**Adapting Large Language Model for**

**JSON Extraction from Text Corpora**

|  |
| --- |
| Van Tuan-Tran1\*, Chin-Shiuh Shieh 2, Mong-Fong Horng3 |
| 1 National Kaohsiung University of Science and Technology / Student 1  2 National Kaohsiung University of Science and Technology / Professor 2,3  \*Email: {1f111169109, 2csshieh, 3mfhorng@nkust.edu.tw} |
| NSTC-112-2221-E-992-045, NSTC-112-2221-E-992-057-MY3  NSTC-112-2622-8992-009-TD1. |

# **Abstract**

This paper explores the adaptation of large language models (LLMs) for the task of extracting structured data in JSON format from extensive and unstructured text corpora. Specifically, we fine-tuned the Llama-2-7 billion model utilizing a combination of advanced techniques, including QLoRA (Quantized Low-Rank Adaptation), Fully Sharded Data Parallel (FSDP) training, and distributed training across multiple GPUs. Our approach addresses the challenges associated with processing large and complex datasets, which include HTML-tagged content, diverse textual paragraphs, and various formatting irregularities. By employing QLoRA, we achieve efficient low-rank adaptation, which helps in reducing the computational burden while retaining model performance. FSDP enables us to handle large-scale data by sharding model parameters and gradients, thus optimizing memory usage and speeding up the training process. Multi-GPU training further enhances scalability and accelerates the fine-tuning process, allowing us to manage and process extensive text corpora effectively. The fine-tuning process transforms raw textual information into structured JSON outputs, facilitating automated data extraction and processing. Our results demonstrate that the adapted model significantly improves the accuracy and efficiency of extracting relevant information from complex text sources, compared to traditional methods. This work not only showcases the effectiveness of combining these advanced techniques but also offers a scalable solution for applications such as web scraping, data parsing, and large-scale information retrieval.

**Keywords:** Fully Sharded Data Parallel (FSDP), Meta-Llama/Llama-2-7b, Meta-Llama/Llama-3-8b, Meta-Llama/Llama-3.1-8b, QLoRA, LoRA, wandb, tensorboard.

**Training Hyper-Parameter**

|  |  |
| --- | --- |
| Hyper-Parameter | Values |
| Model name | meta-llama/Llama-2-7b-hf  meta-llama/Llama-3-8b-hf  meta-llama/Llama-3.1-8b-hf |
| New model name | llama-2-7b-json\_extract-lora\_adapter  llama-3-8b-json\_extract-lora\_adapter  llama-31-8b-json\_extract-lora\_adapter |
| Dataset | chwenjun225/Instruction\_top\_5\_insurance  \_brands\_june\_news\_and\_twitter\_only |
| Batch size | 2 |
| Context length | 4096 |
| Precision | torch.float16 |
| Gradient accumulation steps | 4 |
| Epochs | 5 |
| Training type | QLoRA |
| Inference type | LoRA adapter |

**Meta-llama/Llama-2-7b**

A screen shot of a graph

Description automatically generated

**Figure 1.1:** The loss curves of finetuning Llama-2-7b-json\_extract-lora\_adapter

/////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Task** | **Version** | **Filter** | **n-shot** | **Metric** | | **Value** | | **Stderr** |
| arc\_challenge | 1 | none | 0 | acc | ↑ | 0.4258 | ± | 0.0144 |
|  |  | none | 0 | acc\_norm | ↑ | 0.4556 | ± | 0.0146 |
| arc\_easy | 1 | none | 0 | acc | ↑ | 0.7445 | ± | 0.0089 |
|  |  | none | 0 | acc\_norm | ↑ | 0.7235 | ± | 0.0092 |
| hellaswag | 1 | none | 0 | acc | ↑ | 0.5674 | ± | 0.0049 |
|  |  | none | 0 | acc\_norm | ↑ | 0.7531 | ± | 0.0043 |
| openbookqa | 1 | none | 0 | acc | ↑ | 0.3260 | ± | 0.0210 |
|  |  | none | 0 | acc\_norm | ↑ | 0.4420 | ± | 0.0222 |
| piqa | 1 | none | 0 | acc | ↑ | 0.7818 | ± | 0.0096 |
|  |  | none | 0 | acc\_norm | ↑ | 0.7802 | ± | 0.0097 |

**Table 1:** Llama-2-7b-json\_extract-lora\_adapter’s evaluation

/////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////

A graph of numbers and a number of numbers

Description automatically generated with medium confidence

**Figure 1.2:** Llama-2-7b-json\_extract-lora\_adapter - Metrics for Various Tasks

**Meta-llama/Llama-3-8b**

A graph with a line

Description automatically generated

**Figure 2.1:** The loss curves of finetuning Llama-3-8b-json\_extract-lora\_adapter

/////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Task** | **Version** | **Filter** | **n-shot** | **Metric** | | **Value** | | **Stderr** |
| arc\_challenge | 1 | none | 0 | acc | ↑ | 0.4172 | ± | 0.0144 |
|  |  | none | 0 | acc\_norm | ↑ | 0.4428 | ± | 0.0145 |
| arc\_easy | 1 | none | 0 | acc | ↑ | 0.6869 | ± | 0.0095 |
|  |  | none | 0 | acc\_norm | ↑ | 0.6553 | ± | 0.0098 |
| hellaswag | 1 | none | 0 | acc | ↑ | 0.5468 | ± | 0.0050 |
|  |  | none | 0 | acc\_norm | ↑ | 0.7196 | ± | 0.0045 |
| openbookqa | 1 | none | 0 | acc | ↑ | 0.2900 | ± | 0.0203 |
|  |  | none | 0 | acc\_norm | ↑ | 0.4000 | ± | 0.0219 |
| piqa | 1 | none | 0 | acc | ↑ | 0.7454 | ± | 0.0102 |
|  |  | none | 0 | acc\_norm | ↑ | 0.7508 | ± | 0.0101 |

**Table 2:** Llama-3-8b-json\_extract-lora\_adapter’s evaluation

///////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////A graph of numbers and a number of numbers

Description automatically generated with medium confidence

**Figure 2.2:** Llama-3-8b-json\_extract-lora\_adapter - Metrics for Various Tasks

**Meta-llama/Llama-3.1-8b**

A graph with a line

Description automatically generated

**Figure 2.1:** The loss curves of finetuning Llama-31-8b-json\_extract-lora\_adapter

///////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Task** | **Version** | **Filter** | **n-shot** | **Metric** | | **Value** | | **Stderr** |
| arc\_challenge | 1 | none | 0 | acc | ↑ | 0.4087 | ± | 0.0144 |
|  |  | none | 0 | acc\_norm | ↑ | 0.4309 | ± | 0.0145 |
| arc\_easy | 1 | none | 0 | acc | ↑ | 0.6999 | ± | 0.0094 |
|  |  | none | 0 | acc\_norm | ↑ | 0.6595 | ± | 0.0097 |
| hellaswag | 1 | none | 0 | acc | ↑ | 0.5502 | ± | 0.0050 |
|  |  | none | 0 | acc\_norm | ↑ | 0.7272 | ± | 0.0044 |
| openbookqa | 1 | none | 0 | acc | ↑ | 0.3040 | ± | 0.0206 |
|  |  | none | 0 | acc\_norm | ↑ | 0.3940 | ± | 0.0219 |
| piqa | 1 | none | 0 | acc | ↑ | 0.7481 | ± | 0.0101 |
|  |  | none | 0 | acc\_norm | ↑ | 0.7590 | ± | 0.0100 |

**Table 3:** Llama-3.1-8b-json\_extract-lora\_adapter’s evaluation

///////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////

A graph of numbers and a number of numbers

Description automatically generated with medium confidence

**Figure 3:** Llama-3.1-8b-json\_extract-lora\_adapter - Metrics for Various Tasks