

## A APPENDIX

### A.1 DISCUSSION ON OUTLIER THRESHOLDS OF WEIGHT QUANTIZER

In this section, we discuss the selection and updating strategies for outlier thresholds in dense-and-sparse quantizers. We first report the memory and accuracy of dense-and-sparse quantizers using different percentage thresholds, and the results are shown in Table 4. The accuracy, i.e., the degree of distributional approximation of the quantizers, is evaluated by  $L_2$  distance between de-quantized weights  $\hat{W}$  and full-precision weights  $W$ , where the quantized weights are from the final down projection layer.

Table 4: Comparison of memory (in GB) and accuracy of dense-and-sparse quantizers using different percentage thresholds for weights. Here, accuracy is measured by  $L_2$  distance between de-quantized  $\hat{W}$  and full-precision  $W$ .

Percentile	0	0.45%	1.0%	3.0%	5.0%
Memory	7.06	7.23	7.42	8.23	9.16
$L_2$ Distance	436	0.846	0.619	0.566	0.479

The benefits of employing matrix decomposition in dense-and-sparse quantizers are readily evident from the results. When the percentile is set to 0, the quantizer no longer filters out the outliers and degenerates into a standard uniform quantizer, resulting in intolerable quantization errors. A noteworthy value suggested in SqueezeLLM (Kim et al., 2023) is 0.45%. However, it’s essential to acknowledge that this tight bound is primarily designed for inference scenarios. To accommodate potential fluctuations during training, some level of relaxation in the thresholds is necessary. To this end, we assess the relaxed constraints at percentiles of 1.0%, 3.0%, and 5.0%. The findings reveal that, although the 3% and 5% percentiles yield a slight boost in accuracy, they also incur higher memory usage. Consequently, we advocate for the 1% percentile as a more favorable choice for achieving a balance between memory conservation and accuracy.

We also explore the update frequency of the thresholds. In order to maintain a high training throughput, we adopt a strategy of lazy updates, meaning that the thresholds set in the first iteration are reused in subsequent iterations. This approach is viable due to the smooth numerical updates of the weights during the fine-tuning process across neighboring iterations. Additionally, to our surprise, we discover that the initial thresholds remain effective for an entire epoch. As a result, we only update the thresholds once at the beginning of each epoch.

### A.2 DISCUSSION ON TRAINING STABILITY

In this section, we delve into the essentiality of maintaining full-precision weight copies in mixed-precision training and provide a comprehensive illustration of the benefits of QFT through a comparative analysis. In mixed-precision training, both forward and backward propagation involve weights and gradients in the FP16 format. However, during parameter updates within the optimizer, the weights are involved in calculations using the FP32 format. This approach is employed to circumvent potential data overflow issues that may arise in the FP16 format (Micikevicius et al., 2017), given its limited representation range. Consequently, maintaining full-precision weight copies becomes indispensable to ensure the stability of the training process.

Conversely, QFT effectively addresses the aforementioned issue without necessitating extra storage for weight copies. Specifically, unlike the rudimentary numerical truncation encountered in the FP16 format, the quantization procedure in QFT maps FP32 format values as a cohesive unit to a set of evenly distributed integer values, as illustrated in Equation 1. Consequently, when performing the calculations, these integer values are dynamically de-quantized into the FP32 format, without any potential risk of data overflow.

### A.3 QUALITATIVE ANALYSIS OF CONVERSATIONAL ABILITIES

In the experimental section, we demonstrate the effectiveness of QFT through quantitative evaluations of LLMs' performance on a variety of benchmarks. Despite the large number of benchmarks, the results of the machine evaluations are not considered to be comprehensive, and in some cases, their reliability may be subject to scrutiny, e.g., the "shortcut" issue. Consequently, in the following, we intuitively show and compare examples generated by LLMs, aiming to further verify the conversational abilities through qualitative analysis.

#### [Writing Ability]

Question: Write a captivating movie review for a recently released science fiction film, discussing its plot, characters, and special effects.

**Vicuna-7B\*:** "The Last Voyage" is a thrilling science fiction film that follows a group of astronauts on a mission to explore a distant planet. When they arrive, they discover that the planet is inhabited by a mysterious alien race. As they try to communicate with the aliens and gather data, they realize that their mission may be more dangerous than they ever imagined. With stunning special effects and a gripping plot, "The Last Voyage" is a must-see for science fiction fans.

**LLaMA-2-7B-QFT:** "Intergalactic Intrigue" is a thrilling science fiction film that takes viewers on a journey through the stars. The film follows a group of space explorers as they embark on a mission to uncover the secrets of an ancient alien civilization. The film's special effects are top-notch, with stunning visuals that transport viewers to a world of wonder and discovery. The characters are well-developed and likable, with each member of the crew bringing their own unique skills and personalities to the table.

#### [Roleplay Ability]

Question: As a superhero, how would you explain your origin story to a curious child?

**Vicuna-7B\*:** I was just like any other kid, until one day I was exposed to a strange chemical that gave me incredible powers. I learned to control the elements and use them to protect the innocent. I became known as the Elemental Guardian, and I've been fighting crime and saving the day ever since.

**LLaMA-2-7B-QFT:** I was just like any other kid, playing in the park and having fun. But one day, I was hit by a strange ray that gave me superpowers. I could fly, turn invisible, and even control the weather! I decided to use my powers to help people and fight crime. That's how I became the superhero you know and love!

#### [Common-sense Ability]

Question: In a world where automation is becoming increasingly prevalent, is it more important to prioritize job creation or technological progress?

**Vicuna-7B\*:** Job creation is essential for a thriving economy, as it provides individuals with employment opportunities and a source of income. While technological progress is important for driving innovation and improving efficiency, it is crucial to balance technological advancements with the creation of new jobs. This can be achieved by investing in education and training programs that help workers develop the skills necessary to adapt to new technologies and by promoting policies that support small businesses and entrepreneurship. By prioritizing both job creation and technological progress, we can ensure a more stable and prosperous future for all.

**LLaMA-2-7B-QFT:** In the short term, job creation is a critical concern for policymakers and businesses. The unemployment rate remains high, and many people are struggling to find work. However, in the long term, technological progress is essential for driving economic growth and improving the standard of living for everyone. Automation has the potential to increase productivity and efficiency, which can lead to higher wages and better working conditions for employees. However, it can also lead to job displacement, particularly in industries where automation is most prevalent.

#### [Fermi Ability]

Question: How many snowflakes fall during a typical winter? Try to explain your answer. Your explanation should take the reader through your reasoning step-by-step.

**Vicuna-7B\***: It is difficult to determine the average number of snowflakes that fall during a typical winter, as it can vary greatly depending on location, climate, and other factors. However, it is estimated that on average, a single snowflake can have up to 100,000 ice crystals per flake. Over the course of a typical winter, this could result in billions of snowflakes falling from the sky.

**LLaMA-2-7B-QFT**: To answer this question, we need to consider the average snowfall for a typical winter in a given location. In the United States, for example, snowfall tends to be heaviest in the northern states, where the average snowfall for a typical winter is around 60 inches. In the southern states, snowfall is much less common and the average snowfall for a typical winter is around 10 inches.

**[Counterfactual Ability]**

**Question**: What if the Internet had been invented during the Renaissance period?

**Vicuna-7B\***: The Renaissance would have been a much different time period if the internet had been invented. The flow of information would have been much faster and more efficient, leading to a greater exchange of ideas and innovations. The world would have been connected in ways never before seen, allowing for the rapid spread of new ideas and technologies.

**LLaMA-2-7B-QFT**: The Renaissance was characterized by a renewed interest in classical learning and the arts, and the Internet would have provided a way for people to access and share this knowledge. It would have also facilitated the exchange of ideas and the development of new technologies, leading to even greater advancements in the fields of science, medicine, and engineering.

**[Generic Ability]**

**Question**: Can you explain the basics of quantum computing?

**Vicuna-7B\***: Quantum computing is a type of computing that uses quantum bits (qubits) to perform calculations, unlike traditional computers that use binary.

**LLaMA-2-7B-QFT**: Quantum computing is a type of computing that uses quantum-mechanical phenomena, such as superposition and entanglement, to perform operations on data. These phenomena allow quantum computers to perform certain calculations much faster than classical computers.

**[Knowledge Ability]**

**Question**: Explain the process of natural selection and how it contributes to the evolution and adaptation of species.

**Vicuna-7B\***: Natural selection is the process by which species adapt to their environment through the survival and reproduction of individuals with favorable traits. Over time, these favorable traits become more common in the population, leading to the evolution of new species.

**LLaMA-2-7B-QFT**: Natural selection is a process where species that are better adapted to their environment tend to survive and reproduce more than those that are less adapted. This leads to the survival of the fittest and the gradual evolution of species over time.