Can Large Language Models Understand Context?

Anonymous ACL submission

Abstract

Understanding context is key to understanding human language, an ability which Large Language Models (LLMs) have been increasingly seen to demonstrate to an impressive extent. However, though the evaluation of LLMs encompasses various domains within the realm of Natural Language Processing, limited attention has been paid to probing their linguistic capability of understanding contextual features. This paper introduces a context understanding benchmark by adapting existing datasets to suit the evaluation of generative models. This benchmark comprises of four distinct tasks and ninedatasets, all featuring prompts designed to assess the models' ability to understand context. First, we evaluate the performance of LLMs under the in-context learning pretraining scenario. Experimental results indicate that pre-trained dense models struggle with understanding more nuanced contextual features when compared to state-of-the-art fine-tuned models. Second, as LLM compression holds growing significance in both research and real-world applications, we assess the context understanding of quantized models under in-context-learning settings. We find that 3-bit quantization leads to varying degrees of performance reduction on our benchmark. We conduct an extensive analysis of these scenarios to substantiate our experimental results.

1 Introduction

Discourse understanding, as one of the fundamental problems in NLP, focuses on modeling linguistic features and structures that go beyond individual sentences (Joty et al., 2019). Understanding discourse requires resolving the relations between words/phrases (coreference resolution) and discourse units (discourse parsing and discourse relation classification) in the previous context, identifying carry-over information for the following context (dialogue state tracking), and recognizing discourse-specific phenomena (ellipsis).

LLMs have garnered substantial attention from both academia and the industry due to their remarkable capability in comprehending language and world knowledge. Their unparalleled performance across a diverse range of benchmarks and datasets has firmly established their significance in a relatively short period of time. As LLMs continue to push the boundaries of scale and capability, the evaluation of their multifaceted abilities becomes an equally vital endeavor. Consequently, the development of robust evaluation methodologies to assess specific aspects of LLMs becomes imperative. In addition, these methodologies should focus on helping achieve a comprehensive understanding of their advancement while clearly delineating their limitations. However, recently published LLMs, such as OPT (Zhang et al., 2022), LLaMA (Touvron et al., 2023) and GPT-4 (OpenAI, 2023), are only evaluated on limited benchmarks, and have a significant drawback: they neglect the inclusion of discourse-related datasets for evaluation, thereby limiting the comprehensive assessment of their language understanding capabilities.

To provide a comprehensive evaluation, plenty of benchmarks and datasets address various facets of language understanding, including benchmarks that delve into common sense knowledge (Hendrycks et al., 2021a; Kwiatkowski et al., 2019), as well as linguistic capabilities like sentiment analysis, natural language inference, summarization, text classification, and more (Bang et al., 2023b; Liang et al., 2022). These general benchmarks and specific dataset evaluations exhibit certain limitations. Despite the requirement for contextual information in these benchmarks to effectively tackle tasks (for example, sentiment analysis requires an understanding of polarities within the given text), none of these benchmarks cater to tasks that demand a nuanced comprehension of linguistic features within a provided context.

On the other hand, recent LLMs, by virtue of possessing billions of parameters, have led to an exponential surge in computational and storage costs (Brown et al., 2020b), which hinders the deployment of large models to personal devices and restricts the on-device performance of language understanding tasks. To address this challenge, model compression methods, which can reduce memory and disk requirements of both model training and inference, have gained attention. Existing compression techniques, such as 3-bit quantization (Frantar et al., 2022), have demonstrated the potential to reduce model sizes with only marginal performance trade-offs. However, the evaluation of quantization methods suffers from two deficiencies. Firstly, quantization methods are primarily evaluated on limited benchmarks and datasets, such as Lambada (Paperno et al., 2016), ARC (Boratko et al., 2018), PIQA (Tata and Patel, 2003), BoolQ (Clark et al., 2019), and StoryCloze (Mostafazadeh et al., 2017). Secondly, previous work has not delved into a linguistic analysis to identify where the model efficacy wanes.

Given the above shortcomings, this paper evaluates LLMs on a context understanding benchmark constructed from varied discourse understanding datasets. We conduct an extensive analysis of LLM performance on this benchmark, including models of varying sizes and those subjected to compression techniques, aiming to provide a more comprehensive understanding of context understanding capability of the LLMs. The contributions of this paper can be summarized as follows:

- Our work introduces a contextual understanding benchmark, including four tasks, for the evaluation of LLMs. We also present prompts designed for in-context learning on each task.
- We evaluate LLMs of varying sizes from different model families and provide an analysis on these models' capability for context understanding.
- We evaluate post-training compressed models in ICL settings and conduct an analysis of the reduction in context understanding capability compared to dense models.

2 Related Work

2.1 In-context Learning Evaluation

The paradigm of ICL (Brown et al., 2020a) is rapidly gaining importance. Studies have demonstrated that the generalization of LLMs to various downstream NLP tasks, such as MMLU (Hendrycks et al., 2021b), is significantly enhanced when provided with a small number of examples as prompts (Brown et al., 2020a; Chowdhery et al., 2022; Hoffmann et al., 2022; Rae et al., 2022; Anil et al., 2023; Touvron et al., 2023; OpenAI, 2022, 2023). Recent research has extensively evaluated the performance of LLMs across a spectrum of language-related tasks, spanning from text generation to understanding input sequences. This assessment contains a wide array of benchmarks, including SUPER-GLUE (Wang et al., 2019; Laskar et al., 2023), and tasks such as question answering, information retrieval, sentiment analysis (Bang et al., 2023b; Liang et al., 2022), dialogue (Heck et al., 2023), and text classification (Yang and Menczer, 2023).

2.2 Model Compression for LLMs

Model compression techniques can be broadly categorized into three main approaches: compression during training, compression associated with fine-tuning, and post-training methods. In terms of quantization during training, this technique enables LLMs to adapt to low-precision representations during the training process (Liu et al., 2023). Model compression with fine-tuning involves quantization awareness into the fine-tuning stage (Kim et al., 2023; Dettmers et al., 2023). Post-training techniques, on the other hand, are applied after the completion of an LLMs training phase and typically involve the use of calibration data. This category comprises two primary approaches: pruning, which removes redundant or non-salient weights to induce sparsity (Frantar and Alistarh, 2023), and quantization, which employs low-precision numeric representations of weights and activations (Nagel et al., 2020; Frantar et al., 2022; Yuan et al., 2023). Prior research shows that quantization outperforms pruning in several settings (Kuzmin et al., 2023), thus in this work, we focus on model quantization and its impact on the selected context-aware tasks.

3 Task Selection & Design

Our contextual understanding benchmark includes four tasks with nine datasets, as presented in Table 1. In the following sections, we provide detailed explanations of each task and the corresponding datasets, along with the designed prompts for ICL evaluations.

Type	Task	Dataset	Context
Doc	Coreference	WSC273	Nominal & eventual reference
		OntoNotes	
	Discourse	PDTB-3	Relations between discourse units
Dial.	DST	MultiWoz	Entity carryover within context
	Query Rewrite	MuDoCo	Ellipsis and reference
		QReCC	
		InCar	
		GECOR	
		CANARD	

Table 1: Tasks and datasets in the context understanding benchmark.

Instruction: Please carefully read the following passages. For each passage and the options, you must identify which option the mention marked in *bold* refers to. If the marked mention does not have any antecedent, please select "no antecedent".
Context: ... To express *its* determination ... the Chinese securities regulatory department ... this stock reform ...
Choices:
A. no antecedent
B. the Chinese securities regulatory department
C. this stock reform
...
Question: What does *its* refer to?
Answer: *B*

Table 2: An OntoNotes example of prompt and *answer*.

3.1 Coreference Resolution

The coreference resolution task (CR) contributes to achieving a coherent understanding of the overall meaning conveyed within the text. Thus, it plays a critical role in diving into language models' capability to grasp coreference relations as well as contextual nuances within documents. We select two coreference datasets: WSC273 (Levesque et al., 2012) and OntoNotes 5.0 (Pradhan et al., 2013).

WSC273, which contains the first 273 examples from the Winograd Schema Challenge, is a dataset that requires the system to read a sentence with an ambiguous pronoun and select the referent of that pronoun from two choices. OntoNotes is a human-annotated corpus of documents annotated with multiple layers of linguistic information including syntax, propositions, named entities, word sense, and in-document coreference. As it is one of the most frequently used datasets for training coreference models, prior research has achieved significant advancements under the supervised fine-tuning paradigm (Lee et al., 2017; Joshi et al., 2020; Bohnet et al., 2023). However, these model designs cannot be extended to generative models under ICL settings. Recently, Le and Ritter (2023) have leveraged document templates for LLMs; however, their evaluation is confined to prominent models such as InstructGPT (Ouyang et al., 2022), neglecting the fact that smaller models lack the generative capacity required to accomplish such tasks. Due to these limitations, we propose a novel multiple-choice task design. In this design, we provide the mentions and evaluate the model on resolution. Each option represents a potentially markable span.[1] Table 2 presents an example of the input to the model[2]. The entire prompt consists of five parts: (1) an instruction that provides guidance to the model for the task, (2) a document containing plain text with a selected mention span highlighted using a bold symbol, (3) a list of choices, which includes all the gold mentions present in the document, (4) a question that directs the model's attention, and (5) a guiding word *answer* that prompts for the output. We experiment with multiple instructions and prompts and provide the one with the best performance. Linking scores are computed for each question and the results are subsequently aggregated for evaluation. We utilize the official evaluation metrics from the CoNLL-2012 shared task (Pradhan et al., 2012), which employs the CoNLL F1 score, derived from the averaging of three coreference metrics: MUC, B^3, and $CEAF_{\phi 4}$.

3.2 Dialogue State Tracking

Dialogue state tracking (DST) is an important task in the area of task-oriented dialogue (TOD) modeling (Young et al., 2013), where the dialogue agent tracks the key information provided by the user as the conversation progresses. Table 3 provides an example from MultiWOZ (Budzianowski et al., 2018) where the user expresses the constraints when looking for a restaurant. The output of DST is typically maintained in slot-value pair format.

Previous research has explored ICL capabilities on MultiWOZ and demonstrated promising results compared to fine-tuning models (Hu et al., 2022; Heck et al., 2023). However, these studies either involve partial training or are untested with smaller and quantized models. Here we adopt a straightforward and simplified ICL approach proposed by

[1] Considering the inferior performance of small models on the mention detection task, we utilize gold markable spans coreference linking.

[2] Detailed examples for each task design can be found in Appendix A.

Ontology:
{“slots”: {“restaurant-pricerange”: “price budget for the restaurant”, ... },
“categorical”: {“restaurant-pricerange”: [‘cheap’, ‘expensive’, ‘moderate’], ...} }
Instruction: Now consider the following dialogue between two parties called the “system” and “user”. Can you tell me which of the “slot” was updated by the “user” in its latest response to the “system”? Present the updates in JSON format. If no “slots” were updates, return an empty JSON list. If you encounter “slot” that was requested by the “user” then fill them with “?”. If a user does not seem to care about a discussed “slot” fill it with “dontcare”.
[Previous Dialogue State]
[Conversation]:
“system”: “”
“user”: “I’m looking for a moderately priced place to eat that’s in the centre of town.”
Output: *{“restaurant-pricerange”: “moderate”, “restaurant-area”: “centre”}*

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

Heck et al. (2023), and test it on MultiWOZ v2.2 (Zang et al., 2020). The prompt to the model consists of domain knowledge from ontology, an instruction, previous dialogue state (the belief state accumulated until the previous user turn) and the conversation proceeding to the current turn. The ontology could be very long if we consider all domains in the dataset; thus, given the input length constraint of LLMs, only the knowledge relevant to the conversation is provided. Following literature, we report joint goal accuracy (JGA) (Mrkšić et al., 2017) for evaluating the performance of DST.

3.3 Implicit Discourse Relation Classification

Discourse demonstrates its importance beyond individual sentences, which emphasizes the ways in which different segments of a text interconnect and structure themselves to convey a coherent and meaningful message. The PDTB-3 corpus, as introduced by Webber et al. (2019), annotates implicit discourse relations across elementary discourse units (EDUs)[3]. These relations imply connections between EDUs and may be made explicit by inserting a connective. Within the context of the understanding benchmark, we opt for the implicit discourse relation classification task for two primary reasons. Firstly, the order of the two EDUs is provided, enabling the model to directly utilize this information. Secondly, the connective triggering the relation is implicit, increasing the task’s complexity. In this task (Disc.), two EDUs are fed as input, and the objective of the task is to

[3]EDU refers to the smallest segment of a text that conveys a complete and coherent meaning within larger discourse.

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 4: A PDTB example of prompt and *answer*.

correctly identify the relation between them. Due to the nuanced differences between each relation and the demand for annotators with rich linguistic knowledge and extensive annotation training, the classification task poses challenges to fine-tuned classification models.

The PDTB3 corpus classifies discourse relations into four categories - `Temporal`, `Contingency`, `Comparison`, and `Expansion`. We convert this task into a multiple-choice question and experiment with *classes* as options. In the *classes* scenario, the task offers four options, with each representing a distinct discourse relation class. Table 4 exhibits the components of the prompt. It includes an instruction at the beginning, followed by a concise description of each relation, a context with two arguments, a question along with answer choices, and a trigger word. We evaluate each model’s performance on this dataset using accuracy as the metric.

3.4 Query Rewrite

While document-based CR (OntoNotes, Section 3.1) covers various types of coreference relations across multiple genres, it does not allow the ability to evaluate certain aspects which are important to understand context. Firstly, the CR task typically focuses on document-based coreference chains , neglecting mention resolution in dialogues. Secondly, ellipsis, which is the omission of one or more words from a clause while still allowing it to be understood in context, is a crucial linguistic phenomenon frequently encountered in speech and conversation. It is essential for language models to grasp and accurately identify ellipses within context. Incorporating these features into the benchmark is thus pivotal when evaluating context understanding.

Query rewrite (QR) is the task of rewriting the last utterance of a user into a context-free, independent utterance that can be interpreted without dia-

Instruction: Rewrite the last query following interaction into a well-formed, context independent query. Resolve any disfluencies or grammatical errors in the query.
Input:
User: Try to reach Forbes now .
Bot: Forbes at Washington Post ? Or Forbes of Publishing Division ?
User: Publishing Division .
Rewrite: *Forbes of Publishing Division*

Table 5: A query rewrite example of prompt and *answer*.

log context. The objective of the task is to identify the entity or events references from the previous query, whether through a pronoun or an omitted word/phrase, and then generate a new query that includes the previous context directly.

We incorporate five QR datasets in the proposed benchmark: MuDoCo (Martin et al., 2020), QReCC (Anantha et al., 2021), InCar (Regan et al., 2019), GECOR (Quan et al., 2019), and CANARD (Elgohary et al., 2019). These datasets span multiple genres and domains in dialogues. We experiment with various prompts used for fine-tuning models and present the results with the best selections. Table 5 presents a concise prompt comprising an instruction along with context for each dialogue. To assess the quality of generated queries, we follow the metrics from previous research, particularly BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004).

4 Experiments

The evaluation was conducted on a computational infrastructure comprising 8 $\times$ A100 GPUs. We experiment with three model families. For smaller models, we consider OPT (Zhang et al., 2022), ranging from 125M to 2.7B. Although OPT also offers larger models, we opt for LLaMA (Touvron et al., 2023) as the mid-sized LMs, spanning from 7B to 65B parameters, due to showcased superior performance by prior works. For large-scale LMs, we leverage `GPT-3.5-turbo`[4]. For each model, on every dataset, we assess five different settings: zero-shot, one-shot, 5-shot, 8-shot, and 10-shot. We randomly select the examples from the training set for the few-shot prompting.[5]

[4]https://platform.openai.com/docs/models/gpt-3-5

[5]WSC273 itself is a test set and it does not have any fine-tuning scores, so we only report the zero-shot results in the table.

4.1 Dense Model

Results of the three model families are reported in Table 6, along with results of fine-tuned (FT) models to help better interpret how well the pre-trained models behave with ICL. For each, we present the N-shot setting that yields the highest score (see Appendix B for details). Overall, performance improves as the model size increases and pre-trained models with ICL struggle to catch up with FT models on most tasks.

Coreference Resolution Larger models exhibit promising performance on the WSC273 task, indicating that LLMs can effectively handle "simple" coreference relations within limited contexts and mentions. However, when it comes to document-based CR with complex clusters, their performance substantially drops [6]. Even on providing ground-truth mentions, the highest-performing GPT is only on par with rule-based coreference systems (Manning et al., 2014) and is far from the end-to-end fine-tuned SpanBERT (Joshi et al., 2020). The gap between ICL and FT results highlights that under the ICL setting, LLMs struggle to build coreference chains without adequate domain-specific examples. Specifically, models except GPT perform significantly worse on the MUC metric. Error analysis reveals that these models are inclined to create more clusters, including singleton clusters. This implies that pre-trained LLMs encounter difficulties in understanding long-range contextual information.

DST A similar trend is observed as CR where OPT and LLaMA models fall behind GPT-3.5 significantly. This suggests that these models fail to extract key information as the conversation proceeds, even with the provision of 5 to 10 demonstrations and the distilled relevant domain ontology in prompt. Our error analysis indicates that most of the errors happen due to the misdetection of slots or the wrong predicted value in a slot-value pair. Only GPT-3.5 reaches the level of FT results which is a fine-tuned T5 base model (Bang et al., 2023a).

Implicit discourse relation classification We observe an increase in scores when the model size exceeds 7B. However, even the best-performing

[6]Note that the OntoNotes dataset is substantially larger than the others. We observe that inference on the entire test set becomes extremely time-consuming, particularly with the larger models; further, the cost of running inference on GPT-3.5 starts becoming non-negligible. Consequently, we propose limiting the OntoNotes test set to a 10% sub-sample, which is the setting we consistently adopt.

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	31.8	33.56	56.32	77.26
		B^3	53.8	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	38.0	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	49.2	41.12	61.15	66.51	57.14	80.31
		ROUGE1	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
		ROUGE1	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
		ROUGE1	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
		ROUGE1	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
		ROUGE1	9.82	45.63	49.36	62.73	38.17	46.61	69.73	74.61	81.06

Table 6: 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.

model, GPT, achieves > 30 points lower than the current SOTA fine-tuned model (Liu and Strube, 2023). We carefully examine the predictions for each model and found that all models tend to predict the same relation class for every example, albeit with their individual preferences for the selected relation. This suggests that the models struggle to distinguish the nuances between different relation classes and fail to correctly identify relations across EDUs within context.

Query Rewriting The gap between small and large models is significantly huge, compared to the other tasks. For instance, OPT-125M cannot even complete the rewriting task. Analysis on predictions of small models indicates that the model is not capable of following the instructions or learning patterns from the few-shot examples. We identify a few major error types: (1) generating the next sentence, instead of rewriting; (2) rewriting the wrong user turn from the conversation; (3) copying the last user utterance without any rewriting. These errors get reduced as the model size increases. However, similar to the previous three tasks, the best ICL results achieved by GPT is far from the fine-tuned models.[7] It is worth noting that OPT-2.7B performs on par or notably better than LLaMA-7B, which is somewhat not aligned with the findings in Beeching et al. (2023) where LLaMA-7B even outperforms OPT-66B in many tasks, including ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), and MMLU (Hendrycks et al., 2021b).

All in all, this section presents a holistic comparison of LLMs' behaviors on the target context understanding tasks. On the tasks with structured outputs such as CR or DST, even small models show a certain level of context understanding and seem to follow the task instruction. Classification tasks such as discourse relation selection are deemed the easiest among all tasks; however, the small models are even worse than a random guess (25%). As for the generative task, the ability to complete query rewriting can be only observed in the case of larger models, as the model has the freedom to generate arbitrary content that does not follow the prompt. Except for DST, FT models demonstrate marked superiority over pre-trained models, highlighting the potential for improving LLMs' competence on these context understanding tasks.

4.2 Model Compression Technique

As we focus on evaluating context understanding of LLMs in an ICL setup, we evaluate models quantized using GPTQ (Frantar et al., 2022) , which is an efficient one-shot weight quantization algorithm based on approximate second-order information that compresses the model post-training. It enables a reduction in memory and disk requirements by up to 80%, compared to the pre-quantized model.

[7] In literature, the best FT results come from different models across five QR datasets, where some are not even LLM based. To ensure fair comparison, we fine-tuned a T5 large model on each QR dataset.

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
	ROUGE1	56.07	71.38	77.88
QReCC	BLEU	28.19	53.72	58.68
	ROUGE1	48.27	74.13	78.74
InCar	BLEU	28.20	39.69	48.58
	ROUGE1	49.96	56.32	64.18
GECOR	BLEU	53.27	70.41	83.36
	ROUGE1	69.23	73.80	86.03
CANARD	BLEU	21.34	45.07	47.24
	ROUGE1	38.17	67.15	69.73

Table 7: 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.

4.3 Quantized Model Results

GPTQ (Frantar et al., 2022) has been shown to effectively reduce the model size to 3 bits without incurring substantial performance losses across a range of NLP tasks, such as MMLU, ARC, StoryCloze. However, whether this performance preservation can be extended to contextual understanding was unclear.

Table 7 presents the comparison between the dense and 3-bit quantized LLaMA models. In contrast to previous studies on 3-bit quantization, we observed that quantization leads to fluctuated drops in performance across the four tasks. Specifically, in WSC273, MultiWoz, and CANARD, post-training quantization incurs only a marginal performance drop ($\sim$1.7 points). However, in the remaining datasets, quantization results in significant performance drops.

The results further show that the quantized LLaMA-30B model consistently outperforms the dense LLaMA-7B model across all tasks despite being comparable in disk and memory requirements. For CR, the 30B quantized model achieves significantly higher scores on the OntoNotes dataset across all metrics. The MUC metric shows the most substantial improvement, indicating that the quantized 30B model partially overcomes the tendency to create small clusters for mentions. For DST on MultiWOZ, the 30B quantized model show a 30% relative improvement over the 7B model in JGA. On discourse parsing with PDTB-3, the accuracy of quantized 30B model is almost double, 17.16% vs 31.29%. Across all QR datasets, the quantized 30B model substantially improves NLG scores compared to the dense 7B model, with relative gains ranging from 15-50%. The largest gap is observed on GECOR.

In general, we show that the quantized 30B LLaMA model consistently and significantly outperforms the dense 7B model as a result of the increased scale, despite using 3-bit quantization. The benefits of greater model scale thus outweigh the impacts of quantization in understanding discourse. We believe this finding would be beneficial when deploying LLMs in real-world applications with disk and runtime constraints.

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 8: Comparison between OPT (O.) and LLaMA (L.) across five query rewrite datasets

5 Case Study: Query Rewrite

In this section, we provide in-depth analysis by comparing the two open-sourced model families OPT and LLaMA, and the impact of quantization, using query rewrite as the target task.

5.1 OPT vs. LLaMA

Prior works (Beeching et al., 2023) have consistently shown that, under the same model size, LLaMA outperforms OPT. However, their performance on QR, as shown in Table 8, does not follow this pattern. When the model size is around 7B, OPT consistently performs better than LLaMA by a significant margin across the five QR datasets. The two models perform on par with each other at 13B. The superiority of LLaMA is only obvious with 30B model size. From another perspective, although we expect performance to improve as model size increases, we observe this trend on LLaMA, but not on OPT. These results suggest that it may not be correct to conclude the overall superiority between two model families by only comparing on

Context
`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

Table 9: An example of two major types of errors found in the query rewrite task.

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

a certain range of model sizes or on a certain set of tasks.

5.2 Dense vs. Quantized

We conduct a quantitative analysis on the error types of query rewriting to investigate the performance gap between dense and quantized models with comparable computing requirements. Across the five datasets, we identify two main error types that account for nearly 80% of the total errors. First, the model *repeats* the last query without rewriting the referred entity. In this case, the model seems to understand the instruction but fails to comprehend the referred entity within the context. This type of error can be primarily associated with the model's context understanding capability. Second, the model treats the task as a language modeling (*LM*) task, as shown in Table 9, where it provides a response to the last query. In this scenario, the model appears to struggle to understand the task instruction, even with several few-shot examples. We classify this error type as more related to the model's ICL ability.

We perform manual error annotations on the five QR datasets[8]. Table 10 illustrates the number of errors for the three selected model settings in LLaMA for each dataset. A consistent trend is observed across all QR datasets. In terms of *repeat* errors, the 30B dense model exhibits fewer errors, around half, compared to the 7B dense model (297 vs. 469). However, 3-bit GPTQ quantization leads to an increase in this type of error, reaching a similar error count to the 7B dense model (458 vs. 469). This suggests that 3-bit quantization reduces the model's ability to comprehend the context. Regarding *LM* errors, the 30B dense model also significantly outperforms the 7B dense model, with 106 errors compared to 232. It is to be noted that the quantized model generates 125 *LM* errors, slightly more than the 30B dense model. However, it generates significantly fewer (around 50%) errors compared to the 7B dense model. This indicates that 3-bit quantization maintains the ICL capability when evaluated on our benchmark.

[8] 10% test data on QReCC and CANARD was graded.

6 Conclusion

This paper introduces a contextual understanding benchmark designed to assess the performance of LLMs. We collect nine existing datasets spanning four tasks, each carefully tailored to suit generative models. This benchmark encompasses essential elements for assessing linguistic comprehension within context, including both document and dialog based contextual understanding. Experimental results under an in-context learning setting reveal that LLMs struggle with nuanced linguistic features within this challenging benchmark, exhibiting inconsistencies with other benchmarks that emphasize other aspects of language. To the best of knowledge, we are also the first to compare dense models and post-training quantization models in contextual understanding tasks. This comparison highlights that 3-bit post-training quantization reduces the general understanding capacity of context, particularly in complex references and task-oriented dialogue states. Our proposed contextual comprehension benchmark thus provides a unique perspective on the contextual dimension of language understanding and offers a valuable addition to existing LLM evaluations.

Limitations

This work provides an evaluation of various pre-trained LLMs, including OPT, LLaMA, and GPT, on our understanding benchmark. However, we have not evaluated other LLMs designed for longer input scenarios, such as LongLLaMA (Tworkowski et al., 2023).

In our evaluation, we focus on the GPTQ quantization method, analyzing its performance on our benchmark. We do not include other post-training quantization techniques, such as RPTQ (Yuan et al., 2023), in this work.

Our evaluation concentrates on English datasets, primarily utilizing LLMs pre-trained with English data. All of the four tasks on our benchmark have datasets from other languages. The coreference dataset OntoNotes 5.0 contains annotations of Arabic and Chinese. In addition, recent releases such as CorefUD (Nedoluzhko et al., 2022) promote standardization of multilingual coreference annotations. In DST, CrossWOZ (Zhu et al., 2020) is a cross-domain wizard-of-oz task-oriented dataset. Long et al. (2020) develop TED-CDB, a Chinese discourse relation dataset. The query rewrite task also has datasets in other languages, such as REWRITE (Su et al., 2019) and Restoration-200K (Pan et al., 2019). Finally, specific LLMs optimized for individual languages, such as ChatGLM (Du et al., 2022), exist and are not a part of our evaluation.

References

Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 520–534, Online. Association for Computational Linguistics.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Namo Bang, Jeehyun Lee, and Myoung-Wan Koo. 2023a. Task-optimized adapters for an end-to-end task-oriented dialogue system. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7355–7369, Toronto, Canada. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023b. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard. `https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard`.

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226.

Michael Boratko, Harshit Padigela, Divyendra Mikkilineni, Pritish Yuvraj, Rajarshi Das, Andrew McCallum, Maria Chang, Achille Fokoue-Nkoutche, Pavan Kapanipathi, Nicholas Mattei, Ryan Musa, Kartik Talamadupula, and Michael Witbrock. 2018. A systematic classification of knowledge, reasoning, and context within the ARC dataset. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 60–70, Melbourne, Australia. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind

Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? learning to rewrite questions-in-context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5918–5924, Hong Kong, China. Association for Computational Linguistics.

Elias Frantar and Dan Alistarh. 2023. SparseGPT: Massive language models can be accurately pruned in one-shot. *arXiv preprint arXiv:2301.00774*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. GPTQ: Accurate post-training compression for generative pretrained transformers. *arXiv preprint arXiv:2210.17323*.

Michael Heck, Nurul Lubis, Benjamin Ruppik, Renato Vukovic, Shutong Feng, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, and Milica Gasic. 2023. ChatGPT for zero-shot dialogue state tracking: A solution or an opportunity? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 936–950, Toronto, Canada. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A. Smith, and Mari Ostendorf. 2022. In-context learning for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2627–2643, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Gabriel Murray. 2019. Discourse analysis and its applications. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 12–17, Florence, Italy. Association for Computational Linguistics.

Jeonghoon Kim, Jung Hyun Lee, Sungdong Kim, Joonsuk Park, Kang Min Yoo, Se Jung Kwon, and Dongsoo Lee. 2023. Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization.

Andrey Kuzmin, Markus Nagel, Mart van Baalen, Arash Behboodi, and Tijmen Blankevoort. 2023. Pruning vs quantization: Which is better?

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469, Toronto, Canada. Association for Computational Linguistics.

Nghia T. Le and Alan Ritter. 2023. Are large language models robust zero-shot coreference resolvers?

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *13th International Conference on the Principles of Knowledge Representation and Reasoning, KR 2012*, Proceedings of the International Conference on Knowledge Representation and Reasoning, pages 552–561. Institute of Electrical and Electronics Engineers Inc.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Wei Liu and Michael Strube. 2023. Annotation-inspired implicit discourse relation classification with auxiliary discourse connective generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15696–15712, Toronto, Canada. Association for Computational Linguistics.

Zechun Liu, Barlas Oguz, Changsheng Zhao, Ernie Chang, Pierre Stock, Yashar Mehdad, Yangyang Shi, Raghuraman Krishnamoorthi, and Vikas Chandra. 2023. Llm-qat: Data-free quantization aware training for large language models.

Wanqiu Long, Bonnie Webber, and Deyi Xiong. 2020. TED-CDB: A large-scale Chinese discourse relation dataset on TED talks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2793–2803, Online. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *ACL 2014 System Demonstrations*, pages 55–60.

Scott Martin, Shivani Poddar, and Kartikeya Upasani. 2020. MuDoCo: Corpus for multidomain coreference resolution and referring expression generation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 104–111, Marseille, France. European Language Resources Association.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. LSDSem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, Valencia, Spain. Association for Computational Linguistics.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788.

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. 2020. Up or down? adaptive rounding for post-training quantization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Amir Zeldes, and Daniel Zeman. 2022. CorefUD 1.0: Coreference meets Universal Dependencies. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4859–4872, Marseille, France. European Language Resources Association.

OpenAI. 2022. Optimizing language models for dialogue.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Zhufeng Pan, Kun Bai, Yan Wang, Lianqiang Zhou, and Xiaojiang Liu. 2019. Improving open-domain dialogue systems via multi-turn incomplete utterance restoration. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1824–1833, Hong Kong, China. Association for Computational Linguistics.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. 2019. GECOR: An end-to-end generative ellipsis and co-reference resolution model for task-oriented dialogue. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4547–4557, Hong Kong, China. Association for Computational Linguistics.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis & insights from training gopher.

Michael Regan, Pushpendre Rastogi, Arpit Gupta, and Lambert Mathias. 2019. A dataset for resolving referring expressions in spoken dialogue via contextual query rewrites (cqr). *ArXiv*, abs/1903.11783.

Hui Su, Xiaoyu Shen, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. 2019. Improving multi-turn dialogue modelling with utterance ReWriter. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 22–31, Florence, Italy. Association for Computational Linguistics.

S. Tata and J.M. Patel. 2003. Piqa: an algebra for querying protein data sets. In *15th International Conference on Scientific and Statistical Database Management, 2003.*, pages 141–150.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Szymon Tworkowski, Konrad Staniszewski, Mikołaj Pacek, Yuhuai Wu, Henryk Michalewski, and Piotr Miłoś. 2023. Focused transformer: Contrastive training for context scaling.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf.

Kai-Cheng Yang and Filippo Menczer. 2023. Large language models can rate news outlet credibility.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

Zhihang Yuan, Lin Niu, Jiawei Liu, Wenyu Liu, Xinggang Wang, Yuzhang Shang, Guangyu Sun, Qiang Wu, Jiaxiang Wu, and Bingzhe Wu. 2023. Rptq: Reorder-based post-training quantization for large language models.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, ACL 2020*, pages 109–117.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence?

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. CrossWOZ: A large-scale Chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295.

A Task Design Examples

Table 11 presents the input example for each task. For CR, we only show examples from OntoNotes.

B Few-shot Settings

Table 12 shows the number of examples for each dataset that yields the best scores. All datasets except WSC273 and PDTB3 use randomly selected examples from the training set. Since WSC273 does not include any train or validation set, we use the zero-shot setting, as scores presented in Table 6. For each class in PDTB3, we randomly select two examples from the training set for prompting. For some particular datasets, such as OntoNotes, experiments are only performed in the zero-shot and one-shot settings due to the limitation on input length.

Coreference Resolution
Instructions: Please carefully read the following passages. For each passage and the options, you must identify which option the mention marked in *bold* refers to. If the marked mention does not have any antecedent, please select "no antecedent".

[Few-shot examples]

Context: — basically , it was unanimously agreed upon by the various relevant parties . To express *its* determination , the Chinese securities regulatory department compares this stock reform to a die that has been cast . It takes time to prove whether the stock reform can really meet expectations , and whether any deviations that arise during the stock reform can be promptly corrected . Dear viewers , the China News program will end here . This is Xu Li . Thank you everyone for watching . Coming up is the Focus Today program hosted by Wang Shilin . Good-bye , dear viewers .
Choice:
A. the Chinese securities regulatory department
B. this stock reform
C. the stock reform
D. you
E. everyone
F. no antecedent
Question: What does *its* refers to?
Answer: *A*

Dialogue State Tracking
Consider the following list of concepts, called "slots" provided to you as a json list.

"slots": {"restaurant-pricerange": "price budget for the restaurant",
"restaurant-area": "area or place of the restaurant",
"restaurant-food": "the cuisine of the restaurant you are looking for",
...
"hotel-postcode": "postal code of the hotel",
'hotel-ref": "reference number of the hotel booking"
}

Some "slots" can only take a value from predefined list:

"categorical": {"restaurant-pricerange": ['cheap', 'expensive', 'moderate'],
"restaurant-area": ['centre', 'east', 'north', 'south', 'west'],
"restaurant-bookday": ['monday', 'tuesday', 'wednesday', 'thursday', 'friday', 'saturday', 'sunday'],
...
"hotel-internet": ['free', 'no', 'yes'], "hotel-area": ['centre', 'east', 'north', 'south', 'west']
}

Now consider the following dialogue between two parties called the "system" and "user". Can you tell me which of the "slot" was updated by the "user" in its latest response to the "system"? Present the updates in JSON format. If no "slots" were updates, return an empty JSON list. If you encounter "slot" that was requested by the "user" then fill them with "?". If a user does not seem to care about a discussed "slot" fill it with "dontcare".

Input:
Previous state: {}
"system": ""
"user": "I'm looking for a moderately priced place to eat that's in the centre of town."
Output: *{"restaurant-pricerange": "moderate", "restaurant-area": "centre"}*

Implicit Discourse Relation Classification
Instructions: Given two arguments and a list of connective words, please select the most likely connective between two arguments.

Below are the descriptions of four discourse relation labels. Please find the correct label for each example.

Temporal: The tag temporal is used when the situations described in the arguments are intended to be related temporally.
Contingency: The tag Contingency is used when the situation described by one argument provides the reason, explanation or justification for the situation described by the other.
Comparison: The tag Comparison is used when the discourse relation between two arguments highlights their differ- ences or similarities, including differences between expected consequences and actual ones.
Expansion: The label Expansion is used for relations that expand the discourse and move its narrative or exposition forward.

[Few-shot examples]

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?
A. Temporal
B. Contingency
C. Comparison
D. Expansion
Answer: *C*

Query Rewrite
Instructions: Rewrite the last query following interaction into a well-formed, context independent query. Resolve any disfluencies or grammatical errors in the query.

[Few-shot examples]

Input:
User: Try to reach Forbes now .
Bot: Forbes at Washington Post ? Or Forbes of Publishing Division ?
User: Publishing Division .
Rewrite: *Forbes of Publishing Division*

Table 11: Examples of task design for each task in the context understanding benchmark.

Task	Coreference		DST	Discourse	Query Rewrite				
Dataset	WSC273	OntoNotes	MultiWOZ	PDTB3	MuDoCo	QReCC	InCar	GECOR	CANARD
N-example	Zero-shot	One-shot	5-shot	8-shot	10-shot	5-shot	10-shot	10-shot	5-shot

Table 12: N-shot settings for each task & dataset that yields the highest scores. For each task and model, we use consistent N-shot settings for comparison.