the energy consumed per token was 489.3% higher for the quantized model, making it less efficient for inference on this architecture.

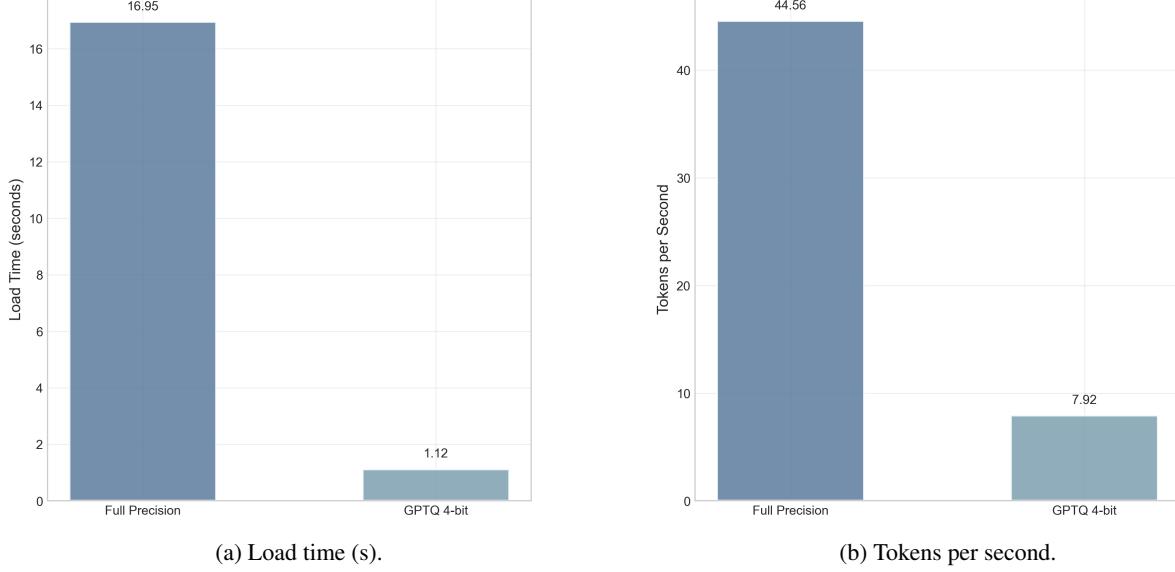


Figure 3: GPU speed profile on NVIDIA T4. GPTQ loads much faster but runs with lower throughput.

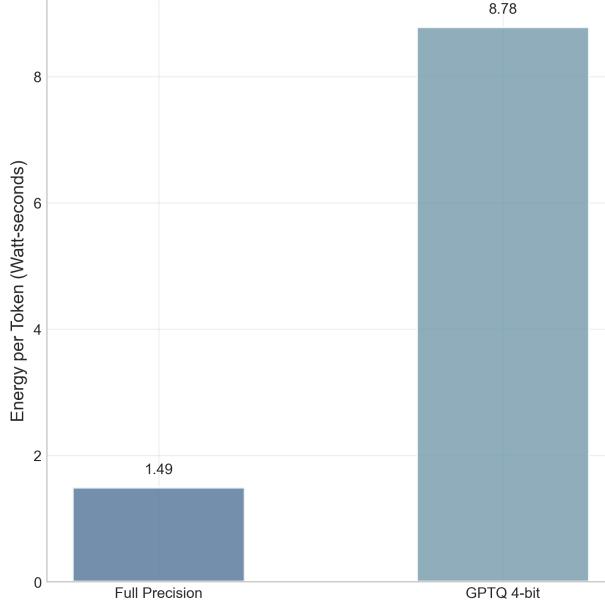


Figure 4: Energy per token on NVIDIA T4. GPTQ consumes substantially more energy per generated token.

4.4 CPU Performance Profile (GGUF)

In stark contrast to the GPU results, the GGUF-quantized models demonstrated exceptional performance gains on a CPU (AMD Ryzen 7 5800HS), leveraging the highly optimized ‘llama.cpp’ library.

As shown in Figure 5, the FP16 model was impractically slow, achieving only 2.6 tokens/second. In contrast, all quantized GGUF versions provided a massive speedup. The 4-bit (Q4_K_M) version was the fastest, peaking at 47.9 tokens/second, representing an 18x improvement compared to the FP16 baseline.

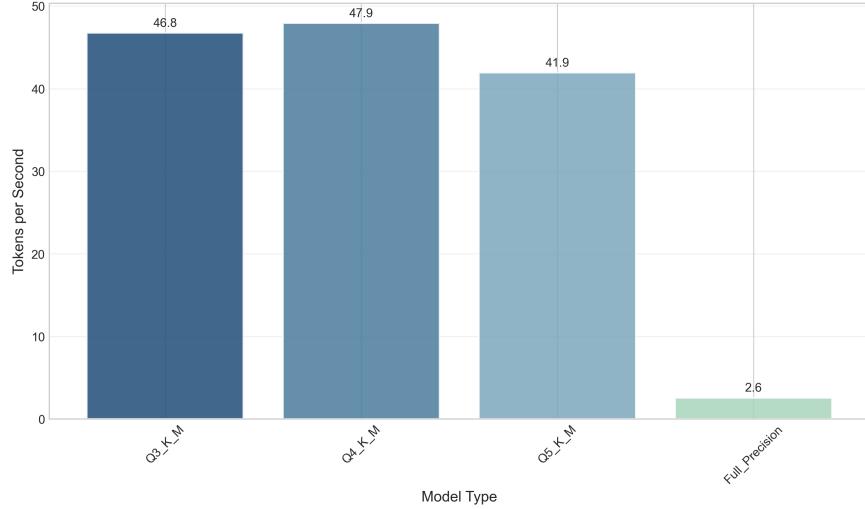


Figure 5: CPU Inference Performance. All GGUF-quantized models offer a dramatic speedup over the full-precision baseline.

These speed advantages were complemented by dramatic reductions in RAM consumption and load times. Figure 6 shows that the RAM footprint was reduced from 14.39 GB for the FP16 model to around 1.15-1.51 GB for the GGUF versions - a reduction of over 90%. This makes it feasible to run the model on standard consumer hardware. Consequently, load times also saw a significant improvement, as shown in Figure 7.

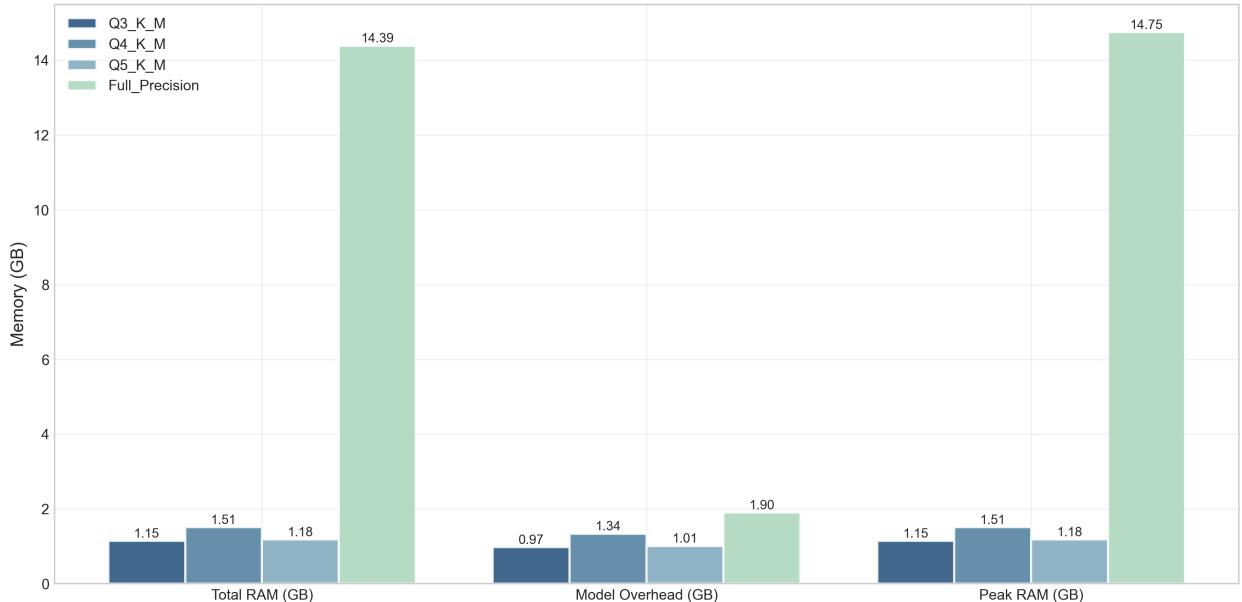


Figure 6: CPU Memory Usage Comparison. GGUF quantization reduces RAM consumption by over 90%, from 14.39 GB to less than 1.6 GB.

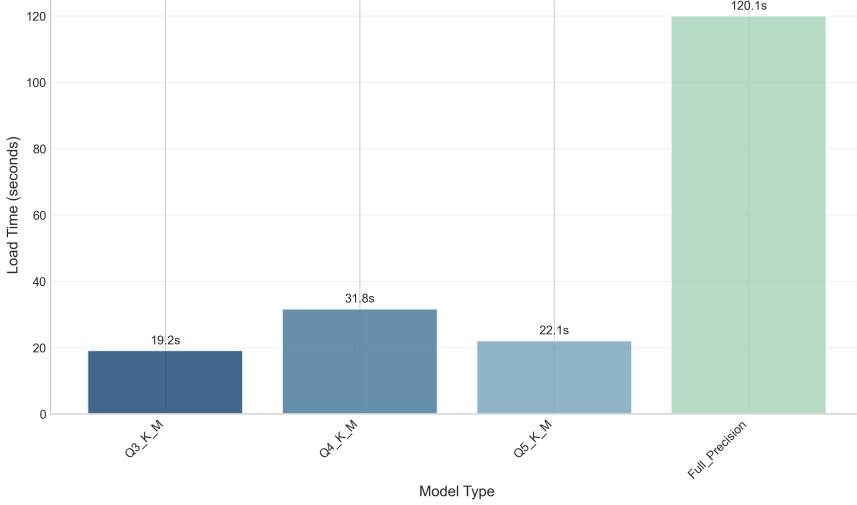


Figure 7: CPU Load Time Comparison. Quantized GGUF models load significantly faster than the full-precision version.

4.5 Accuracy vs. Efficiency Trade-offs on CPU

The CPU performance results allow for a clear visualization of the trade-offs between accuracy, memory, and speed. Figures 8 and 9 plot the Pareto frontiers for the GGUF variants. These charts illustrate the optimal choices depending on application priorities.

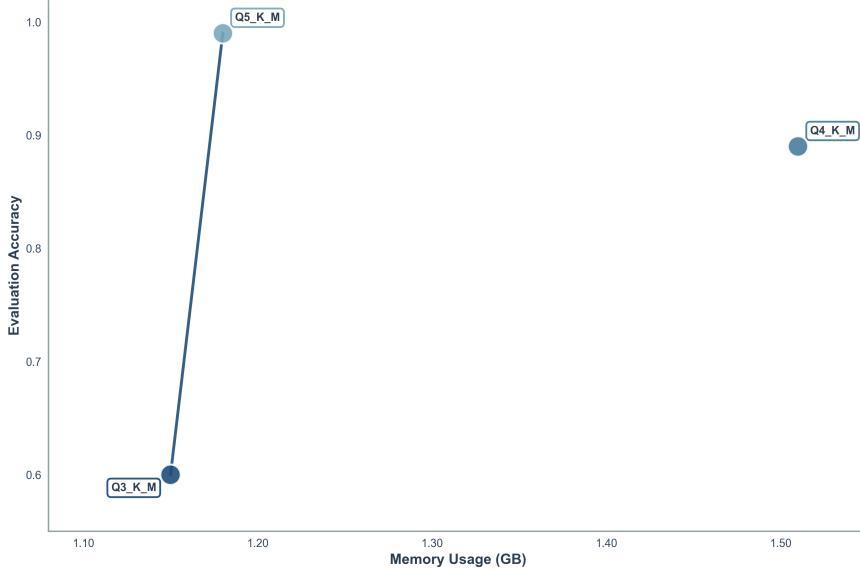


Figure 8: Pareto Frontier: Accuracy vs. Memory Usage (CPU). The Q5_K_M model offers the highest accuracy with minimal memory usage, while the Q3_K_M model is suboptimal due to its low accuracy.

The analysis reveals two optimal candidates for CPU deployment. The **Q5_K_M (5-bit)** model provides the highest possible accuracy (0.99) with excellent speed (around 42 tokens/s) and minimal memory usage. The **Q4_K_M (4-bit)** model offers the peak inference speed (nearly 48 tokens/s) at the cost of a moderate reduction in accuracy (0.89). The 3-bit version is clearly suboptimal, as it provides no significant memory or speed advantage over the other quantized versions but suffers a severe accuracy penalty.

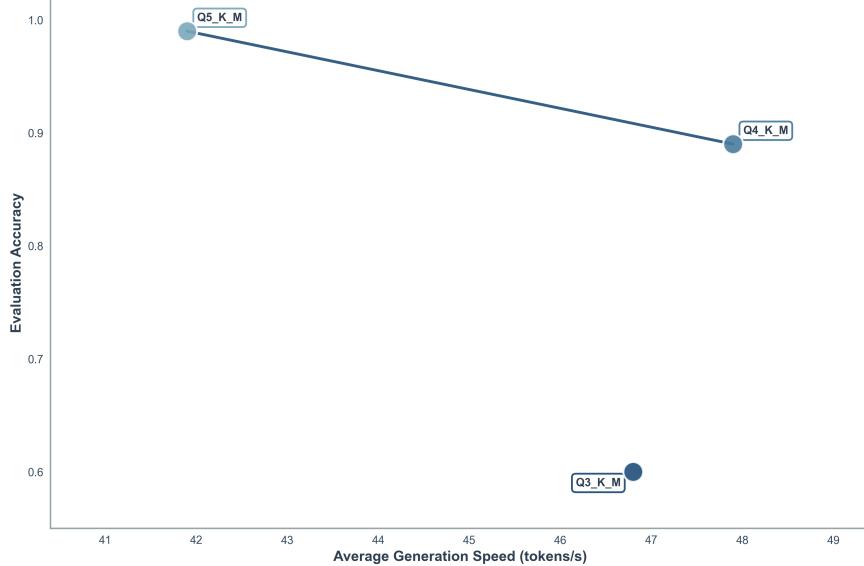


Figure 9: Pareto Frontier: Accuracy vs. Generation Speed (CPU). The Q4_K_M model provides the highest speed, while the Q5_K_M model offers the best accuracy, presenting a clear trade-off.

5 Discussion

The results of our study offer a multi-faceted view of the opportunities and challenges associated with deploying small, optimized language models. Our findings confirm that specialization can elevate a small model to state-of-the-art performance; however, they also reveal that the efficiency gains from optimization techniques such as quantization are deeply intertwined with the underlying hardware and software ecosystem.

5.1 Small Models Can Achieve State-of-the-Art Accuracy

Our central finding - that a 1B parameter Llama 3.2 model can match the 99% accuracy of GPT-4.1 after specialized fine-tuning - contributes to a growing body of evidence challenging the notion that larger models are constantly superior. This result underscores the power of domain specialization. While massive models possess a broad, generalist knowledge base, a smaller model trained on a high-quality, task-specific dataset can develop a deep, expert-level competency within its narrow domain. This principle has been demonstrated across diverse fields, from outperforming GPT-4 on arithmetic tasks [12] to achieving comparable performance in pedagogical applications [14] and medical language understanding [40]. The comprehensive study in LoRA Land, which found that over 200 fine-tuned small models surpassed GPT-4 on specific tasks, solidifies this paradigm [15]. For many business applications, particularly in e-commerce where user interactions follow predictable patterns [8], deploying a fleet of small, expert models is a more cost-effective, private, and computationally efficient strategy than relying on a single, oversized generalist model.

5.2 Quantization is Not a Free Lunch: The Hardware-Software Synergy

Perhaps the most critical insight from our performance analysis is that the benefits of quantization are not intrinsic to the algorithm itself but emerge from the synergy between the model format, the inference engine, and the hardware architecture. The stark contrast between our GPU and CPU results illustrates this point perfectly.

The 82% slowdown of the 4-bit GPTQ model on the NVIDIA T4 GPU, despite a 41% reduction in VRAM, highlights a common pitfall. The slowdown suggests that the T4 GPU did not execute computations directly in 4-bit. Instead, it appears that the quantized weights were converted back to higher precision during inference, adding overhead that negated the benefits of reduced memory bandwidth. This finding aligns with systematic characterizations of LLM quantization, which show that performance and energy gains are highly dependent on the interplay between the quantization scheme and the GPU's capabilities [21, 10]. On modern GPUs designed with native support for low-precision arithmetic, these results would likely be reversed, leading to significant speedups [26].

Conversely, the massive success of the GGUF formats on the CPU (achieving over 18x speedup) is a testament to software optimization. The llama.cpp library is purpose-built to leverage CPU vector instructions (such as AVX) for

highly efficient low-bit integer matrix multiplications. This demonstrates that with the right software, even ubiquitous consumer hardware can become a powerful platform for LLM inference. This principle of hardware-software co-design is a central theme in efficient AI deployment, from large-scale data centers dynamically managing resources [41] to hyper-efficient accelerators for edge devices [42, 43]. Our results provide a clear, practical example of this principle in action.

5.3 Practical Recommendations and the Pareto Frontier

Our analysis of the accuracy-vs-efficiency trade-offs, particularly for CPU deployment (Figures 8 and 9), allows us to formulate concrete recommendations for practitioners. The choice of which model variant to deploy is not a simple one but a strategic decision based on application-specific priorities.

- **For Maximum Accuracy:** In scenarios where correctness is paramount, the 5-bit GGUF (Q5_K_M) model emerges as the optimal choice for CPU deployment. It delivers the full 99% accuracy of the original fine-tuned model while still providing significant benefits in RAM reduction and a substantial inference speedup. For GPU deployment, using the FP16 model or a GPTQ-quantized version on modern hardware with native low-bit support would be the recommended path.
- **For Maximum Speed:** When lowest latency is the primary goal and a slight dip in accuracy is acceptable, the 4-bit GGUF (Q4_K_M) model is the superior option, offering the highest throughput on our CPU testbed.
- **Suboptimal Choices:** The 3-bit GGUF model is clearly a suboptimal choice. It is "dominated" on the Pareto frontier, offering no significant performance advantage over the 4-bit and 5-bit versions while suffering a catastrophic loss of accuracy. This finding aligns with broader studies showing that performance degradation at very low bit-depths can be non-linear and severe [22].

This decision-making process, balancing multiple objectives such as accuracy, latency, and energy, resonates with calls to move beyond single metrics and adopt more holistic evaluation frameworks, such as considering energy-per-token [44, 24].

5.4 Limitations of the Study

While our findings provide valuable insights, it is important to acknowledge the limitations of this study, which also point to avenues for future research. First, our model was trained and evaluated exclusively on **synthetically generated data**. While carefully designed with strategic noise injection, synthetic data may not fully capture the complexity and unpredictability of real-world user queries. The challenges of ensuring realism and avoiding distributional mismatch in synthetic data are well-documented [29]. Second, our **hardware scope** was limited to a single older-generation GPU (NVIDIA T4) and one type of consumer CPU. As discussed, performance results for GPTQ are expected to be significantly different and more favorable on newer GPUs. Third, our investigation focused on a **single, well-defined task**. The model's excellent performance in intent extraction does not guarantee similar success on other, more complex e-commerce tasks, such as those benchmarked in ShoppingBench, which require multi-step reasoning and tool use [45]. Finally, our use of a strict **Exact Match Accuracy** metric, while appropriate for this task, is binary and does not capture nuances of partially correct answers.

6 Conclusion

This paper investigated the feasibility of using small, optimized open-weight language models as a practical alternative to large commercial systems for a specialized e-commerce task. Our findings demonstrate that this approach is not only viable but highly effective, offering a pathway to building state-of-the-art AI solutions that are both powerful and resource-efficient.

We have successfully shown that through parameter-efficient fine-tuning on a high-quality synthetic dataset, a 1 billion parameter Llama 3.2 model can achieve 99% exact match accuracy in a structured intent recognition task. This result places its performance on par with the significantly larger, state-of-the-art GPT-4.1 model, confirming our primary hypothesis. However, our analysis of post-training quantization revealed a critical layer of complexity: the operational benefits are profoundly dependent on the deployment environment. On an older GPU architecture, 4-bit GPTQ quantization, while drastically reducing memory, led to a counter-intuitive 82% slowdown in inference due to dequantization overhead. In stark contrast, GGUF formats on a CPU achieved over an 18x improvement in inference throughput and approximately 90% lower RAM usage compared to the FP16 baseline, making sophisticated LLM inference feasible on consumer-grade hardware.

The primary conclusion of this work is that the future of many applied AI solutions may not lie solely with ever-larger generalist models, but in a diverse ecosystem of smaller, highly specialized, and hardware-aware models. For developers and organizations, our results provide a clear directive: the choice of an optimization strategy cannot be made in isolation from the target hardware. A model that is highly efficient in one context can be impractical in another. This paradigm enables organizations to develop customized solutions that are more cost-effective, private, and computationally efficient.

Building on these findings, several avenues for future research emerge. First, validating the fine-tuned model's performance on a large corpus of real-world, anonymized user data would provide the definitive confirmation of its practical utility. Second, replicating the GPU performance analysis on modern architectures (e.g., NVIDIA Ampere or Hopper) is essential to quantify the potential inference speedups that GPTQ can offer when paired with native low-precision hardware support. Finally, the scope of this work could be expanded by fine-tuning the model to handle a broader range of e-commerce tasks, such as product recommendation or order status inquiries, and by exploring alternative optimization techniques, such as AWQ quantization or other PEFT methods like DoRA.

Ultimately, this research serves as a practical demonstration that with the right specialization and optimization strategies, small models can indeed stand shoulder-to-shoulder with giants, offering a more accessible and sustainable path for the widespread adoption of advanced AI.

References

- [1] Ashish Vaswani et al. *Attention Is All You Need*. Aug. 2023. DOI: 10.48550/arXiv.1706.03762. arXiv: 1706.03762 [cs].
- [2] Aaron Grattafiori et al. *The Llama 3 Herd of Models*. Nov. 2024. DOI: 10.48550/arXiv.2407.21783. arXiv: 2407.21783 [cs].
- [3] Gemma Team et al. *Gemma 3 Technical Report*. Mar. 2025. DOI: 10.48550/arXiv.2503.19786. arXiv: 2503.19786 [cs].
- [4] An Yang et al. *Qwen3 Technical Report*. May 2025. DOI: 10.48550/arXiv.2505.09388. arXiv: 2505.09388 [cs].
- [5] Justina Sidlauskiene, Yannick Joye, and Vilte Auruskeviciene. "AI-based Chatbots in Conversational Commerce and Their Effects on Product and Price Perceptions". In: *Electronic Markets* 33 (May 2023). DOI: 10.1007/s12525-023-00633-8.
- [6] Sina Gholamian et al. *LLM-Based Robust Product Classification in Commerce and Compliance*. Oct. 2024. DOI: 10.48550/arXiv.2408.05874. arXiv: 2408.05874 [cs].
- [7] Wendi Zhou et al. *A Usage-centric Take on Intent Understanding in E-Commerce*. Feb. 2024. DOI: 10.48550/arXiv.2402.14901. arXiv: 2402.14901 [cs].
- [8] Christian Herold et al. *LiLiuM: eBay's Large Language Models for e-Commerce*. June 2024. DOI: 10.48550/arXiv.2406.12023. arXiv: 2406.12023 [cs].
- [9] Nidhal Jegham et al. *How Hungry Is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference*. May 2025. DOI: 10.48550/arXiv.2505.09598. arXiv: 2505.09598 [cs].
- [10] Jared Fernandez et al. *Energy Considerations of Large Language Model Inference and Efficiency Optimizations*. Apr. 2025. DOI: 10.48550/arXiv.2504.17674. arXiv: 2504.17674 [cs].
- [11] Erik Johannes Husom et al. *The Price of Prompting: Profiling Energy Use in Large Language Models Inference*. July 2024. DOI: 10.48550/arXiv.2407.16893. arXiv: 2407.16893 [cs].
- [12] Tiedong Liu and Bryan Kian Hsiang Low. *Goat: Fine-tuned LLaMA Outperforms GPT-4 on Arithmetic Tasks*. May 2023. DOI: 10.48550/arXiv.2305.14201. arXiv: 2305.14201 [cs].
- [13] Hong Jia et al. *Beyond Scale: Small Language Models Are Comparable to GPT-4 in Mental Health Understanding*. July 2025. DOI: 10.48550/arXiv.2507.08031. arXiv: 2507.08031 [cs].
- [14] Lorenzo Lee Solano et al. *Narrowing the Gap: Supervised Fine-Tuning of Open-Source LLMs as a Viable Alternative to Proprietary Models for Pedagogical Tools*. July 2025. DOI: 10.48550/arXiv.2507.05305. arXiv: 2507.05305 [cs].
- [15] Justin Zhao et al. *LoRA Land: 310 Fine-tuned LLMs That Rival GPT-4, A Technical Report*. Apr. 2024. DOI: 10.48550/arXiv.2405.00732. arXiv: 2405.00732 [cs].
- [16] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. Oct. 2021. DOI: 10.48550/arXiv.2106.09685. arXiv: 2106.09685 [cs].
- [17] Tim Dettmers et al. *QLoRA: Efficient Finetuning of Quantized LLMs*. May 2023. DOI: 10.48550/arXiv.2305.14314. arXiv: 2305.14314 [cs].

- [18] Elias Frantar et al. *GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers*. Mar. 2023. DOI: 10.48550/arXiv.2210.17323. arXiv: 2210.17323 [cs].
- [19] Ji Lin et al. *AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration*. July 2024. DOI: 10.48550/arXiv.2306.00978. arXiv: 2306.00978 [cs].
- [20] Jemin Lee et al. *Exploring the Trade-Offs: Quantization Methods, Task Difficulty, and Model Size in Large Language Models From Edge to Giant*. June 2025. DOI: 10.48550/arXiv.2409.11055. arXiv: 2409.11055 [cs].
- [21] Tianyao Shi and Yi Ding. *Systematic Characterization of LLM Quantization: A Performance, Energy, and Quality Perspective*. Aug. 2025. DOI: 10.48550/arXiv.2508.16712. arXiv: 2508.16712 [cs].
- [22] Jiaqi Zhao et al. *Benchmarking Post-Training Quantization in LLMs: Comprehensive Taxonomy, Unified Evaluation, and Comparative Analysis*. May 2025. DOI: 10.48550/arXiv.2502.13178. arXiv: 2502.13178 [cs].
- [23] Yulong Mao et al. *DoRA: Enhancing Parameter-Efficient Fine-Tuning with Dynamic Rank Distribution*. June 2024. DOI: 10.48550/arXiv.2405.17357. arXiv: 2405.17357 [cs].
- [24] Paul Joe Maliakel, Shashikant Ilager, and Ivona Brandic. *Investigating Energy Efficiency and Performance Trade-offs in LLM Inference Across Tasks and DVFS Settings*. Sept. 2025. DOI: 10.48550/arXiv.2501.08219. arXiv: 2501.08219 [cs].
- [25] Guangxuan Xiao et al. *SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models*. Mar. 2024. DOI: 10.48550/arXiv.2211.10438. arXiv: 2211.10438 [cs].
- [26] Jemin Lee et al. *A Comprehensive Evaluation of Quantized Instruction-Tuned Large Language Models: An Experimental Analysis up to 405B*. Sept. 2024. DOI: 10.48550/arXiv.2409.11055. arXiv: 2409.11055 [cs].
- [27] Zhen Li et al. *Quantization Meets Reasoning: Exploring LLM Low-Bit Quantization Degradation for Mathematical Reasoning*. Feb. 2025. DOI: 10.48550/arXiv.2501.03035. arXiv: 2501.03035 [cs].
- [28] Yao Fu et al. *Quantized but Deceptive? A Multi-Dimensional Truthfulness Evaluation of Quantized LLMs*. Aug. 2025. DOI: 10.48550/arXiv.2508.19432. arXiv: 2508.19432 [cs].
- [29] Mihai Nadas, Laura Diosan, and Andreea Tomescu. “Synthetic Data Generation Using Large Language Models: Advances in Text and Code”. In: *IEEE Access* 13 (2025), pp. 134615–134633. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2025.3589503. arXiv: 2503.14023 [cs].
- [30] Yizhong Wang et al. *Self-Instruct: Aligning Language Models with Self-Generated Instructions*. May 2023. DOI: 10.48550/arXiv.2212.10560. arXiv: 2212.10560 [cs].
- [31] Arindam Mitra et al. *AgentInstruct: Toward Generative Teaching with Agentic Flows*. July 2024. DOI: 10.48550/arXiv.2407.03502. arXiv: 2407.03502 [cs].
- [32] Haris Riaz et al. *MetaSynth: Meta-Prompting-Driven Agentic Scaffolds for Diverse Synthetic Data Generation*. June 2025. DOI: 10.48550/arXiv.2504.12563. arXiv: 2504.12563 [cs].
- [33] Seungone Kim et al. “Evaluating Language Models as Synthetic Data Generators”. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Wanxiang Che et al. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 6385–6403. ISBN: 979-8-89176-251-0. DOI: 10.18653/v1/2025.acl-long.320.
- [34] Jialin Yang et al. *StructEval: Benchmarking LLMs’ Capabilities to Generate Structural Outputs*. May 2025. DOI: 10.48550/arXiv.2505.20139. arXiv: 2505.20139 [cs].
- [35] Connor Shorten et al. *StructuredRAG: JSON Response Formatting with Large Language Models*. Aug. 2024. DOI: 10.48550/arXiv.2408.11061. arXiv: 2408.11061 [cs].
- [36] Bhavik Agarwal, Ishan Joshi, and Viktoria Rojkova. *Think Inside the JSON: Reinforcement Strategy for Strict LLM Schema Adherence*. Feb. 2025. DOI: 10.48550/arXiv.2502.14905. arXiv: 2502.14905 [cs].
- [37] Orlando Marquez Ayala et al. *Fine-Tune an SLM or Prompt an LLM? The Case of Generating Low-Code Workflows*. July 2025. DOI: 10.48550/arXiv.2505.24189. arXiv: 2505.24189 [cs].
- [38] Zhi Rui Tam et al. *Let Me Speak Freely? A Study on the Impact of Format Restrictions on Performance of Large Language Models*. Aug. 2024. DOI: 10.48550/arXiv.2408.02442. arXiv: 2408.02442 [cs].
- [39] Arash Ahmadian et al. “Intriguing Properties of Quantization at Scale”. In: *Advances in Neural Information Processing Systems* 36 (Dec. 15, 2023), pp. 34278–34294.
- [40] Yujuan Velvin Fu et al. *BioMistral-NLU: Towards More Generalizable Medical Language Understanding through Instruction Tuning*. Mar. 2025. DOI: 10.48550/arXiv.2410.18955. arXiv: 2410.18955 [cs].
- [41] Jovan Stojkovic et al. *DynamoLLM: Designing LLM Inference Clusters for Performance and Energy Efficiency*. 2024. arXiv: 2408.00741 [cs.AI].

- [42] Chunlin Tian et al. “CLONE: Customizing LLMs for Efficient Latency-Aware Inference at the Edge”. In: *Proceedings of the 2025 USENIX Annual Technical Conference (USENIX ATC 25)*. 2025.
- [43] Ye Qiao et al. *TELLMe: An Energy-Efficient Ternary LLM Accelerator for Prefilling and Decoding on Edge FPGAs*. Apr. 2025. DOI: 10.48550/arXiv.2504.16266. arXiv: 2504.16266 [cs].
- [44] Patrick Wilhelm, Thorsten Wittkopp, and Odej Kao. “Beyond Test-Time Compute Strategies: Advocating Energy-per-Token in LLM Inference”. In: *Proceedings of the 5th Workshop on Machine Learning and Systems (EuroMLSys '25)*. 2025. DOI: 10.1145/3721146.3721953.
- [45] Jiangyuan Wang et al. *ShoppingBench: A Real-World Intent-Grounded Shopping Benchmark for LLM-based Agents*. Aug. 2025. DOI: 10.48550/arXiv.2508.04266. arXiv: 2508.04266 [cs].