**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

**Fig 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

|  |
| --- |
| Input |
| The final leg of the Kiss "End of the Road" tour begins in Cincinnati. The iconic band are wrapping up a 50-year career with a North American tour that starts at Heritage Bank Center in Cincinnati, and ends at New York City's Madison Square Garden. Tickets go on sale Friday, June 9, 2023. The end of the road begins in Cincinnati. The legendary rock 'n' roll band Kiss is closing out a 50-year career, but before the band packs away its iconic makeup and wild costumes, the boys are taking one last ride around the world with a final tour, fittingly titled the "End of the Road" tour. It will span 50 dates around the world, and the North American leg kicks off Oct. 19 right here in Cincinnati. Tickets go on sale Friday, June 9, for the show, which will take place at Heritage Bank Center (100 Broadway, Downtown). The tour wraps up in December with a massive final show at Madison Square Garden in New York City. Concert dates:Cincinnati's full 2023 concert calendar 🎵 Kiss was formed in New York City in 1973 by members Paul Stanley, Gene Simmons, Ace Frehley and Peter Criss. With greasepaint makeup and outrageous costumes, the bandmembers took on the personae of comic book-style characters, and their "shock-rock" style live performances have been known to feature fire-breathing, blood-spitting, levitating drum kits and pyrotechnics. Considered one of the most influential rock bands of all time and one of the best-selling bands of all time, Kiss has sold more than 75 million records worldwide, earned 30 gold albums, and all four original members have been inducted into the Rock and Roll Hall of Fame. The current lineup includes Stanley, Simmons, guitarist Tommy Thayer and drummer Eric Singer. Need a break? Play the USA TODAY Daily Crossword Puzzle.  Kiss 2023 North American End of the Road tour dates:  Oct. 19: Cincinnati, Heritage Bank Center  Oct. 20: Detroit, Little Caesars Arena  Oct. 22: Cleveland, Rocket Mortgage FieldHouse  Oct. 23: Nashville, Bridgestone Arena  Oct. 25: St. Louis, Enterprise Center  Oct. 27: Fort Worth, Texas, Dickies Arena  Oct. 29: Austin, Moody Center  Nov. 1: Palm Springs, Calif. Acrisure Arena  Nov. 3: Los Angeles, Hollywood Bowl  Nov. 6: Seattle, Climate Pledge Arena  Nov. 8: Vancouver, Rogers Arena  Nov. 10: Edmonton, Alberta, Rogers Place  Nov. 12: Calgary, Alberta, Scotiabank Saddledome  Nov. 13: Saskatoon, Saskatchewan, SaskTel Centre  Nov. 15: Winnipeg, Manitoba, Canada Life Centre  Nov. 18: Montreal, Quebec, Centre Bell  Nov. 19: Quebec, Videotron Centre  Nov. 21: Ottawa, Ontario, Canadian Tire Centre  Nov. 22: Toronto, Ontario, Scotiabank Arena  Nov. 24: Knoxville, Tenn., Thompson-Boling Arena  Nov. 25: Indianapolis, Gainbridge Fieldhouse  Nov. 27: Rosemont, Illinois, Allstate Arena  Nov. 29: Baltimore, CFG Bank Arena  Dec. 1: New York City, Madison Square Garden  Dec. 2: New York City, Madison Square Garden |
| Output of Llama-2-7b-Json\_extract-Lora\_adapter |
| ```{'Country': 'United States', 'Headline': 'Pucker up! Kiss to open final 'End of the Road' tour in Cincinnati 💋', 'Hit Sentence': '... Stanley, Simmons, guitarist Tommy Thayer and drummer Eric Singer. Need a break? Play the USA TODAY Daily Crossword Puzzle. Kiss 2023 ...', 'Influencer': 'Luann Gibbs', 'Language': 'English', 'Opening Text': 'The final leg of the Kiss "End of the Road" tour begins in Cincinnati. The iconic band are wrapping up a 50-year career with a North American tour ...', 'Source': 'Cincinnati Enquirer', 'URL': 'https://www.cincinnati.com/story/entertainment/music/2023/06/07/kiss-end-of-the-road-tour-cincinnati/71228854007/'}``` |
| Output of Llama-3-8b-Json\_extract-Lora\_adapter |
| ```<|eot\_id|><|start\_header\_id|>Response<|end\_header\_id|>:{'Country': 'United States', 'Headline': "Pucker up! Kiss to open final 'End of the Road' tour in Cincinnati 💋", 'Hit Sentence': '..., Saskatchewan, SaskTel Centre Nov. 15: Winnipeg, Manitoba, Canada Life Centre Nov. 18: Montreal, Quebec, Centre Bell Nov. 19: Quebec,...', 'Influencer': 'Luann Gibbs', 'Language': 'English', 'Opening Text': 'The final leg of the Kiss "End of the Road" tour begins in Cincinnati. The iconic band are wrapping up a 50-year career with a North American...', 'Source': 'Cincinnati Enquirer', 'URL': 'https://www.cincinnati.com/story/entertainment/music/2023/06/07/kiss-final-end-of-the-road-tour-comes-to-cincinnati/70283619007/'}<|end\_of\_text|>``` |
| Output of Llama-3.1-8b-Json\_extract-Lora\_adapter |
| ```<|eot\_id|><|start\_header\_id|>Response<|end\_header\_id|>: {'Country': 'United States', 'Headline': "Pucker up! Kiss to open final 'End of the Road' tour in Cincinnati 💋", 'Hit Sentence': '..., Saskatchewan, SaskTel Centre Nov. 15: Winnipeg, Manitoba, Canada Life Centre Nov. 18: Montreal, Quebec, Centre Bell Nov. 19: Quebec,...', 'Influencer': 'Luann Gibbs', 'Language': 'English', 'Opening Text': 'The final leg of the Kiss "End of the Road" tour begins in Cincinnati. The iconic band are wrapping up a 50-year career with a North American...', 'Source': 'Cincinnati Enquirer', 'URL': 'https://www.cincinnati.com/story/entertainment/music/2023/06/07/kiss-final-end-of-the-road-tour-comes-to-cincinnati/70283619007/'}<|end\_of\_text|>``` |

**Table 5:** Inference Result of three models in English language