# Can Large Language Models Understand Context?

Yilun Zhu[1], Joel Ruben Antony Moniz[2], Shruti Bhargava[2], Jiarui Lu[2]
Dhivya Piraviperumal[2], Site Li[2], Yuan Zhang[2], Hong Yu[2], Bo-Hsiang Tseng[2]

[1]Georgetown University [2]Apple

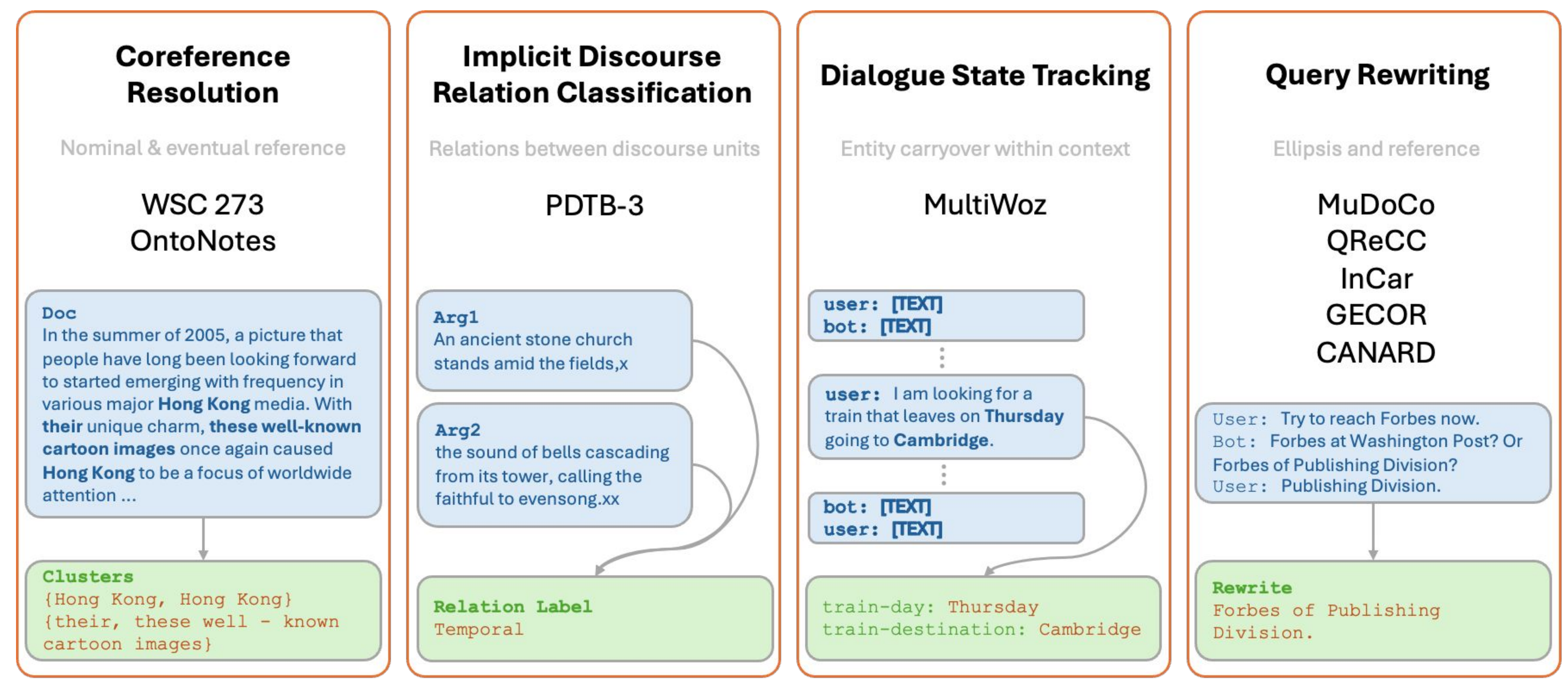
GEORGETOWN UNIVERSITY

EACL 2024

## ▶ Introduction

- Motivation
  - Understanding context is key to understanding human language, and an essential ability for Large Language Models (LLMs).
  - The size of LLMs hinders the deployment of large models to personal devices and restricts the on-device performance of language understanding tasks.
- NLP tasks that demand a nuanced comprehension of linguistic features within a provided context are under studied in previous LLM evaluations.
- Our context understanding benchmark aims to provide a comprehensive and in-depth evaluation of LLMs from multiple linguistic perspectives.
- We study the performance of various dense and 3-bit quantized LLMs on the query rewriting task.

## ▶ The Context Understanding Benchmark




## ▶ Experiments

- Three model families
  - OPT (Zhang et al., 2022)
  - LLaMA (Touvron et al., 2023)
  - GPT (OpenAI, 2023)
- Various model sizes
  - OPT: 125M, 350M, 1.3B, 2.7B, 6.7B, 13B, 30B
  - LLaMA: 7B, 13B, 30B
  - GPT: 3.5-turbo

**Instruction**: Given two arguments and a list of connective words, please select the most likely connective between two arguments.
**[Relation Description]**
**Input**:
Arg 1: Amcore, also a bank holding company, has assets of $1.06 billion.
Arg 2: Central's assets are $240 million.
**Question**: What is the connective that best describes the relation between two arguments?
**Choices**:
A. Temporal B. Contingency C. Comparison D. Expansion
**Answer**: *C*

Table 3: A PDTB example of prompt and *answer*.

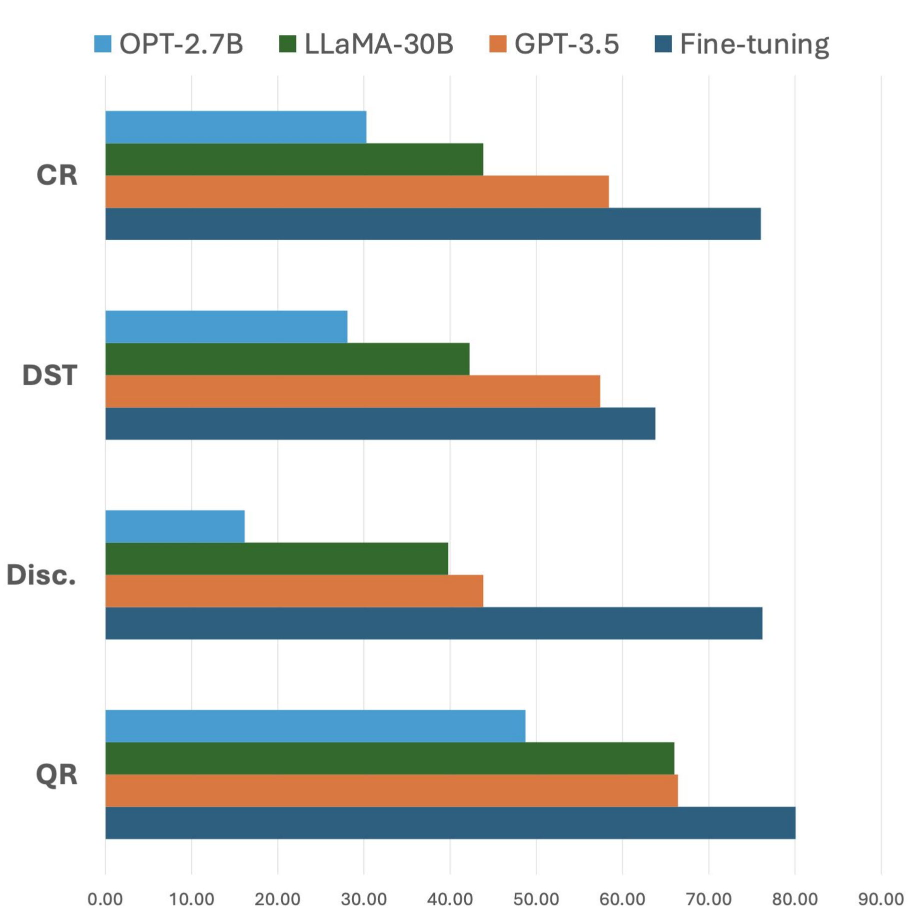



Figure 2: Comparison between commercial/non-commercial models and fine-tuning models for each task in the context understanding benchmark.

| Task | Dataset | Metrics | OPT | | | | LLaMA | | | GPT | FT |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 125M | 350M | 1.3B | 2.7B | 7B | 13B | 30B | 3.5-turbo | |
| CR | WSC273 | Acc | 58.24 | 66.67 | 76.19 | 77.66 | 86.81 | 89.38 | 89.01 | 88.64 | N/A |
| | OntoNotes | MUC | 12.66 | 7.58 | 13.21 | 8.29 | 10.31 | 31.80 | 33.56 | 56.32 | 77.26 |
| | | $B^3$ | 53.80 | 52.26 | 53.54 | 52.41 | 52.20 | 58.43 | 58.66 | 68.20 | 73.43 |
| | | $CEAF_{\phi 4}$ | 31.09 | 29.49 | 31.40 | 30.10 | 32.63 | 38.00 | 39.27 | 50.72 | 74.46 |
| | | Avg. F1 | 32.52 | 29.78 | 32.72 | 30.27 | 31.71 | 42.74 | 43.83 | 58.41 | 76.03 |
| DST | MultiWOZ | JGA | 11.11 | 27.96 | 26.61 | 28.08 | 32.30 | 28.12 | 42.24 | 57.40 | 63.79 |
| Disc. | PDTB-3 | Acc | 10.04 | 10.04 | 10.04 | 16.15 | 17.16 | 26.01 | 39.77 | 43.83 | 76.23 |
| QR | MuDoCo | BLEU | 0.46 | 0.36 | 7.02 | 49.20 | 41.12 | 61.15 | 66.51 | 57.14 | 80.31 |
| | | ROUGE | 1.52 | 12.18 | 10.98 | 65.61 | 56.07 | 74.78 | 77.88 | 79.37 | 92.01 |
| | QReCC | BLEU | 4.53 | 31.27 | 26.35 | 40.09 | 28.19 | 38.64 | 58.68 | 55.24 | 58.67 |
| | | ROUGE | 13.91 | 58.18 | 53.10 | 68.32 | 48.27 | 56.40 | 78.74 | 79.98 | 81.75 |
| | InCar | BLEU | 0.00 | 7.66 | 12.71 | 27.42 | 28.20 | 42.13 | 48.58 | 63.66 | 88.45 |
| | | ROUGE | 3.41 | 28.76 | 30.45 | 49.63 | 49.96 | 56.73 | 64.18 | 83.51 | 95.24 |
| | GECOR | BLEU | 0.20 | 26.40 | 26.32 | 49.99 | 53.27 | 66.30 | 73.80 | 63.34 | 82.56 |
| | | ROUGE | 4.06 | 42.13 | 42.57 | 65.89 | 69.23 | 80.99 | 86.03 | 79.00 | 92.63 |
| | CANARD | BLEU | 2.61 | 19.39 | 24.24 | 34.66 | 21.34 | 29.32 | 47.24 | 47.12 | 57.46 |
| | | ROUGE | 9.82 | 45.63 | 49.36 | 62.73 | 38.17 | 46.61 | 69.73 | 74.61 | 81.06 |

Table 5: Few-shot results of two open-sourced models and GPT-3.5 on the context understanding benchmark. The results with the best number of few-shot examples are reported for each task. Fine-tuning (FT) results serves as a reference when evaluating LLMs' capability under ICL setup.

## ▶ Dense vs. Quantized (Query Rewriting)

3-bit post-training quantization GPTQ (Frantar et al., 2022)

Two types of errors
- Error type 1: **repeat the last query**
- Error type 2: **language modeling**

- **Example**

`User:` what is the name of india pakistan border line
`Bot:` The Radcliffe Line was the boundary demarcation line between the Indian and Pakistani portions of the Punjab and Bengal provinces of British India.
`User:` who created the radcliffe line
`Bot:` The Radcliffe Line was named after its architect, Sir Cyril Radcliffe, who was the joint chairman of the two boundary commissions for the two provinces.
`User:` when was the line published

**Gold answer**: when was the radcliffe line published

**Prediction 1 (repeat the last query):** when was the line published

**Prediction 2 (language modeling):** 1947

| Dataset | Metrics | 7B-D | 30B-Q | 30B-D |
|---|---|---|---|---|
| WSC273 | Acc | 86.81 | 87.18 | 89.01 |
| OntoNotes | MUC | 10.31 | 25.37 | 33.56 |
| | $B^3$ | 52.20 | 56.80 | 58.66 |
| | $CEAF_{\phi 4}$ | 32.63 | 36.93 | 39.27 |
| | Avg. F1 | 31.71 | 39.70 | 43.83 |
| MultiWOZ | JGA | 32.30 | 41.99 | 42.24 |
| PDTB-3 | Acc | 17.16 | 31.29 | 39.77 |
| MuDoCo | BLEU | 41.12 | 59.22 | 66.51 |
| | ROUGE | 56.07 | 71.38 | 77.88 |
| QReCC | BLEU | 28.19 | 53.72 | 58.68 |
| | ROUGE | 48.27 | 74.13 | 78.74 |
| InCar | BLEU | 28.20 | 39.69 | 48.58 |
| | ROUGE | 49.96 | 56.32 | 64.18 |
| GECOR | BLEU | 53.27 | 70.41 | 83.36 |
| | ROUGE | 69.23 | 73.80 | 86.03 |
| CANARD | BLEU | 21.34 | 45.07 | 47.24 |
| | ROUGE | 38.17 | 67.15 | 69.73 |

Table 6: Comparison between dense and quantized models. Dense LLaMA-7B and 3-bit quantized LLaMA-30B share similar memory and disk requirements. **D** represents dense model and **Q** denotes quantized model.

| Type | Dataset | 7B D | 30B Q | 30B D |
|---|---|---|---|---|
| Repeat | MuDoCo | 260 | 247 | 194 |
| | QReCC | 86 | 90 | 26 |
| | InCar | 17 | 15 | 8 |
| | GECOR | 59 | 62 | 37 |
| | CANARD | 47 | 44 | 32 |
| | Total | 469 | 458 | 297 |
| LM | MuDoCo | 71 | 29 | 16 |
| | QReCC | 80 | 28 | 16 |
| | InCar | 19 | 20 | 15 |
| | GECOR | 6 | 1 | 0 |
| | CANARD | 127 | 76 | 59 |
| | Total | 232 | 125 | 106 |

Table 9: Number of the major two types errors on three LLaMA models (7B dense, 30B quantized, and 30B dense) found in query rewriting. *Repeat* stands for repeat-the-last-query error and *LM* denotes language modeling error.

## ▶ OPT vs. LLaMA (Query Rewriting)

- Prior works (Beeching et al., 2023) have consistently shown that, under the same model size, LLaMA outperforms OPT
- Our findings (query rewriting)
  - Model size ~7B
    - OPT > LLaMA
  - Model size ~13B
    - OPT ≈ LLaMA
  - Model size ~30B
    - OPT < LLaMA

| Dataset | 6.7/7B | | 13B | | 30B | |
|---|---|---|---|---|---|---|
| | O. | L. | O. | L. | O. | L. |
| Mudoco | 53.1 | 41.1 | 55.2 | 61.1 | 55.2 | 66.5 |
| | 71.8 | 56.0 | 72.1 | 74.7 | 71.5 | 77.8 |
| QReCC | 46.6 | 28.1 | 43.7 | 38.6 | 43.8 | 58.6 |
| | 73.4 | 48.2 | 71.6 | 56.4 | 71.9 | 78.7 |
| InCar | 40.3 | 28.2 | 41.9 | 42.1 | 44.6 | 48.5 |
| | 64.8 | 49.9 | 62.6 | 56.7 | 65.3 | 64.1 |
| GECOR | 58.8 | 53.2 | 60.9 | 66.3 | 58.2 | 73.8 |
| | 75.7 | 69.2 | 78.3 | 80.9 | 76.1 | 86.0 |
| CANARD | 43.8 | 21.3 | 37.5 | 29.3 | 41.3 | 47.2 |
| | 72.0 | 38.1 | 66.0 | 46.6 | 69.3 | 69.7 |

Table 7: Comparison between OPT (O.) and LLaMA (L.) across five query rewrite datasets. For each dataset, the first and second rows represent BLEU and ROUGE scores respectively.

## ▶ Conclusion

- Introduce a context understanding benchmark designed to assess the performance of LLMs.
- LLMs under in-context learning struggle with nuanced linguistic features within this challenging benchmark, exhibiting inconsistencies with other benchmarks that emphasize other aspects of language.
- 3-bit post-training quantization reduces the general understanding capacity of context to different extent across the 4 tasks.

Paper

SCAN ME!

Poster

SCAN ME!