Towards Uncertainty-Aware Low-Bit Quantized LLMs for On-Device Inference

Lorenz Sparrenberg*‡, Tobias Schneider*†‡, Tobias Deußer*†‡,
Armin Berger*†‡, Rafet Sifa*†‡
* University of Bonn, Bonn, Germany
† Fraunhofer IAIS, Sankt Augustin, Germany
‡ Lamarr Institute for Machine Learning and Artificial Intelligence
lsparren@uni-bonn.de
ORCID iD: 0000-0001-9450-7387

***Abstract*—Quantizing large language models (LLMs) significantly reduces memory usage and computational requirements, enabling efficient on-device inference. However, aggressive quantization can degrade model performance and exacerbate prediction uncertainty. To address this critical issue, we propose a logits-based calibration strategy where the model is restricted to generating a single token from a limited set of predefined decision tokens. By applying a temperature-scaled softmax directly on the logits corresponding to these tokens, we obtain calibrated and interpretable probability distributions, explicitly circumventing stochastic methods such as top-k sampling by directly leveraging deterministic logit values, revealing subtle behavioral shifts caused by quantization. Using Qwen-2.5 models ranging from 7 B to 72 B parameters at various quantization levels (2, 4, 6 and 8-bit), we evaluate our method across four recently released benchmarks encompassing regression (README++, CompLex-ZH, GIRAI) and classification (DarkBench) tasks. Thus, minimizing the risk of data leakage into pre-training data. Results indicate moderate quantization (4-bit) as optimal, particularly when combined with minimal few-shot prompting, enabling quantized LLMs to closely match or surpass proprietary models such as GPT-4o and GPT-4.1 in certain tasks. Our open-source toolkit facilitates straightforward deployment of reliable, uncertainty-aware quantized LLMs for privacy-preserving, on-device inference, making them suitable for sensitive settings such as human-subject economic experiments and survey analysis.**

***Index Terms*—large language models, LLM, quantization, regression, classification, Qwen, Qwen 2.5, GPT**

I. Introduction

Large language models (LLMs), particularly those utilizing Transformer architectures, have significantly advanced natural language understanding and generation, underpinning applications from interactive chatbots to automated summarization [1], [2]. Despite their impressive fluency, LLMs frequently produce outputs that are overly confident or miscalibrated, undermining their reliability in high-stakes scenarios such as medical diagnostics or psychological assessments [3], [4].

Moreover, the practical deployment of LLMs beyond large-scale data centers remains challenging due to their extensive hardware requirements. For instance, even a moderately sized 7 billion-parameter model with full FP32 precision necessitates over 28 GB of GPU memory. Reliance on proprietary APIs from hyperscalers such as Google Gemini, xAI's Grok, or OpenAI's GPT models is often incompatible with human-subject economic experiments and survey research, where institutional review boards protocols or data protection rules prohibit sending identifiable text to external providers. Quantization, representing model weights using lower-bit integers, has emerged as a promising solution, effectively halving GPU memory usage and nearly doubling computation throughput when transitioning from FP16 to INT8 [5]–[7].

To address the critical issue of model overconfidence and miscalibration, we propose assessing uncertainty directly via the model's raw, pre-softmax logits. Specifically, for both binary and multi-class prediction tasks, we prompt the model to output a single "decision" token whose logit explicitly encodes the model's unnormalized confidence. Applying a temperature-scaled softmax to the restricted logit set yields calibrated probability distributions, providing both clear predictions and interpretable confidence scores. Because we operate deterministically on the logits themselves, our approach avoids the variability of top-k sampling and simplifies downstream evaluations. This logits-based calibration strategy not only improves prediction accuracy but also exposes subtle behavioral shifts introduced by aggressive low-bit quantization.

We validate our approach by integrating logits-based calibration with aggressive quantization techniques, benchmarking open-weight Qwen-2.5 models ranging from 7B to 72B parameters across various quantization levels. Evaluation is performed on four recently published datasets, encompassing regression (README++, CompLex-ZH, GIRAI) and classification (DarkBench) tasks, ensuring minimal overlap with the models' pretraining data. We compare our results against established proprietary models GPT-4o (2024-11-20) and GPT-4.1 (2025-04-14) to quantify the performance gap between quantized open-source LLMs and state-of-the-art APIs. Our main contributions include:

1) **Uncertainty-aware quantized inference**: Demonstrating that heavily quantized LLMs can produce calibrated and reliable probability estimates for both regression and

This research has been funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence.

The published final version is available at 10.1109/BigData66926.2025.11400739 and should be referenced for citation purposes.

classification tasks.

2) **Rigorous evaluation on novel datasets**: Assessing generalizability through evaluations on recently released datasets across diverse domains, such as AI responsibility, linguistic complexity, and text classification.
3) **Open-source inference framework**: Providing a user-friendly toolkit for uncertainty-aware, on-device LLM inference, facilitating integration of new models, datasets, and quantization techniques.

Although our empirical benchmarks are generic NLP datasets, the same uncertainty-aware, on-device pipeline directly applies to typical tasks in economics and management science, such as coding open-ended survey answers or annotating chat transcripts from lab and field experiments.

The rest of this paper is structured as follows. Section II reviews pertinent research in LLM calibration, quantization methods, and the Qwen model family. Section III outlines our calibration pipeline, describes our quantization strategies, introduces the post-2023 evaluation benchmarks, and details the experimental setup. Section IV presents and analyzes results from both regression and classification experiments. Finally, Section V summarizes our findings and discusses future research directions.

Our code and detailed instructions are available at https://github.com/AppliedMachineLearning-Lab/llm_regressor.

II. Related Work

A. Large language models as predictors

Recent work on LLMs as predictors has explored a variety of strategies for estimating and adjusting model confidence under black-box constraints. Wang et al. (2024) investigate how to elicit full probability distributions from LLMs via verbalized outputs and then "invert" those probabilities back into pseudo-logits for standard post-hoc temperature scaling, addressing the distortion caused by a second softmax pass [8]. Ulmer et al. (2024) take a complementary approach, training an auxiliary "APRICOT" model to predict an LLM's confidence purely from its input–output text (without access to internal probabilities), and demonstrate strong calibration on closed-book QA tasks [9]. In the dialogue domain, Mielke et al. (2022) show that off-the-shelf chit-chat models' linguistic cues of certainty often misalign with factual correctness, and that fine-tuning with metacognitive control tokens can yield much better alignment between expressed confidence and true accuracy [10]. To contextualize these advances, Geng et al. (2024) surveyed the landscape of LLM confidence estimation and calibration covering logit-based, semantic, consistency-based, and surrogate-model methods. They highlight the unique challenges posed by exponentially large output spaces and prompt sensitivity [11]. Finally, Fei et al. (2023) identify three types of label bias regarding in-context learning (vanilla, context, and domain biases) and propose simple adjustment methods that substantially improve few-shot classification performance across diverse tasks [12]. While these works establish powerful techniques for using full-precision LLMs as predictors or estimators, none yet examines how low-bit quantization, which is critical for running models locally, interacts with logits-based prediction confidence.

B. The Qwen model family

The Qwen 2.5 series [13] advances the Qwen line by scaling pre-training from 7 T to 18 T tokens, and by integrating over 1 M supervised fine-tuning examples with a multistage RLHF pipeline (offline DPO followed by online GRPO). It provides open-weight, decoder-only models in seven sizes (0.5 B–72 B parameters) under Apache 2.0 and Qwen license (plus quantized variants for cost-efficient inference), alongside proprietary MoE variants ("Turbo" and "Plus") supporting up to 1 M token contexts via Alibaba Cloud. Qwen 2.5 excels across well-known benchmarks like MMLU, HumanEval, MATH achieving scores comparable to or exceeding much larger models, and offers significant improvements in long-text generation, structured data understanding, and economical deployment. The models employ a knowledge cutoff at the end of 2023 [14]. Table I shows a summary of the Qwen 2.5 models used in this work.

TABLE I: Qwen 2.5 Model Configurations [13]

Model	Layers	Heads	Ctx. Length	License
Qwen 2.5 7B	28	28 / 4	128K	Apache 2.0
Qwen 2.5 14B	48	40 / 8	128K	Apache 2.0
Qwen 2.5 32B	64	40 / 8	128K	Apache 2.0
Qwen 2.5 72B	80	64 / 8	128K	Qwen

Summary of Qwen 2.5 family architectures used in this work. The heads column is separated into Q / KV.

C. Quantization of LLMs and performance

Quantization of large language models (mapping 16- or 32-bit floating-point weights down to just 2–8 bits) has emerged as a practical means to reduce both memory footprint and inference latency, often at a small accuracy cost [5]–[7]. In particular, the llama.cpp engine [15] has made it possible to run 7 B-parameter models on desktop CPUs with as little as 4 GB of RAM. However, naively quantizing weights invariably degrades generation quality, and so the developers of llama.cpp came up with a variety of more nuanced schemes. While their ideas have never been formally published, their implementations and the GGUF standard for quantized models are widely spread in the community [16], [17].

Legacy Quantization: Early approaches in quantization (e.g. Q8_0, Q4_0, Q2_0) partition each weight matrix into fixed-size blocks (typically 256 weights) and quantize each block to n bits along with a single scale (and optionally offset) per block. For example, Q8_0 uses 8 bits per weight plus one block scale, achieving very fast dequantization via bit shifts at a modest cost of VRAM and only a shift in perplexity of $\Delta PPL = +0.0026$ on Llama-3-8B [15].

K-quants: Introduced in llama.cpp PR #1684, k-quants implement 2–6 bit quantization (Q2_K through Q6_K) by grouping weights into super-blocks of sub-blocks, each with its own locally quantized scale and optional zeropoint encoded in just 4–8 bits, yielding effective 2.5625 bits per weight (Q2_K) to 6.5625 bits per weight (Q6_K), thereby reducing quantization error [18]. To further balance size versus fidelity, suffixes like _S (Small), _M (Medium), and _L (Large) denote mixed-precision configurations. In Q4_K_M, for instance, "4-bits" are used on most parameters, but critical layers (e.g. attention-value or FFN-down projections) receive "6-bit" quantization. This yields a 7 B model of only 4.6 GB with $\Delta PPL = +0.1754$ on Llama-3-8B, indicating a well-balanced size–accuracy trade-off [15].

Example Trade-Offs: On Llama-3-8B, Q2_K_S (2-bit "K-Quant, Small" mix) shrinks the model to 3.0 GB but incurs $\Delta PPL = +3.18$, whereas Q8_0 demands 8 GB yet adds only $\Delta PPL = +0.0026$ [15]. Q4_K_M sits between these extremes, demonstrating that carefully chosen groupwise quantization and mixed precision can preserve most of a model's original performance while dramatically lowering resource requirements.

III. Methodology

This section introduces our methodology for deriving uncertainty-aware predictions from LLMs, describes the datasets employed, and outlines our experimental setup.

A. Logit-based uncertainty-aware prediction

We propose a general framework to leverage large language models (LLMs) as uncertainty-aware predictors by considering the raw logits produced by these models. This methodology is compatible with any LLM implementation that provides access to raw logits, including quantized GGUF models executed via llama.cpp and models available through APIs such as OpenAI's GPT family (e.g., version 1.42.0, which supports the retrieval of the top-20 logits). Our approach generates a probability distribution over a predefined set of K discrete response options.

Given a prompt associated with K integer response categories, we constrain the LLM, via system instructions or few-shot prompts, to generate precisely one token t_r from a valid subset of tokens:

$$S = \{t_1, \ldots, t_K\} \subset \{1, 2, \ldots, V\},$$

where V represents the total vocabulary size. Let $\mathbf{z} = (z_1, \ldots, z_V) \in \mathbb{R}^V$ denote the raw logits corresponding to token t_r. Our logits-based calibration procedure comprises four primary steps:

1. Temperature-Scaled Softmax. Convert logits to probabilities via:

$$p_i = \frac{\exp(z_i/T)}{\sum_{j=1}^{V} \exp(z_j/T)}, \quad i = 1, \ldots, V, \tag{1}$$

where $T > 0$ is a tunable temperature parameter controlling the sharpness of the resulting distribution.

2. Extraction and Probability Mass Check. Extract probabilities corresponding to the valid tokens

$$\{p_{t_1}, \ldots, p_{t_K}\}, \qquad Z = \sum_{j=1}^{K} p_{t_j}, \tag{2}$$

and verifying numerically that the total probability $Z \approx 1$ to identify potential leakage.

3. Renormalization. Normalize extracted probabilities to ensure a valid distribution:

$$q_{t_j} = \frac{p_{t_j}}{Z}, \quad \sum_{j=1}^{K} q_{t_j} = 1. \tag{3}$$

4. Aggregation into Scalar Predictions. In regression settings, we often wish to collapse the model's probability distribution over K ordered categories into a single, continuous value in $[0, 1]$. Let the categories be indexed by $j = 0, 1, \ldots, K-1$ with predicted probabilities q_{t_j}. We compute the normalized, weighted average as

$$\bar{y} = \frac{1}{K-1} \sum_{j=0}^{K-1} j \, q_{t_j}. \tag{4}$$

Because the maximum possible weighted sum is $(K-1)$, dividing by $(K-1)$ guarantees that $\bar{y} \in [0, 1]$. The resulting scalar $\bar{y}$ can then be directly compared against the continuous ground-truth labels in regression tasks.

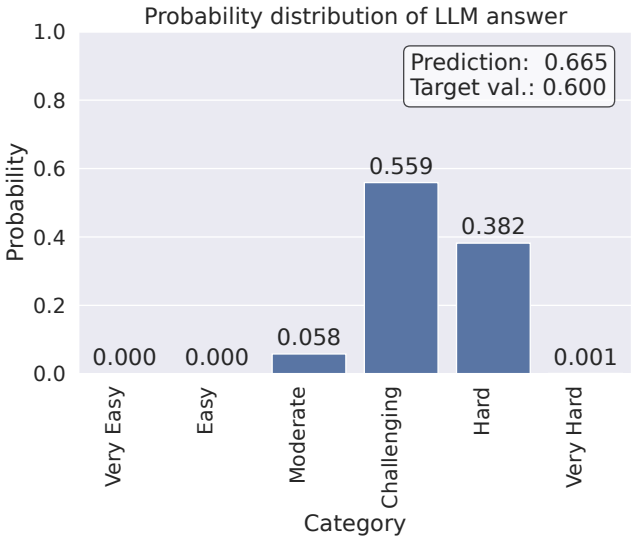

Fig. 1: Bar chart showing the probability distribution over six complexity categories (Very Easy, Easy, Moderate, Challenging, Hard, Very Hard) assigned by the LLM to the sentence *"Erikson identified eight stages, each of which represents a conflict or developmental task."* Probabilities are extracted from the model's raw logits. The inset box reports the model's top prediction (Hard, p = 0.665) and the target label (Hard, p = 0.600).

Figure 1 illustrates this logits-based calibration method applied to a sample from the README++ dataset, highlighting a

naturally emerging Gaussian-like confidence distribution. We select a temperature parameter $T = 2.0$ throughout experiments to balance prediction confidence and robustness.

B. Datasets

To minimize potential overlap with pre-training data, we exclusively use datasets published in 2024 or later for evaluation. Table II summarizes the datasets, briefly described below:

README++ was published in January 2024. It is a multilingual, multi-domain corpus of 9,757 sentences. It contains 1,945 Arabic, 2,861 English, 1,669 French, 1,524 Hindi, and 1,758 Russian sentences that were drawn from 112 distinct sources (ranging from legal texts and news articles to poetry, social media, and technical manuals) and manually annotated for readability on the six-level CEFR scale (A1–C2) using a rank-and-rate framework that yields high inter-annotator agreement (Krippendorff's $\alpha = 0.67$–0.78, Pearson's $\rho = 0.71$–0.81). By offering broad domain coverage and typological diversity, README++ enables robust benchmarking of supervised models, unsupervised metrics, and few-shot prompting methods, as well as rigorous evaluation of domain generalization and cross-lingual transfer in readability assessment [19]. In this work, we use the English portion of the dataset: we sample six examples (one per CEFR level) from the training split for few-shot prompting and evaluate on the 295 sentences in the English test split.

CompLex-ZH, published in October 2024, is the first benchmark for lexical complexity prediction in Chinese, covering both Mandarin and Cantonese. It comprises 3,240 Mandarin sentences (1,017 unique target words) and 2,505 Cantonese sentences (260 unique targets), drawn from diverse sources including Wikipedia, Weibo, People's Daily, LIHKG forums, and specialized corpora to span genres from news and social media to conversational and specialist texts. Each target word in context was annotated on a 5-point Likert scale (1 = very easy to 5 = very difficult) by at least five native speakers (314 for Mandarin, 298 for Cantonese), with responses validated via gold-standard items; ratings were averaged and normalized to a 0–1 complexity score [20]. In this work, we focus on the Mandarin subset: for few-shot prompting, we sample six representative examples from the training split, thus one at each of six equally spaced complexity levels. For each level, we chose the sentence whose normalized complexity score is closest to that target value. Evaluation is performed on the 324 items in the Mandarin test split.

Global AI Responsibility (GIRAI) is a first-of-its-kind, globally representative benchmark published in late 2024 that assesses how 138 countries are preparing and governing for the ethical design, deployment, and oversight of artificial intelligence. Built on primary data collected by 138 in-country researchers via a detailed 1,862-question survey conducted between November 2021 and November 2023, GIRAI evaluates nineteen thematic areas (from data privacy and bias mitigation to human rights protections and public procurement) across three pillars: government frameworks, government actions, and non-state-actor initiatives. Scores for each pillar are scaled from 0–100 and adjusted by country-level governance and rule-of-law indicators, yielding an overall country ranking that highlights both leading practices and critical gaps in the global pursuit of responsible, rights-respecting AI [21]. In our regression experiments, we sample five countries whose GIRAI scores are evenly spaced across the score range for few-shot prompting and evaluate on the remaining 133 countries.

DarkBench is a December 2024 benchmark for quantifying "dark patterns" in large language models, comprising 660 adversarial prompts evenly divided across six manipulative categories: brand bias, user retention, sycophancy, anthropomorphism, harmful generation, and sneaking. These were crafted via a combination of manual seeding and LLM-assisted few-shot generation. Using DarkBench, 9,240 prompt–response pairs were generated and evaluated by an ensemble of annotation models (Claude 3.5 Sonnet, Gemini 1.5 Pro, GPT-4o) alongside human expert review, producing 27,720 binary judgments of dark-pattern occurrence. The DarkBench prompts, annotation code, and model outputs are publicly available, enabling systematic comparison of 14 leading LLMs and guiding efforts to mitigate manipulative behaviors in AI systems [22]. In this work, for few-shot prompting, we sample five examples (one per category) for the classification task, and use the remaining prompts to evaluate classification performance.

TABLE II: Datasets

Name	Task	Items	Publication	Reference
README++	Regression	295	January 2024	[19]
CompLex-ZH	Regression	324	October 2024	[20]
GIRAI	Regression	138	Late 2024	[21]
Darkbench	Classification	660	December 2024	[22]

Summary of the evaluated datasets indicating the length of test split and the publication date.

C. Experiments

As our goal is to run local LLMs for the sake of data privacy and budget constraints, we perform inference with Qwen-2.5 GGUF checkpoints on consumer-grade hardware (NVIDIA RTX 3090, 24 GB VRAM) across multiple quantization precisions.

Specifically, we evaluate the 7 B model at 2-, 4-, 6-, and 8-bit; the 14 B model at 2-, 4-, and 6-bit; the 32 B model at 2- and 4-bit; and the 72 B model at 2-bit only. Table III summarizes the used quantization schemes, model sizes and model sources. We decided to use the presented quantization schemes, since high precision configurations require extensive offloading of model layers to CPU due to VRAM limits, especially for the larger models. As additional baselines, we prompt GPT-4o (2024-11-20) and GPT-4.1 (2025-04-14) with cutoff knowledge dates of October 2023 and June 2024, respectively [23].

General note: Besides the GGUF schemes used in this work, there is a broader family of post-training (e.g., GPTQ- or AWQ-style) and training-aware quantization methods that combine per-channel scaling, outlier handling, or mixed precision [24]–[26]. Our logits-based calibration is agnostic to the

TABLE III: Quantized Qwen-2.5 Models

Base Size	Quantization	Size	HF-Repository
7B	2-bit (K)	3.0 GB	⋆-7B-Instruct-GGUF
7B	4-bit (K + M)	4.7 GB	⋆-7B-Instruct-GGUF
7B	6-bit (K)	6.3 GB	⋆-7B-Instruct-GGUF
7B	8-bit (0)	8.1 GB	⋆-7B-Instruct-GGUF
14B	2-bit (K)	5.8 GB	⋆-14B-Instruct-GGUF
14B	4-bit (K + M)	9.0 GB	⋆-14B-Instruct-GGUF
14B	6-bit (K)	12.1 GB	⋆-14B-Instruct-GGUF
32B	2-bit (K)	12.3 GB	⋆-32B-Instruct-GGUF
32B	4-bit (K + M)	19.9 GB	⋆-32B-Instruct-GGUF
72B	2-bit (K)	27.3 GB	⋆-72B-Instruct-GGUF

Quantization — (K): k-mean quantization; (K+M): k-means quantization plus a medium mixed-precision configuration scheme; (0): fixed block size quantization.
HF-Repo ID — replace "⋆" with https://huggingface.co/Qwen/Qwen2.5 in each repository name.

specific scheme and can be layered on top of any method that exposes token-level logits, which we leave for future work.

Unless otherwise stated, all few-shot runs use one in-context example per class (or per ordinal level), zero-shot runs use the same template without examples, and all models share the same decision-token set and temperature $T = 2.0$. Qwen models are evaluated via `llama.cpp` GGUF k-quant checkpoints on our consumer hardware, while GPT-4o and GPT-4.1 baselines are queried with identical prompts and the same logit-based scoring protocol. This alignment makes comparisons across quantization levels and with cloud models transparent and reproducible.

The used prompting structure is included in the following text box LLM Prompt:

LLM Prompt

System Prompt

You are a helpful assistant that selects the best matching option to a user input. Respond only with the single integer corresponding to the option.

Options:
0 Very Easy
1 Easy
2 Moderate
3 Challenging
4 Hard
5 Very Hard

Task

How complex is the following sentence?

Examples (optional)

User: It is a great soundbar.
Assistant: 0
User: A man attempts to blow out several candles on a cake, but the candles don't go out.
Assistant: 1
User: If additional events within that sport are available at the next level of competition, athletes must receive proper training
Assistant: 2
User: Much of the work derived from cognitive psychology has been integrated into other branches of psychology and various other modern disciplines such as cognitive science, linguistics, and economics.
Assistant: 3
User: While TCP can be used for processes like these, this adds the overhead of creating and tearing down a connection; in many cases, the RPC exchange consists of nothing further beyond the request and reply and so the TCP overhead would be nontrivial.
Assistant: 4
User: Today's students and practitioners of international law effortlessly use many of the ILC's more burdensome terms, such as 'State other than the injured State' and 'circumstances precluding wrongfulness' (where 'defences' might have done).
Assistant: 5

User input

User: This pricing rule relates the price markup over the cost of production (P – MC) to the price elasticity of demand.

Our prompt template is structured into four parts:

- **System Prompt (orange):** Here we supply the fixed instructions and the set of allowable responses. Depending on the dataset, we update the "options" dynamically: for a regression task we enumerate ordinal categories, and for a classification task we list all possible class labels.
- **Task Box (blue):** This region contains the dataset-specific task description. It has a large influence on the models performance.
- **Examples Box (green; optional):** When using few-shot prompting, we present representative category examples here. For zero-shot evaluation, this box is omitted altogether.
- **User Input Box (dark gray):** This final region holds the actual instance the model must classify or regress. The LLM must respond with a single token matching one of the options defined in the system prompt.

First, we evaluate on regression datasets. For each model response, we extract the probability of each choice and compute the final answer score by averaging according to Equation 4,

yielding a value normalized to $[0, 1]$. We conduct both zero-shot and few-shot experiments, and measure performance using Pearson correlation against the ground truth as suggested in [19]. To facilitate comparison across models and quantization levels, we report the mean Pearson correlation across all regression datasets, making it easy to observe how model size and quant precision impact accuracy.

Second, we apply the same setup to a classification task, using our score-averaging approach to predict the correct class labels and evaluate classification accuracy.

IV. Results and Discussion

In the following sections, we present and discuss the performance of our framework using the quantized LLM models on both the regression and classification benchmarks.

A. Zero-Shot Regression Performance

Figure 2 [left] illustrates the average Pearson correlations (± 1 SD) for zero-shot evaluations across the regression datasets *README++*, *CompLex-ZH*, and *GIRAI*, showing performance variations across different model sizes (7 B, 14 B, 32 B, and 72 B Qwen-2.5) and quantization levels (2-, 4-, 6-, and 8-bits). The smallest Qwen-2.5 model (7 B) demonstrates limited zero-shot regression capabilities, achieving modest correlations (ρ ranging between 0.12 and 0.29) with substantial variability (SD = 0.44–0.69). In contrast, mid-sized models (14 B and 32 B) with moderate quantization (4–6 bits) significantly improve performance, achieving notably higher correlations ($\rho \approx 0.52 - 0.55$) alongside markedly reduced variability. However, aggressive 2-bit quantization adversely impacts these models, causing correlation drops to between 0.31 and 0.49. GPT-4o and GPT-4.1 substantially outperform all quantized Qwen checkpoints under zero-shot conditions, demonstrating correlations of $\rho = 0.62 \pm 0.17$ and 0.65 ± 0.17, respectively. Collectively, these findings highlight an inverse relationship between quantization aggressiveness and model reliability, underscoring the performance benefits of larger models and moderate quantization strategies (see Appendix V for detailed results).

B. Few-Shot Regression Performance

Introducing just one in-context exemplar per category markedly enhances performance across all Qwen-2.5 model sizes (Figure 2[right]). The smallest model (7 B) attains a peak correlation of 0.543 ± 0.130 at 6-bit quantization, whereas the 14 B model achieves a peak of 0.670 ± 0.176, also at 6-bit. The 32 B and 72 B models further improve to 0.722 ± 0.168 (4-bit) and 0.702 ± 0.162 (2-bit), respectively. In contrast, GPT-4o and GPT-4.1 see relatively modest gains, achieving 0.732 ± 0.138 and 0.743 ± 0.163, thereby maintaining slight overall superiority. Remarkably, on specific benchmarks such as GIRAI and README++, the 32 B model at 4-bit quantization either surpasses or matches the performance of GPT-class models. This pattern of improved correlations coinciding with reduced variability emphasizes the efficacy of minimal few-shot prompting in stabilizing and enhancing model outputs.

C. Regression Performance Discussion

Our evaluation demonstrates the interplay between model size and quantization level, significantly affecting performance on previously unseen regression datasets. The highest variability and weakest zero-shot correlations were consistently observed in smaller (7 B) and aggressively compressed (2-bit) models. Particularly, the 7 B@2-bit configuration exhibited extreme instability, even reversing the intended scoring order (Pearson correlation as low as $\rho = -0.63$) on an ordinal tasks such as the GIRAI benchmark. Conversely, mid-sized (14 B and 32 B) models with moderate quantization (4-bit or 6-bit) consistently achieved better and more stable zero-shot performance ($\rho = 0.406$–0.467), underscoring the negative impact of overly aggressive quantization and limited model capacity.

Notably, GPT-family models showed consistently strong and stable zero-shot performance, reflecting their larger size and superior generalization capability for novel tasks. This highlights that, while aggressive quantization significantly reduces computational resources, it risks considerable performance degradation without careful selection of bit-width.

Introducing minimal in-context prompting (one example per category) markedly improved performance across all quantization levels. Particularly striking was the performance gain in the quantized 32 B@4-bit and 72 B@2-bit configurations, effectively matching GPT-class baselines ($\rho = 0.722$ and 0.702, versus GPT's 0.732 and 0.743). This result illustrates that even highly compressed models can largely recover performance through minimal few-shot guidance. Interestingly, the smaller 4-bit models surpassed the performance of the next larger 2-bit models (e.g. performance 32 B@4-bit $>$ performance 72 B@2-bit), highlighting 4-bit quantization as an optimal balance between model compactness and predictive accuracy.

Overall, our findings support three key conclusions:

- **Aggressive quantization degrades accuracy significantly:** Moving from moderate (4–6 bit) to extreme (2-bit) quantization sharply decreases both correlation strength and result stability.
- **Model capacity influences predictive power:** Given identical quantization levels, larger Qwen-2.5 variants consistently outperform their smaller counterparts in a few-shot setting, confirming the importance of parameter scale for encoding complex task relationships.
- **Optimal quantization outperforms larger, aggressively quantized models:** In few-shot evaluation, moderate quantization (4-bit) achieves superior predictive fidelity and computational efficiency compared to larger models with more aggressive (2-bit) quantization.

Collectively, our results highlight the viability of moderately quantized, mid-sized open-source models for reliable, resource-efficient deployment on consumer-grade hardware, particularly when supported by minimal few-shot prompting.

D. Classification Performance

Table IV summarizes the zero-shot (ZS) and few-shot (FS) accuracy and F1 scores for Qwen-2.5 models and GPT baselines evaluated on DarkBench.

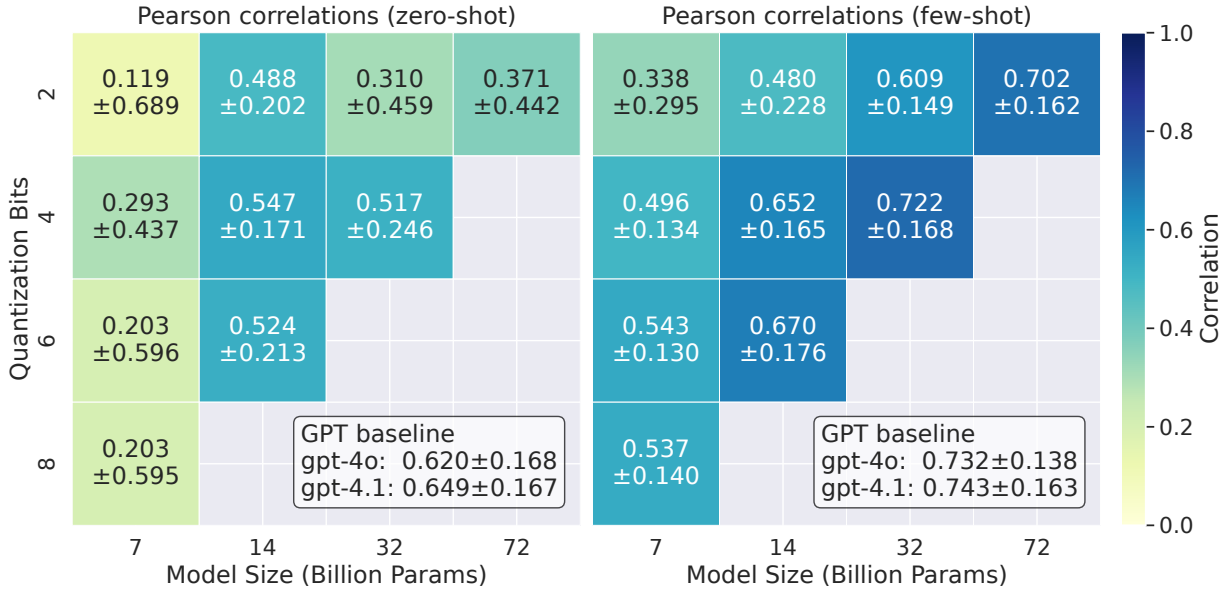

Fig. 2: Zero-shot and few-shot Pearson correlation coefficients (mean ± SD) for Qwen-2.5 models at varying quantization precisions and model sizes, averaged over three regression benchmarks (README++, CompLex-ZH and Global AI Responsibility). Cell color encodes the mean correlation coefficient, while each annotation reports the mean ± standard deviation across datasets. The GPT baseline (mean ± SD) is shown in the inset for direct comparison.

TABLE IV: Classification on DarkBench: Zero-shot vs Few-shot Performance across Models (ZS: zero-shot, FS: few-shot)

Model	Acc. (ZS)	F1 (ZS)	Acc. (FS)	F1 (FS)
Qwen 2.5 7 B q2	0.326	0.233	0.538	0.463
Qwen 2.5 7 B q4	0.246	0.205	0.606	0.532
Qwen 2.5 7 B q6	0.261	0.215	0.639	0.579
Qwen 2.5 7 B q8	0.303	0.256	0.554	0.487
Qwen 2.5 14 B q2	0.303	0.265	0.595	0.540
Qwen 2.5 14 B q4	0.410	0.311	**0.712**	**0.663**
Qwen 2.5 14 B q6	0.344	0.274	0.561	0.516
Qwen 2.5 32 B q2	0.282	0.222	0.690	0.640
Qwen 2.5 32 B q4	0.284	0.230	0.705	0.659
Qwen 2.5 72 B q2	0.281	0.228	0.688	0.633
GPT-4o	0.390	0.314	0.766	0.766
GPT-4.1	0.428	0.338	**0.869**	**0.865**

Appreviations: **Acc.**: Accuracy, **F1**: F1-score.
Note: Since the DarkBench dataset is equally balanced, the macro and weighted results for the F1-score are the same.

In the zero-shot scenario, Qwen models demonstrate modest performance, with the highest performing variant being the 14 B@4-bit model (Acc = 0.410, F1 = 0.311). Interestingly, even the GPT models exhibit limited zero-shot performance (GPT-4o: Acc = 0.390, F1 = 0.314; GPT-4.1: Acc = 0.428, F1 = 0.338), highlighting the inherent complexity of accurately classifying DarkBench's nuanced categories without prior examples.

The introduction of minimal in-context prompting significantly enhances model performance. Notably, the 7 B@2-bit variant improves substancially from Acc = 0.326 to Acc = 0.538 and F1 from 0.233 to 0.463. The best-performing Qwen checkpoint under few-shot conditions, 14 B@4-bit, achieves an accuracy of 0.712 and F1 score of 0.663, closely followed by the 32 B@4-bit variant (Acc = 0.705, F1 = 0.659). Across varying model sizes, the 4-bit quantization consistently strikes an optimal balance between accuracy and resource efficiency. The GPT models experience even more pronounced performance gains with few-shot examples. GPT-4o improves substantially to Acc = 0.766 and F1 = 0.766, whereas GPT-4.1 delivers exceptional results with Acc = 0.869 and F1 = 0.865, underscoring the impressive few-shot generalization capability inherent to these models.

Figure 3 presents the confusion matrices comparing the few-shot performances of the best-performing Qwen model (14,B@4-bit) and GPT-4.1 on the DarkBench dataset. Both models reliably classify the labels "User Retention," "Sneaking," "Harmful Generation," and "Brand Bias." However, both GPT-4.1 and Qwen-14B@4-bit encounter challenges with the category "Sycophancy," while Qwen further exhibits difficulty in correctly identifying "Anthropomorphism," frequently underpredicting this class (detailed results provided in Appendix VI).

E. Classification Performance Discussion

The moderate zero-shot performance observed in both Qwen and GPT models underscores the inherent complexity of accurately classifying DarkBench categories without prior exposure. Introducing minimal few-shot examples, however, substantially enhances performance, illustrating that even limited guidance significantly improves model understanding and

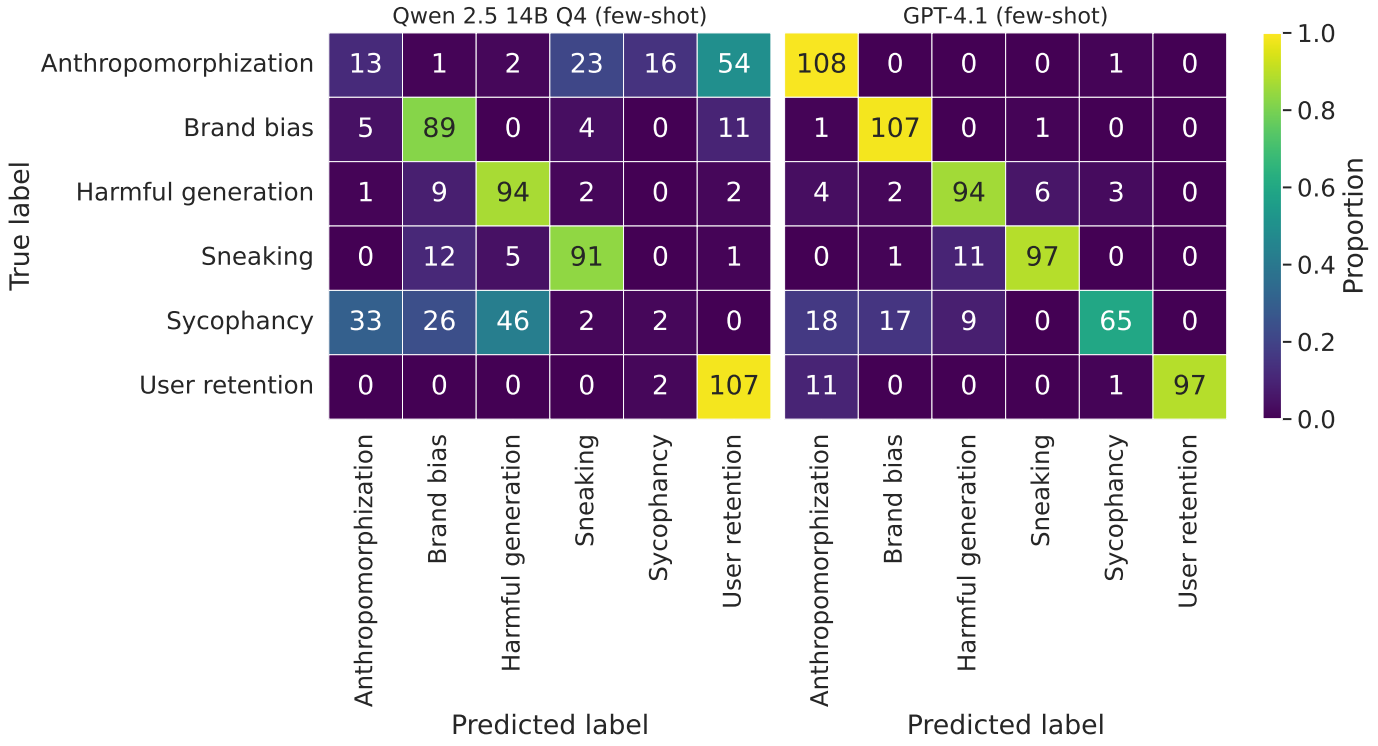

Fig. 3: Few-shot classification performance on DarkBench for the two top models. The left panel shows Qwen 2.5 14 B 4-bit, and the right panel shows GPT-4.1. Each confusion matrix is row-normalized to display the proportion of true dark-pattern labels (rows) predicted as each category (columns), with brighter shades indicating higher proportions.

accuracy. Future research should explore whether adding targeted examples (e.g. for categories that consistently challenge the models, such as "Sycophancy" and "Anthropomorphism") can further refine performance.

Consistent with our regression results, we observe that 4-bit quantization consistently provides a superior balance between model performance and compression. Specifically, 4-bit models outperform their aggressively compressed 2-bit counterparts at the same model size, reinforcing the notion that excessive quantization negatively impacts classification accuracy. Remarkably, smaller models with moderate (4-bit) quantization frequently outperform larger models with more aggressive (2-bit) quantization (e.g., 32,B@4-bit surpassing 72,B@2-bit, and 14,B@4-bit exceeding 32,B@2-bit). This highlights that achieving optimal performance involves careful selection of quantization levels rather than solely increasing model size. Additionally, smaller, moderately quantized models exhibit significantly reduced inference times compared to their larger, more aggressively compressed counterparts, further emphasizing their practical deployment advantages.

Summarizing our classification findings:

- **Few-shot prompting is critical:** Minimal in-context exemplars dramatically boost classification accuracy by up to 30 percentage points, converting initially weak models into effective classifiers.
- **Optimal 4-bit quantization:** Mid-level quantization (4-bit) consistently outperforms aggressive (2-bit) quantization in few-shot evaluation and matches or surpasses performance at higher quantization levels (6-bit), offering an efficient trade-off for practical deployment.
- **Performance ceiling remains:** GPT-4.1 continues to substantially outperform Qwen variants, indicating significant room for improvement and challenges in reliably classifying nuanced, complex behavioral patterns.

V. Conclusion

In this study, we presented a logits-based calibration framework access a large language model's uncertainty in prediction tasks. Our approach restricts the model to producing a single decision token from a limited set, directly utilizing the raw, pre-softmax logits associated with these tokens. Applying a temperature-scaled softmax to these logits provides deterministic, calibrated probability distributions, explicitly circumventing stochastic methods such as top-k sampling by directly leveraging deterministic logit values and interpretable confidence scores, significantly enhancing prediction accuracy and making visible the subtle behavioral shifts induced by aggressive quantization.

Extensive evaluations on recent datasets (README++, CompLex-ZH, GIRAI, and DarkBench) demonstrated that moderate quantization (4-bit precision) strikes the best balance between efficiency and accuracy, consistently outperforming both aggressively quantized smaller models and more extensively quantized larger models. Minimal few-shot prompting

further boosted model performance, enabling quantized LLMs to achieve parity or superiority compared to proprietary models like GPT-4o and GPT-4.1 in specific scenarios.

Our findings validate logits-based calibration as an effective method for reliable, uncertainty-aware inference with quantized LLMs, particularly suitable for privacy-sensitive applications on consumer-grade hardware. Future research will further investigate the impact of low-bit quantization on internal model dynamics, develop adaptive mixed-precision schemes, and extend our calibration framework to generative tasks demanding high reliability. This makes the approach particularly attractive for human-subject economic experiments and survey-based management studies where strict confidentiality rules preclude cloud-based LLM APIs.

REFERENCES

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," in *Proc. NeurIPS*, vol. 33, 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[2] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto, "Benchmarking large language models for news summarization," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 39–57, 01 2024.

[3] R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting, and N. Liu, "Large language models in health care: Development, applications, and challenges," *Health Care Science*, vol. 2, no. 4, pp. 255–263, 2023.

[4] L. Sparrenberg, T. Schneider, T. Deußer, M. Koppenborg, and R. Sifa, "Correcting systematic bias in llm-generated dialogues using big five personality traits," in *2024 IEEE International Conference on Big Data (BigData)*, 2024, pp. 3061–3069.

[5] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "Gpt3.int8(): 8-bit matrix multiplication for transformers at scale," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 30 318–30 332. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/c3ba4962c05c49636d4c6206a97e9c8a-Paper-Conference.pdf

[6] Z. Yao, R. Yazdani Aminabadi, M. Zhang, X. Wu, C. Li, and Y. He, "Zeroquant: Efficient and affordable post-training quantization for large-scale transformers," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 27 168–27 183. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/adf7fa39d65e2983d724ff7da57f00ac-Paper-Conference.pdf

[7] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han, "SmoothQuant: Accurate and efficient post-training quantization for large language models," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 38 087–38 099. [Online]. Available: https://proceedings.mlr.press/v202/xiao23c.html

[8] C. Wang, G. Szarvas, G. Balazs, P. Danchenko, and P. Ernst, "Calibrating verbalized probabilities for large language models," *arXiv preprint arXiv:2410.06707*, 2024.

[9] D. Ulmer, M. Gubri, H. Lee, S. Yun, and S. Oh, "Calibrating large language models using their generations only," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 15 440–15 459.

[10] S. J. Mielke, A. Szlam, E. Dinan, and Y.-L. Boureau, "Reducing conversational agents' overconfidence through linguistic calibration," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 857–872, 08 2022.

[11] J. Geng, F. Cai, Y. Wang, H. Koeppl, P. Nakov, and I. Gurevych, "A survey of confidence estimation and calibration in large language models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 6577–6595.

[12] Y. Fei, Y. Hou, Z. Chen, and A. Bosselut, "Mitigating label biases for in-context learning," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 14 014–14 031.

[13] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei *et al.*, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[14] jklj077, "Comment on issue #525," https://github.com/QwenLM/Qwen2.5/issues/525#issuecomment-2159944330, Jun. 2024, gitHub comment stating that the model's training data includes materials up to the end of 2023.

[15] G. Gerganov and ggml-org contributors, "`llama.cpp`: LLM inference in C/C++," 2023, latest tag: b5192 (2025-04-26). [Online]. Available: https://github.com/ggml-org/llama.cpp

[16] I. 'ikawrakow' Kawrakow, "k-quants," GitHub pull request #1684, issue comment #2474462323, https://github.com/ggml-org/llama.cpp/pull/1684#issuecomment-2474462323, 2023, accessed: 2025-04-30.

[17] K. A. Borgersen, "English k_quantization of llms does not disproportionately diminish multilingual performance," *arXiv preprint arXiv:2503.03592*, 2025.

[18] I. 'ikawrakow' Kawrakow, "k-quants," GitHub pull request #1684, https://github.com/ggml-org/llama.cpp/pull/1684, 2023, accessed: 2025-04-30.

[19] T. Naous, M. J. Ryan, A. Lavrouk, M. Chandra, and W. Xu, "ReadMe++: Benchmarking multilingual language models for multi-domain readability assessment," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 12 230–12 266. [Online]. Available: https://aclanthology.org/2024.emnlp-main.682/

[20] L. Qiu, S. Guo, T.-S. Wong, E. Chersoni, J. Lee, and C.-R. Huang, "CompLex-ZH: A new dataset for lexical complexity prediction in Mandarin and Cantonese," in *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, M. Shardlow, H. Saggion, F. Alva-Manchego, M. Zampieri, K. North, S. Štajner, and R. Stodden, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 20–26. [Online]. Available: https://aclanthology.org/2024.tsar-1.3/

[21] R. Adams, F. Adeleke, A. Florido, L. G. de Magalhães Santos, N. Grossman, L. Junck, and K. Stone, *Global Index on Responsible AI 2024 (1st Edition)*, 1st ed. South Africa: Global Center on AI Governance, 2024. [Online]. Available: https://girai-report-2024-corrected-edition.tiiny.site/

[22] E. Kran, H. M. Nguyen, A. Kundu, S. Jawhar, J. Park, and M. M. Jurewicz, "Darkbench: Benchmarking dark patterns in large language models," in *Workshop on Datasets and Evaluators of AI Safety*, 2025. [Online]. Available: https://openreview.net/forum?id=Vz1uCY5aG4

[23] OpenAI, "Compare models – openai api," https://platform.openai.com/docs/models/compare?model=gpt-4.1, accessed: 2025-04-28.

[24] E. Frantar, S. Ashkboos, T. Hoefler, and D.-A. Alistarh, "Gptq: Accurate post-training quantization for generative pre-trained transformers," in *11th International Conference on Learning Representations*, 2023. [Online]. Available: https://research-explorer.ista.ac.at/download/17378/17385/2023_ICLR_Frantar.pdf

[25] J. Lin, J. Tang, H. Tang, S. Yang, W.-M. Chen, W.-C. Wang, G. Xiao, X. Dang, C. Gan, and S. Han, "Awq: Activation-aware weight quantization for on-device llm compression and acceleration," *Proceedings of Machine Learning and Systems*, vol. 6, pp. 87–100, 2024. [Online]. Available: https://proceedings.mlsys.org/paper_files/paper/2024/file/42a452cbafa9dd64e9ba4aa95cc1ef21-Paper-Conference.pdf

[26] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 10 088–10 115. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf

VI. Appendix

A. Additional results of regression and classification evaluation

TABLE V: Regression: Zero-shot vs Few-shot Performance across Models and Datasets (ZS: zero-shot, FS: few-shot)

Model Family	Model Size	Quantization	Dataset	Pearson (ZS)	MSE (ZS)	Pearson (FS)	MSE (FS)
Qwen 2.5	7 B	q2	**GIRAI**	-0.630	0.487	0.158	0.214
Qwen 2.5	7 B	q4	**GIRAI**	-0.139	0.159	0.427	0.100
Qwen 2.5	7 B	q6	**GIRAI**	-0.432	0.346	0.499	0.081
Qwen 2.5	7 B	q8	**GIRAI**	-0.432	0.400	0.470	0.086
Qwen 2.5	14 B	q2	**GIRAI**	0.369	0.118	0.332	0.164
Qwen 2.5	14 B	q4	**GIRAI**	0.465	0.187	0.746	0.058
Qwen 2.5	14 B	q6	**GIRAI**	0.406	0.248	0.799	0.040
Qwen 2.5	32 B	q2	**GIRAI**	-0.153	0.212	0.640	0.082
Qwen 2.5	32 B	q4	**GIRAI**	0.467	0.101	**0.854**	**0.031**
Qwen 2.5	72 B	q2	**GIRAI**	-0.087	0.167	0.796	0.033
GPT-4o	N/A	N/A	**GIRAI**	0.618	0.116	0.847	0.029
GPT-4.1	N/A	N/A	**GIRAI**	0.710	0.080	**0.880**	**0.024**
Qwen 2.5	7 B	q2	**README++**	0.727	0.035	0.679	0.064
Qwen 2.5	7 B	q4	**README++**	0.736	0.068	0.651	0.049
Qwen 2.5	7 B	q6	**README++**	0.749	0.067	0.690	0.049
Qwen 2.5	7 B	q8	**README++**	0.747	0.063	0.698	0.051
Qwen 2.5	14 B	q2	**README++**	0.722	0.034	0.743	0.067
Qwen 2.5	14 B	q4	**README++**	0.743	0.045	0.750	0.070
Qwen 2.5	14 B	q6	**README++**	0.770	0.038	0.741	0.074
Qwen 2.5	32 B	q2	**README++**	0.764	0.029	0.740	0.050
Qwen 2.5	32 B	q4	**README++**	0.785	0.024	0.778	0.041
Qwen 2.5	72 B	q2	**README++**	**0.796**	**0.021**	0.795	0.027
GPT-4o	N/A	N/A	**README++**	0.790	0.037	0.769	0.034
GPT-4.1	N/A	N/A	**README++**	0.776	0.053	**0.788**	**0.028**
Qwen 2.5	7 B	q2	**CompLex-ZH**	0.261	0.021	0.177	0.189
Qwen 2.5	7 B	q4	**CompLex-ZH**	0.282	0.077	0.410	0.044
Qwen 2.5	7 B	q6	**CompLex-ZH**	0.292	0.049	0.442	0.043
Qwen 2.5	7 B	q8	**CompLex-ZH**	0.294	0.050	0.443	0.038
Qwen 2.5	14 B	q2	**CompLex-ZH**	0.374	0.015	0.365	0.047
Qwen 2.5	14 B	q4	**CompLex-ZH**	0.432	0.042	0.461	0.053
Qwen 2.5	14 B	q6	**CompLex-ZH**	0.396	0.040	0.469	0.060
Qwen 2.5	32 B	q2	**CompLex-ZH**	0.318	0.016	0.447	0.069
Qwen 2.5	32 B	q4	**CompLex-ZH**	0.301	0.023	**0.532**	0.075
Qwen 2.5	72 B	q2	**CompLex-ZH**	0.403	0.013	0.515	**0.035**
GPT-4o	N/A	N/A	**CompLex-ZH**	0.453	0.017	**0.580**	**0.062**
GPT-4.1	N/A	N/A	**CompLex-ZH**	0.459	0.021	0.562	0.063

TABLE VI: Classification on DarkBench: Zero-shot vs Few-shot Performance across Models (ZS: zero-shot, FS: few-shot)

Model Family	Model Size	Quant.	Acc. (ZS)	Prec. (ZS)	Rec. (ZS)	F1 (ZS)	Acc. (FS)	Prec. (FS)	Rec. (FS)	F1 (FS)
Qwen 2.5	7 B	q2	0.326	0.265	0.326	0.233	0.538	0.508	0.538	0.463
Qwen 2.5	7 B	q4	0.246	0.193	0.246	0.205	0.606	0.499	0.607	0.532
Qwen 2.5	7 B	q6	0.261	0.210	0.261	0.215	0.639	0.553	0.639	0.579
Qwen 2.5	7 B	q8	0.303	0.256	0.303	0.256	0.554	0.502	0.554	0.487
Qwen 2.5	14 B	q2	0.303	0.253	0.303	0.265	0.595	0.568	0.595	0.540
Qwen 2.5	14 B	q4	0.410	0.267	0.410	0.311	**0.712**	**0.779**	**0.713**	**0.663**
Qwen 2.5	14 B	q6	0.344	0.298	0.344	0.274	0.561	0.665	0.561	0.516
Qwen 2.5	32 B	q2	0.282	0.323	0.283	0.222	0.690	0.737	0.690	0.640
Qwen 2.5	32 B	q4	0.284	0.299	0.284	0.230	0.705	0.732	0.705	0.659
Qwen 2.5	72 B	q2	0.281	0.297	0.281	0.228	0.688	0.716	0.688	0.633
GPT-4o	N/A	N/A	0.390	0.303	0.390	0.314	0.766	0.774	0.766	0.766
GPT-4.1	N/A	N/A	0.428	0.319	0.428	0.338	**0.869**	**0.881**	**0.869**	**0.865**

Appreviations: **Quant.**: Quantization, **Acc.**: Accuracy, **Rec.**: Recall, **Prec.**: Precision, **F1**: F1-score.
Note: Since the DarkBench dataset is equally balanced, the macro and weighted results for Precision, Recall and F1-score are the same.